Detecting Patient Deception and Adherence in Diabetes Support Using AI-Generated Conversation Summaries

Leveraging chat summaries to Enhance Doctor-Patient Communication

Hugo Koot
Supervisor(s): Prof. Catholijn M. Jonker, J.D. Top, Msc
[1]EEMCS, Delft University of Technology, The Netherlands
[2]Bernouli Institute, University of Groningen, The Netherlands

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

Name of the student: Hugo Koot
Final project course: CSE3000 Research Project
Thesis committee: Prof. Catholijn M. Jonker, J.D. Top, Msc, Dr. Avishek Anand

An electronic version of this thesis is available at http://repository.tudelft.nl/.

Abstract

Unreliable patient self-reporting complicates diabetes management. This study investigates how AI-generated summaries of patient-chatbot conversations can be structured to help healthcare professionals detect deception and non-adherence. To address this, we developed a novel pipeline by first identifying four key behavioral indicators from literature and then using advanced prompt engineering to automatically flag these in structured summaries. The system's effectiveness was evaluated in an annotation experiment using synthetic chat data. The results showed that the summaries did not improve detection accuracy, increased annotation time, and revealed a critically low inter-annotator agreement. These findings highlight the inherent subjectivity and complexity of the detection task, demonstrating that the developed summarization method is not an effective intervention. Although the approach was unsuccessful, this research contributes a novel summarization pipeline, an open-source annotation tool, and a synthetic dataset, establishing a baseline for future work in enhancing doctor-patient communication.

1 Introduction

Diabetes is a chronic condition requiring persistent self-management, including careful monitoring of blood glucose levels, adherence to medication, and lifestyle adjustments. Despite the availability of sophisticated medical treatments and self-care recommendations, adherence rates remain inconsistent and patient self-reports can be unreliable [1, 2]. Patients may intentionally or unintentionally deceive themselves, their healthcare providers, and their support systems about adherence to prescribed regimens, complicating clinical decision-making.

The Netherlands Organization for Applied Scientific Research (TNO) together with the Hybrid Intelligence project (HI) [3], are developing the CHIP system [4]. The CHIP system is a chatbot made for diabetes patients. The goal of CHIP is to help diabetes patients to improve adherence to their prescribed program and medications. Patients are able to tell the CHIP system how they are doing, how they are adhering and receive advice. However, a problem that arises with self-reporting is that patients can start to deceive the system, hurting its effectiveness [2].

The challenge of detecting deception is fundamentally linked to Theory of Mind (ToM), the capacity to reason about others' mental states such as intentions, beliefs, and desires [5, 6]. While directly integrating advanced computational ToM for autonomous deception detection within chat support systems is a complex and currently unsolved problem, an alternative approach is to support the human healthcare provider. By summarizing and structuring patient interactions to highlight behavioral indicators, we might be able to better equip clinicians to apply their own expertise in assessing patient adherence and sincerity.

In this paper, we aim to address this gap by investigating how AI-generated conversation summaries derived from patient bot interactions can assist healthcare professionals in detecting patient deception and non-adherence. We build upon existing work, particularly research in prompt engineering, and utilize an existing diabetes-support application (CHIP) provided by HI and TNO to implement and evaluate our approach [4].

Specifically, this paper addresses the following main question: *How can we capture and structure relevant patient behavior information from interactions between patients and a diabetes-support chatbot to help doctors detect deception and non-adherence?*

To structure our research, we pose the following concrete subquestions:

1. Which behavioral indicators from existing literature can be used to identify potential deception and non-adherence in patient-chatbot chat logs?

2. How can a prompt engineering pipeline for a Large Language Model be designed to generate structured summaries that specifically flag the indicators identified in SQ1?

3. What is the effect of providing AI-generated summaries on the accuracy and decision time of human annotators when classifying synthetic patient chat logs for deception and non-adherence?

We evaluate our proposed approach through an annotation-based experiment using synthetic data to simulate patient interactions. This experiment allows us to demonstrate a proof-of-concept for our summarization technique's ability to flag possible deception and adherence problems effectively. The AI-generated summary will be considered a successful intervention if it helps annotators achieve either of two outcomes: classify deception and non-adherence more accurately than when using the chat logs alone, or classify faster without any loss of accuracy. We also measure annotation consistency to evaluate if the summaries lead to more reliable and objective judgments, this will also serve as an indicator of the reproducability of the experiment. These criteria will be used to determine whether the tool offers any practical benefit.

2 Methodology

This section details the research methodology, which is organized into three stages. First, we establish the foundation with a literature review on deception indicators and prompt engineering techniques (Section 2.1). Second, we describe the implementation of the AI-powered summarization pipeline (Section 2.2). Finally, we outline the synthetic data generation process (Section 2.3) and the design of the annotation experiment (Section 2.4).

2.1 Literature review

We reviewed research on the indicators of deception and non-adherence to build a system that summarizes patient-bot chats while detecting these behaviors. In parallel, we explored recent advances in prompt engineering to improve the quality of the AI-generated summaries.

2.1.1 Key indicators of deception and non-adherence

Identifying signs of deception and non-adherence is important because patients often provide inaccurate or incomplete information about their medication and lifestyle when self-reporting [2]. Detecting these signs would help healthcare providers understand true patient behavior, enabling better care and tailored support.

Below are four key indicators derived from existing research that can help detect deceptive behavior in patient chats.

Inconsistencies Inconsistent answers are a key indicator of deception, as fabricating information burdens a liar's working memory more than telling the truth [7]. A patient, for example, might claim dietary adherence but later contradict themselves by mentioning off-plan eating. While this could be due to misremembering, research confirms that liars produce more inconsistencies than truth-tellers [7]. This will be fairly trivial to detect in our use case, as we can directly compare all chats the patient has had.

Vague or Evasive Language Patients who are deceptive or non-adherent often use language that introduces uncertainty or avoids specific commitments. This includes frequent use of hedging words like "maybe," "should," or "could," and generalizing terms such as "always" or "everyone" that avoid concrete details. They might respond indirectly or provide irrelevant details instead of answering questions clearly [8]. Research specifically highlights that deceptive communication often involves intentionally unclear language, including ambiguous phrasing, increased usage of modal verbs, and passive voice, to avoid providing verifiable specifics [8]. Similarly, a hallmark of deceptive statements is their low specificity and concreteness: deceivers deliberately omit concrete facts like dates, quantities, or names that could be easily verified [9].

Engagement Level Patients' levels of engagement, meaning how short or long the answers are. Can signal deception or non-adherence, depending on context. Research suggests that in asynchronous text chat contexts similar to the patient bot interactions in the CHIP system, deceivers often produce longer and more elaborate messages, as they have more time to craft convincing stories, thereby converting their deception cues into richer, more detailed text [8].

Gaming the System Patients sometimes give overly consistent or perfect answers, claiming flawless adherence to appear cooperative and avoid negative judgment. Research shows that self-reported adherence measures often show "strong tendencies for over-reporting," resulting in ceiling effects where patients frequently claim perfect adherence despite objective evidence to the contrary [2]. Such exaggerated reporting is typically driven by social-desirability bias, where patients aim to present themselves positively rather than truthfully. Therefore, detecting and flagging responses that seem unrealistically good can be an effective indicator of non-adherence and deception.

Together, these four indicators help to effectively detect deception and non-adherence in patient interactions, making them crucial in our summarization to get the full context.

2.1.2 Prompt engineering techniques

A way to affect the usefulness of LLMs is by prompt engineering. Ever since the popularization of LLMs there has been a lot of research published about prompt engineering techniques that should improve the performance of existing models for specific tasks. As a result, prompt engineering has emerged as a powerful and accessible alternative to fine-tuning. This approach allows practitioners to adapt pretrained models for specialized tasks efficiently and cost-effectively [10–12].

The objective of this literature review into prompt engineering techniques is to design a robust system prompt that accurately detects indicators of patient deception and adherence-related behaviors within user bot chats. We reviewed several recent comprehensive surveys on prompt engineering [10–12], identifying and combining three key prompting techniques to use for the final system prompt.

Implicit Retrieval-Augmented Generation (Implicit RAG) Implicit RAG is a simple but powerful prompting technique: the LLM is told to first find the segments that matter most and then answer the question [11, 13]. This small prompt tweak translates directly into higher accuracy, especially when the context of the prompt becomes larger [13]. This

could be an advantage when a summary has to be generated for multiple chats at the same time.

Annotation-Guideline Prompting Annotation-Guideline prompting packages a task prompt into four blocks [14]:

1. baseline task and output specification
2. annotation-guideline rules
3. error-analysis fixes
4. annotated samples for few-shot learning

This way all domain knowledge is available within the prompt itself. In zero-shot clinical-NER tests, this simple redesign lifted GPT-3.5's relaxed-match F_1 by ≈ 0.09 and GPT-4's by ≈ 0.06 without any fine-tuning [14]. For our use case, this technique will be a useful way to impart knowledge about the key indicators for deception and non-adherence within the system prompt.

Hallucination Rail This is a check after execution, where a new system prompt screens the answers to open-ended questions where no external evidence is available [15]. This should be able to catch some hallucinated facts before they are put into the final summary. We perform the summarization task in two separate instances with the exact same prompt. Then we ask the LLM to compare the answers to these prompt and ask whether the answers agree. If the answers do not agree then we mention in the final summary that there is low confidence in the indicators that do not agree.

Our literature review identified four key behavioral indicators of deception and non-adherence: inconsistencies, vague or evasive language, engagement level, and gaming the system. The review also confirmed that prompt engineering is a powerful, accessible method for tailoring Large Language Models (LLMs) to specialized tasks. Therefore, to build a system capable of detecting these specific indicators, we selected a combination of three prompt engineering techniques. We chose Annotation-Guideline Prompting to embed domain knowledge of the four indicators directly into the prompt, Implicit Retrieval-Augmented Generation (Implicit RAG) to improve accuracy when analyzing long chat histories , and a Hallucination Rail to enhance the reliability of the output by checking for fabricated information.

2.2 Implementation of the Summarization Pipeline

This research extended the existing CHIP system, which was introduced in Section 1. To answer the research questions, a custom summarization module was designed and implemented[1]. This module gives the CHIP system the new capabilities to store conversations and generate summaries focused on detecting deception and non-adherence.

The implemented workflow consists of the following steps:

[1] https://github.com/HugoKoot/CHIPsummary2.git

1. **Chat Termination and Storage:** Functionality was added to allow a patient to end a conversation, which automatically stores the complete chat history. This step was necessary because the current CHIP system processes messages individually without retaining conversation context, which is essential for meaningful summarization.

2. **Automated Summarization:** Once stored, it triggers an automated summarization process. Leveraging configurable system prompts, the LLM generates concise summaries capturing relevant patient behavior indicators across the entire interaction sequence. The LLM used can be swapped by adjusting the API called.

3. **Hallucination rail:** To enhance the reliability of the output, the summarization process is executed twice with the identical prompt. A subsequent LLM call compares the two generated summaries for agreement. Any parts of the summaries that disagree are flagged as *low-confidence* in the final output.

4. **Prompt Engineering and Optimization:** An iterative process was used to refine the system prompts and model parameters. The goal of this optimization was to maximize the quality of the summaries, with a specific focus on identifying the key indicators for deception and non-adherence outlined in Section 2.1.1 (e.g., inconsistencies, evasive language, engagement levels, and gaming the system).

2.3 Data generation

This study utilized synthetic data for the annotation experiment due to the unavailability of real patient chat data and to avoid the ethical complexities of handling actual medical information. The primary motivation for using synthetic data was to create a controlled dataset to evaluate the system's ability to flag specific behavioral indicators.

To be useful for this evaluation, the synthetic data was designed to meet several criteria:

- **Validity:** The aim was for each generated conversation to be classifiable into one of the four experimental categories (e.g., Deceptive, Non-Adhering) to serve as a ground truth for measuring accuracy.

- **Realism:** The chat logs needed to realistically embody the behavioral indicators identified in the literature, such as inconsistencies or evasive language.

- **Relevance:** The conversations needed to be relevant to the diabetes management use case, reflecting interactions a patient might have with the CHIP support chatbot.

The specific prompt used for data generation is included in Appendix A. For the (Truthful, Adhering), (Truthful, Non-adhering), and (Deceptive, Non-adhering) categories, the language model was prompted directly to generate a chat that fit the description. However, this direct approach failed for the (Deceptive, Adhering) category, as the model struggled to generate a plausible chat log for a patient who is both compliant and deceitful.

To resolve this, a targeted, scenario-based approach was used *only for the (Deceptive, Adhering) category*. The model was given a specific scenario in which a patient generally follows medical advice but lies about minor, occasional deviations to appear perfectly compliant. This specific intervention was necessary to generate believable and distinct examples for this otherwise problematic category, ensuring all four conditions could be represented in the experiment.

2.4 Annotation Experiment

This experiment is designed to answer the third sub-question: *What is the effect of providing AI-generated summaries on the accuracy and decision time of human annotators when classifying synthetic patient chat logs for deception and non-adherence?*

Experimental Setup Each annotator is shown every chat in exactly one of two conditions:

1. **Chat logs + Summary**
2. **Chat logs only**

Each annotator receives half of the chats in each condition. The assignment of chats to conditions alternates between annotators, ensuring that all chats are evenly distributed across both conditions for the group as a whole.

Materials For every chat, the following items are prepared:

1. The chat logs.
2. An AI-generated summary containing four deception/non-adherence cues.
3. A ground truth label selected from four categories:
    - Adhering & Truthful
    - Adhering & Deceptive
    - Non-adhering & Truthful
    - Non-adhering & Deceptive

Annotators Four fellow bachelor students were recruited as annotators. Each has prior experience with deception detection through their own research projects. Their familiarity with relevant indicators is expected to enhance both the reliability and interpretability of their annotations.

Annotation Session An annotation web application[2] was developed for this experiment. Upon launching the application, annotators are first instructed to thoroughly read the annotation guidelines found in appendix B within the web application. The researcher overseeing the experiment will then fill in an identification number and assign the annotator to either group 1 or group 2.

After a group is selected, the user can click the *Start Annotation* button to begin. A timer is started in the background, and the annotation interface becomes visible. For each chat, annotators can:

- Select one of the four ground truth categories.
- Select any deception/non-adherence cues they observe.

Once an annotation is submitted, the timer stops, and the annotator can proceed to the next chat by clicking *Start Annotation* again. This process continues until all chats have been annotated.

[2] https://github.com/HugoKoot/annotationExperiment.git

Metrics Collected For each annotator, we collect the following metrics:

1. **Accuracy:** The percentage of correct annotations compared to the ground truth labels.
2. **Decision Time:** The time in seconds taken per chat to arrive at a decision.
3. **Inter-Annotator Reliability:** Assessed using *Krippendorff's* α, which supports multiple annotators, categorical labels, and small datasets [16]. The reliability is interpreted as follows:
   - $\alpha \geq 0.80$: Strong agreement.
   - $0.67 \leq \alpha < 0.80$: Acceptable, but interpret with caution.
   - $\alpha < 0.67$: Poor agreement.

3 System Prompt Design and Output

This section describes how the key indicators and prompt-engineering techniques identified in the methodology (Sections 2.1.1 and 2.1.2) were integrated to create the final system prompt. The prompt is designed to instruct a Large Language Model (LLM) to analyze patient-chatbot conversations and produce a structured summary that flags potential deception and non-adherence. The complete prompt can be found in Appendix C.

3.1 Prompt Architecture

The system prompt is constructed from several components, each drawing on specific prompt-engineering techniques to ensure a reliable and accurate output.

Context and Task Definition The prompt begins by providing the LLM with context about its role as a "compliance summarizer" and the intended audience (a doctor). It then outlines the exact tasks to perform, a practice recommended by the Annotation-Guideline Prompting technique. The length of the summary is set as a variable for easy adaptation.

Indicator Guidelines Following the task definition, the prompt provides explicit annotation rules for the four key behavioral indicators (Inconsistencies, Vague/Evasive Language, Engagement Level, and Gaming the System). This step is another core component of Annotation-Guideline Prompting, which embeds the necessary domain knowledge directly into the prompt.

Implicit-RAG Instructions The prompt then instructs the LLM to first find the most relevant parts of the chat log before generating the summary. This follows the Implicit RAG technique, which improves accuracy, especially with long chat histories, by focusing the model's attention.

Hallucination Rail While not part of the primary prompt itself, the Hallucination Rail is a crucial subsequent step in the workflow. Two summaries are generated using the same prompt, and a second LLM call compares them for agreement. Discrepancies are flagged as "low-confidence" to guard against fabricated information. The prompt for this comparison step can be found in Appendix D.

3.2 Resulting Summary Format

The final output is a JSON file containing two main keys:

- `summary`: a plain-text summary of the patient's chat history.
- `flags`: an array of flagged segments. Each flag includes:
  - the name of the indicator
  - a relevant excerpt from the chat
  - an explanation for the flag
  - a `confidence` field set to `"low"` if the two summaries disagreed

An example of this output format can be found in Appendix E.

4 Experimental Results

This section presents and interprets the results from the annotation experiment detailed in Section 2.4. It is important to note at the outset that the experimental validation involved only four annotators, a small sample size that restricts the generalization and reliability of the findings. The experiment was designed to measure the effect of AI-generated summaries on the accuracy, decision time, and reliability of human annotators in detecting patient deception and non-adherence.

4.1 Accuracy

For this analysis, accuracy is defined as the percentage of annotations that correctly matched the pre-defined ground truth label for each synthetic chat log. The primary finding from the experiment shows that the AI-generated summaries had no effect on this overall accuracy. As illustrated in Figure 1, the average accuracy for annotations made with the chat logs and a summary was identical to annotations made with the chat logs alone.

To observe performance changes over the course of the experiment, we also calculated the rolling average accuracy, a metric that averages performance over a sliding window of recent trials. A noteworthy secondary finding as shown in Figure 2, is the presence of a learning curve in this rolling average. The annotators' accuracy improved as they progressed through the chats, which suggests that familiarity with the task increased their performance over time. Because the order of chats and conditions was randomized, it is possible this learning trend may impact the overall results when trying to replicate this experiment with a low number of annotators.

4.2 Decision Time

The experiment revealed that participants took significantly longer to annotate chats when they were provided with a summary in addition to the raw logs. The boxplot in Figure 3 clearly illustrates that the median annotation time, as well as the interquartile range, was higher in the "with summary" condition.

This suggests that instead of making the task easier or faster, the summaries added to the annotators' cognitive load. Rather than relying on the summary for a quicker judgment, participants likely spent additional time cross-referencing the summary with the chat logs

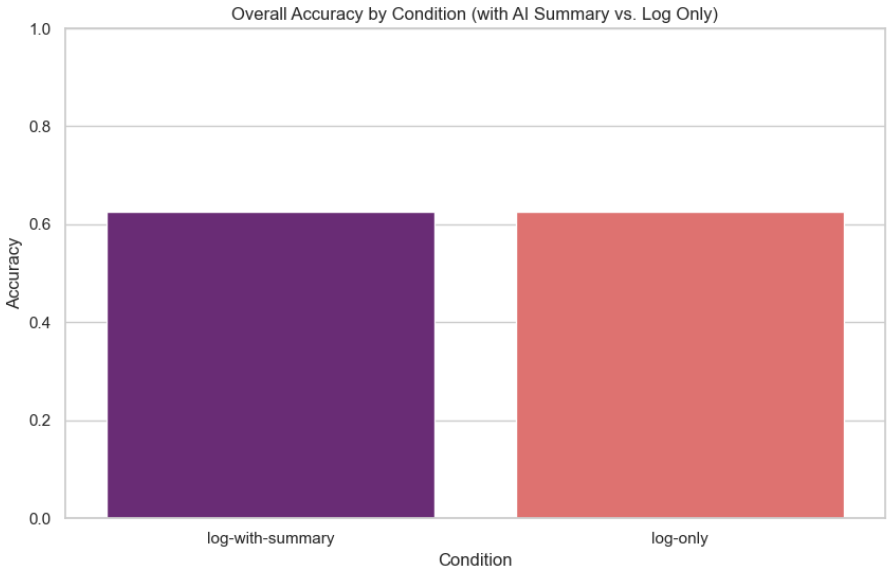

Figure 1: Accuracy by condition (with AI summary vs. log only).

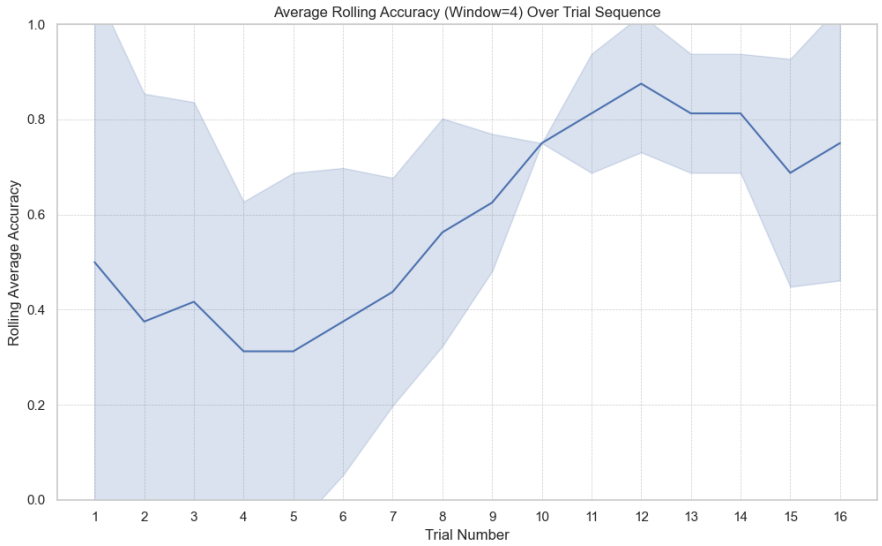

Figure 2: Average rolling accuracy

in an attempt to improve their accuracy. As shown in Figure 4, the annotation times were highest for the initial chats, indicating a "warm-up" period, but the overall trend of longer times with summaries holds.

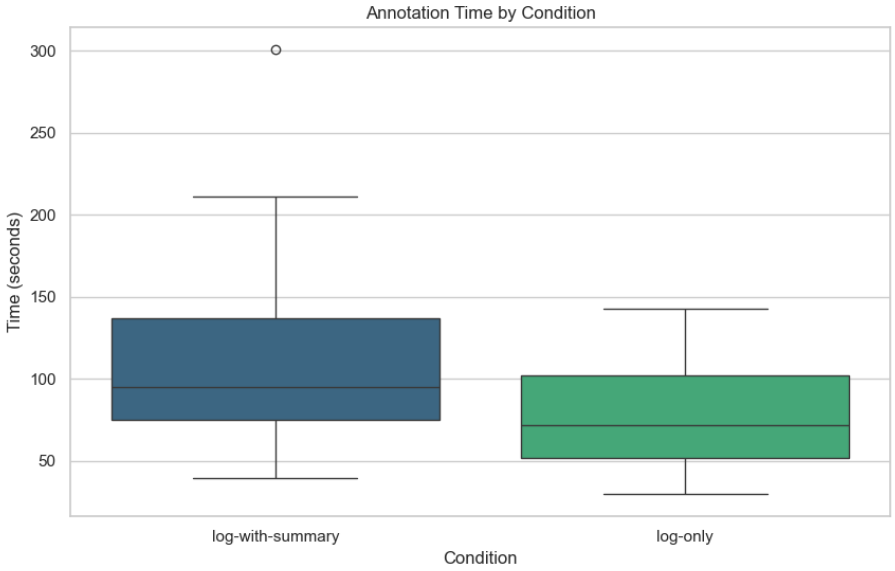

Figure 3: Time by condition

4.3 Reliability

The inter-annotator agreement that was measured using Krippendorff's Alpha [16], was low across all categories. This indicates poor reliability among the annotators (Figure 5). This is a critical finding, as it suggests that the task of identifying deception and non-adherence from these chat logs is highly subjective. The low agreement means that the annotations are highly dependent on the individual annotator, and the process would not be consistently reproducible with a different group of people.

4.4 Summary of Experimental Findings

In summary, the experimental results indicate that the AI-generated summaries did not achieve their intended goal. They failed to improve detection accuracy and, contrary to expectations, they increased the time required for annotation. The low inter-annotator agreement highlights the inherent difficulty and subjectivity of the detection task itself. These findings suggest that the summaries implemented and evaluated in this study, did not provide any benefit to the annotators.

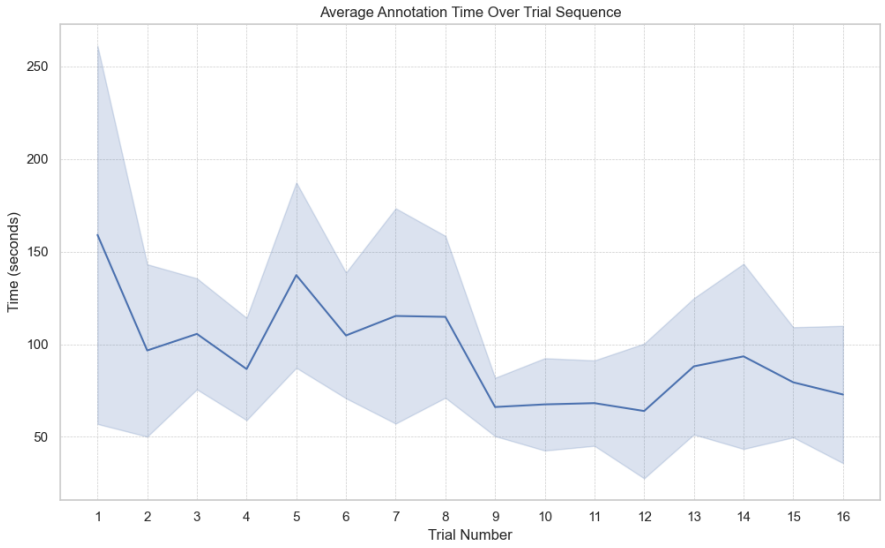

Figure 4: Average annotation time over trial

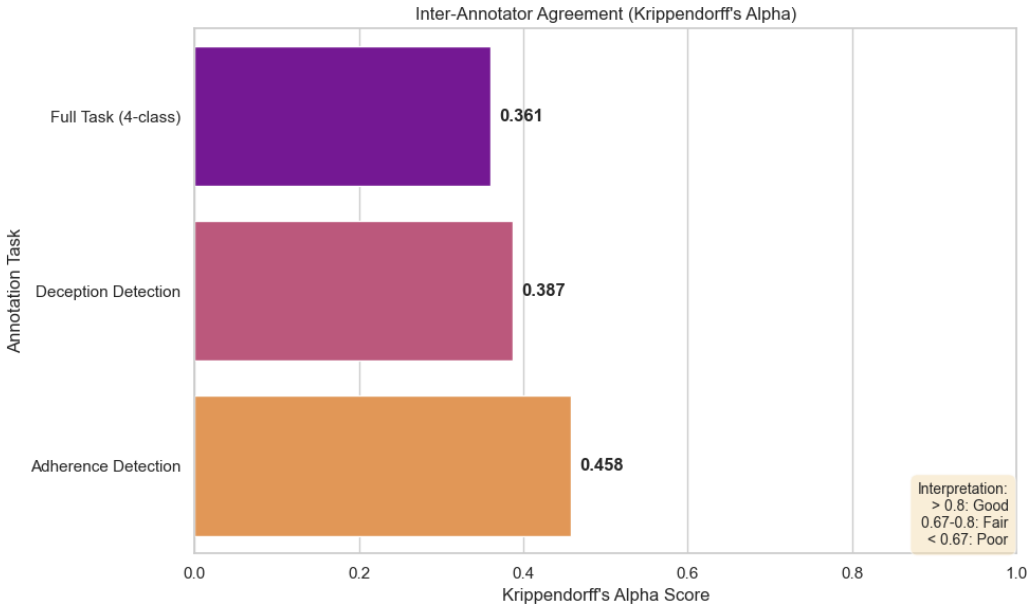

Figure 5: Inter-annotator agreement (Krippendorff's Alpha)

5 Responsible Research

While this research does not utilize patient data and instead relies on synthetic conversation data generated for experimental validation, the potential practical applications of the developed methods raise significant ethical considerations that must be proactively addressed.

If deployed in practice, the proposed enhancement to the CHIP system would require stringent data protection measures for handling real patient medical data. The current architecture, which transmits logs to a third-party LLM (Gemini) for summary generation, is an insufficient safeguard for sensitive patient information. Beyond data privacy, this approach introduces other significant risks inherent to LLMs, including the potential for fabricated "hallucinations" in summaries, a lack of explainability for the model's conclusions, and the perpetuation of algorithmic biases, all of which challenge safe and equitable clinical use.

A more responsible, long-term solution would involve replacing the third-party service with a locally hosted or first-party LLM to ensure sensitive data remains within a secure institutional environment. While this would solve the critical data transmission issue, the risks of hallucinations, bias, and poor explainability would still require robust safeguards. This includes mitigation techniques like the "Hallucination Rail" implemented in this research, alongside continuous monitoring and validation before any deployment.

Patients should explicitly provide informed consent for generating these summaries and sharing them with their healthcare providers. When obtaining consent, it is imperative to clearly communicate the rationale and intended use of these AI-generated summaries, emphasizing their supportive role in patient care.

Additionally, careful thought must be given to how the information from the summaries is communicated to patients by their doctors, ensuring transparency but also taking into account how the patients perceive the feedback. The primary purpose should always remain patient-centered, aiming to support rather than chastise individuals that may be deceitful or non-adhering.

To support scientific transparency, this research is designed to be as reproducible as possible. The methodology is explicitly documented, detailing the data generation procedures, summarization methods, and the specific prompt engineering techniques employed. Furthermore, the complete source code for the summarization pipeline and the annotation tool are open GitHub repositories[3][4], and the prompts used for generation are included in the appendix, allowing future researchers to replicate and build upon this paper's findings. However, a potential challenge to exact replication is the specific LLM used, Gemini 2.5 flash. The availability of this model may change, and future updates could alter its behavior. Consequently, future researchers may need to use a different LLM, which could lead to variations in the results.

To ensure responsible data handling, all annotators provided an informed consent form about which they could ask questions before signing. They were advised of their right to pause or withdraw from the study at any time. The privacy of participants was protected through the anonymization of all collected data, which will be permanently deleted upon the conclusion of the research.

To assist in the writing of this paper ChatGPT versions 4o and 4.5, were used for proofreading. The models were engaged with prompts limited to identifying spelling errors and providing suggestions to improve the flow of the text. These prompts can be found in Ap-

[3]https://github.com/HugoKoot/CHIPsummary2.git
[4]https://github.com/HugoKoot/annotationExperiment.git

pendix F. Care was taken to avoid having the AI rewrite any portion of the text, and no responses from the models were copied verbatim.

6 Discussion

This study aimed to evaluate the effectiveness of AI-generated conversation summaries in assisting healthcare providers to detect deception and non-adherence among diabetes patients. The findings showed no improvement in annotation accuracy when AI-generated summaries were provided. Additionally, annotators required more time when using summaries, suggesting increased cognitive load due to the extra information available.

Low inter-annotator agreement highlighted the subjective and complex nature of textual deception detection. This variability undermines the reliability of the results as it shows even trained annotators looking for specific cues can struggle to consistently interpret indicators.

An observed learning curve effect suggested annotators improved with familiarity, indicating potential for better outcomes with enhanced training and standardized guidelines. However, the reliance on synthetic data limited ecological validity. Future research should validate these methods using real patient interactions.

6.1 Limitations

Several limitations must be acknowledged in interpreting the results of this study. First, the experimental validation involved only four annotators, constituting a small sample size. This limitation restricts the generalization and reliability of the findings.

Secondly, systematic comparison of various LLMs and their configurations was not possible due to time constraints. The indicators selected for detecting deception and non-adherence, as well as the specific prompt engineering techniques applied, were chosen based on theoretical considerations and literature review and were not tested and compared against each other in an experimental setup due to time constraints.

Additionally, this research was conducted exclusively with synthetic data due to the unavailability of actual patient chat logs. Therefore, validation with real-world data remains a critical step for future studies to ensure the practical effectiveness and accuracy of the developed methods.

Lastly, there was no "Summary only" condition in the annotation experiment. This means that the effect of completely replacing the chat logs was not measured and the effect of the summary was not fully explored in the performed experiment.

7 Conclusions and Future Work

This research sought to answer the question: *How can we capture and structure relevant patient behavior information from interactions between patients and a diabetes-support chatbot to help doctors detect deception and non-adherence?* To address this, the study first investigated three concrete sub-questions.

First, the research asked which behavioral indicators could identify deception and non-adherence in chat logs. This was answered through a literature review, which identified four key indicators: inconsistencies, vague or evasive language, engagement level, and gaming the system.

Second, the study questioned how a prompt engineering pipeline could be designed to flag these indicators. This was answered by building and implementing a novel summarization pipeline that leveraged a combination of three specific techniques: Annotation-Guideline Prompting, Implicit Retrieval-Augmented Generation, and a Hallucination Rail to improve output reliability.

The third and central sub-question asked what effect these AI-generated summaries would have on the accuracy and decision time of human annotators. The answer, derived from an annotation experiment, was that the summaries had a negative impact. They failed to improve detection accuracy and increased the time taken to make a decision. This result, however, must be contextualized by the experiment's limitations, including a small sample size of only four annotators and the exclusive use of synthetic data. Furthermore, the experiment revealed a low inter-annotator agreement, a critical finding in itself that highlights the subjective difficulty of the detection task.

Ultimately, these findings provide a direct, but negative, answer to the main research question of how to effectively capture and structure patient information. This study proposed and built a specific method for this task: a summarization pipeline that uses prompt engineering to flag behavioral indicators. The experimental results demonstrated, however, that this particular method of structuring information does not help doctors detect deception and non-adherence more effectively. Therefore, the initial hypothesis that this approach would be beneficial was shown to be incorrect.

Despite this outcome, the study produced several contributions, including the novel summarization pipeline, an open-source annotation tool, and a synthetic dataset, all of which can serve as a foundation for future research.

7.1 Future Work

Based on our findings, we recommend the following directions for future research:

1. **Systematic Component Evaluation:** The current study combined multiple prompt engineering techniques and behavioral indicators. Future work should systematically isolate and test these individual components to determine if any single indicator or technique has a genuinely predictive impact when used alone. This would help identify the most effective elements for generating useful summaries.

2. **Real-World Validation:** This research was conducted exclusively with synthetic data. It is critical that any promising methods be validated using real-world patient data. Furthermore, healthcare professionals should be involved in the evaluation process to ensure the clinical relevance and usability of any developed tools.

3. **Refined Experimental Design:** This study compared annotations with chat logs alone versus logs supplemented by a summary. To better isolate the summary's independent value, a future experiment should compare one group reviewing only the raw chat logs with another group reviewing only the AI-generated summary. This would directly test if the summary can effectively replace the full transcript for making an accurate judgment.

4. **Validating Utility:** This work operated on the assumption that summaries focused on deception and non-adherence would be valuable to clinicians. This assumption should be tested directly. Future research should survey and interview healthcare professionals

to determine what information from patient chat logs they would actually find useful in a clinical setting, ensuring that future tools are aligned with their practical needs.

References

[1] 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", *The American Journal of Medicine*, vol. 77, no. 2, pp. 211–217, 1984. DOI: 10.1016/0002-9343(84)90693-4.

[2] M. J. Stirratt, J. Dunbar-Jacob, H. M. Crane, J. M. Simoni, S. Czajkowski, M. E. Hilliard, J. E. Aikens, C. M. Hunter, D. I. Velligan, K. Huntley, *et al.*, "Self-report measures of medication adherence behavior: Recommendations on optimal use", *Translational behavioral medicine*, vol. 5, no. 4, pp. 470–482, 2015. DOI: 10.1007/s13142-015-0315-2.

[3] 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. Verheij, 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.

[4] B. J. W. Dudzik, J. S. van der Waa, P.-Y. Chen, R. Dobbe, Í. M. D. R. de Troya, R. M. Bakker, M. H. T. de Boer, Q. T. S. Smit, D. Dell'Anna, E. Erdogan, P. Yolum, S. Wang, S. B. Santamaría, L. Krause, and B. A. Kamphorst, "Viewpoint: Hybrid intelligence supports application development for diabetes lifestyle management", English, *Journal of Artificial Intelligence Research*, vol. 80, pp. 919–929, 2024, ISSN: 1076-9757. DOI: 10.1613/jair.1.15916.

[5] 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.

[6] H. de Weerd, R. Verbrugge, and B. Verheij, "How much does it help to know what she knows you know? An agent-based simulation study", *Artificial Intelligence*, vol. 199–200, pp. 67–92, 2013. DOI: 10.1016/j.artint.2013.05.004.

[7] J. J. Walczyk, K. T. Mahoney, D. Doverspike, and D. A. Griffith-Ross, "Cognitive lie detection: Response time and consistency of answers as cues to deception", *Journal of Business and Psychology*, vol. 24, no. 1, pp. 33–49, 2009, ISSN: 1573-353X. DOI: 10.1007/s10869-009-9090-8.

[8] L. Zhou, J. K. Burgoon, J. F. Nunamaker, and D. Twitchell, "Automating linguistics-based cues for detecting deception in text-based asynchronous computer-mediated communications", *Group Decision and Negotiation*, vol. 13, pp. 81–106, 2004. DOI: 10.1023/B:GRUP.0000011944.62889.6f.

[9] R. Loconte, R. Russo, P. Capuozzo, P. Pietrini, and G. Sartori, "Verbal lie detection using large language models", *Scientific Reports*, vol. 13, no. 1, p. 22 849, 2023, ISSN: 2045-2322. DOI: 10.1038/s41598-023-50214-0.

[10] P. Sahoo, A. K. Singh, S. Saha, V. Jain, S. Mondal, and A. Chadha, "A systematic survey of prompt engineering in large language models: Techniques and applications", *arXiv preprint arXiv:2402.07927*, 2024. DOI: 10.48550/arXiv.2402.07927.

[11] S. Vatsal and H. Dubey, "A survey of prompt engineering methods in large language models for different nlp tasks", *arXiv preprint arXiv:2407.12994*, 2024. DOI: 10.48550/arXiv.2407.12994.

[12] X. Amatriain, "Prompt design and engineering: Introduction and advanced methods", *arXiv preprint arXiv:2401.14423*, 2024. DOI: 10.48550/arXiv.2401.14423.

[13] S. Vatsal and A. Singh, "Can gpt redefine medical understanding? evaluating gpt on biomedical machine reading comprehension", *arXiv preprint arXiv:2405.18682*, 2024. DOI: 10.48550/arXiv.2405.18682.

[14] Y. Hu, Q. Chen, J. Du, X. Peng, V. K. Keloth, X. Zuo, Y. Zhou, Z. Li, X. Jiang, Z. Lu, K. Roberts, and H. Xu, "Improving large language models for clinical named entity recognition via prompt engineering", *Journal of the American Medical Informatics Association : JAMIA*, vol. 31, no. 9, pp. 1812–1820, 2024, ISSN: 1527-974X. DOI: 10.1093/jamia/ocad259.

[15] T. Rebedea, R. Dinu, M. Sreedhar, C. Parisien, and J. Cohen, "NeMo guardrails: A toolkit for controllable and safe llm applications with programmable rails", *arXiv:2310.10501*, 2023. DOI: 10.48550/arXiv.2310.10501.

[16] K. Krippendorff, "Reliability in content analysis", *Human Communication Research*, vol. 30, no. 3, pp. 411–433, 2004. DOI: 10.1111/j.1468-2958.2004.tb00738.x.

A Data generation prompt

```
Imagine you are diabetes patient talking to a diabetes support chatbot.

I want to get the logs of the chats you would have with this bot.

In this chat you will talk to the bot about your adherence to the doctor
prescribed lifestyle. Like eating the right things, moving enough and
keeping you insulin at the right levels. You will also log your insulin
levels in the chat and ask for advice. The chatbot does not have information
about the measured insulin or glucose levels, other than the information
the patient provides in the chat. The chatbot does not confront the user
when the user has lied.

However sometimes patients lie to this chatbot about their adherence or
insulin levels. these are the key indicators of deception and
non-adherence:

Inconsistencies: Patients may contradict themselves across conversations
(e.g., claiming dietary adherence but later mentioning off-plan eating).
When patients provide contradictory information it either means they are
being deceptive, or that they are misremembering. This will be fairly
trivial to detect in our use case as we will be able to directly compare
all the chats the patient has had.

Vague or Evasive Language: Patients who are deceptive or non-adherent
often use language that introduces uncertainty or avoids specific
```

commitments. This includes frequent use of hedging words like "maybe," "should," or "could," and generalizing terms such as "al ways" or "everyone" that avoid concrete details. They might respond indirectly or provide irrelevant details instead of answering questions clearly. Research specifically highlights that deceptive communication often involves intentionally unclear language, including am biguous phrasing, increased usage of modal verbs, and passive voice, to avoid providing verifiable specifics. Similarly, a hallmark of deceptive statements is their low specificity and concreteness deceivers deliberately omit concrete facts like dates, quantities, or names that could be easily verified.

Engagement Level: Patients levels of engagement, meaning how short or long the answers are. Can signal deception or non-adherence, depending on context. Research suggests that in asynchronous text chat contexts similar to the patient bot interactions in the CHIP system, deceivers often produce longer and more elaborate messages, as they have more time to craft convincing stories, thereby converting their deception cues into richer, more detailed text.

Gaming the System: Patients sometimes give overly consistent or perfect answers, claiming flawless adherence to appear cooperative and avoid negative judgment. Research shows that self-reported adherence measures often show "strong tendencies for over-reporting," resulting in ceiling effects where patients frequently claim perfect adherence despite objective evidence to the contrary. Such exaggerated reporting is typically driven by social desirability bias, where patients aim to present themselves positively rather than truthfully. Therefore detecting and flagging responses that seem unrealistically good can be an effective indicator of non-adherence and deception.

Together, these four indicators help to effectively detect deception and non-adherence in patient interactions, making them crucial in our summarization to get the full context.

I want to generate several of these chatlogs. Each log would need to fall into one of the following categories:

(Truthfull, adhering), (truthfull, non-adhering), (deceptive, adhering), (deceptive, non-adhering)

Give the chatlog in json. for each message also give the time. You can also give multiple seperate chatlogs representing multiple consecutive chats.

After all the chatlogs you should state what key indicators you used to make the chatlog the truthfull/deceitful/adhering/non-adhering. Als make sure the human does not adress the bot as 'bot' but chats to it the same way people chat to an llm

```
First give the chatlogs for the [Insert desired category] category

You do NOT have to use all key indicators

The logs should follow this format and be multiple chats long of varying
length:

{
 "patient_name": "John",
 "timestamp": "20250520_080443",
 "messages": [
   {
        "message": "message text",
        "user": {
            "name": "John",
            "human": true
        },
        "timestamp": "2025-05-20T08:04:16.761Z"
    },
    {
        "message": "message text",
        "user": {
            "name": "Bot",
            "human": false
        },
        "timestamp": "2025-05-20T08:04:17.289Z"
    }
 ]
}
```

B Annotator Guidelines

Your Task

Your primary task is to read patient-bot chat logs and classify each log based on the evidence you find. You will also be asked to select the specific indicators that led you to your conclusion. Please base your judgments solely on the indicators described below.

Annotation Categories

You will classify each log into one of the following four categories:

- **Truthful, Adhering:** The patient is honest and generally follows the program guidelines.
- **Truthful, Non-Adhering:** The patient is honest about their struggles or failure to follow the program.

- **Deceitful, Adhering:** The patient is generally following the program but is dishonest about certain aspects.
- **Deceitful, Non-Adhering:** The patient is dishonest and not following the program guidelines.

Definition of "Adhering"

Please note that adhering does not mean the patient is 100% perfect. A patient can be considered adhering if they are generally following the program's core principles and taking their participation seriously.

Key Indicators

Please read these definitions carefully. You must annotate according to these specific indicators.

Inconsistencies

Patients may contradict themselves across conversations (e.g., claiming dietary adherence but later mentioning off-plan eating). When patients provide contradictory information it either means they are being deceptive, or that they are misremembering. This will be fairly trivial to detect in our use case as we will be able to directly compare all the chats the patient has had.

Vague or Evasive Language

Patients who are deceptive or non-adherent often use language that introduces uncertainty or avoids specific commitments. This includes frequent use of hedging words like maybe, should, or could, and generalizing terms such as "always" or "everyone" that avoid concrete details. They might respond indirectly or provide irrelevant details instead of answering questions clearly. Research specifically highlights that deceptive communication often involves intentionally unclear language, including ambiguous phrasing, increased usage of modal verbs, and passive voice, to avoid providing verifiable specifics. Similarly, a hallmark of deceptive statements is their low specificity and concreteness; deceivers deliberately omit concrete facts like dates, quantities, or names that could be easily verified.

Engagement Level

Patients' levels of engagement, meaning how short or long the answers are, can signal deception or non-adherence, depending on context. Research suggests that in asynchronous text chat contexts similar to the patient-bot interactions in the CHIP system, deceivers often produce longer and more elaborate messages, as they have more time to craft convincing stories, thereby converting their deception cues into richer, more detailed text.

Gaming the System

Patients sometimes give overly consistent or perfect answers, claiming flawless adherence to appear cooperative and avoid negative judgment. Research shows that self-reported adherence measures often show "strong tendencies for over-reporting," resulting in ceiling

effects where patients frequently claim perfect adherence despite objective evidence to the contrary. Such exaggerated reporting is typically driven by social-desirability bias, where patients aim to present themselves positively rather than truthfully. Therefore detecting and flagging responses that seem unrealistically good can be an effective indicator of non-adherence and deception.

Timed Annotations

For each chat log, the annotation process is timed. The timer will start the moment you click the **Start Annotation & Timer** button, which is also when the chat log and summary (if applicable) will become visible. The timer stops when you click "Submit & Next". This helps us understand the time taken for each annotation.

Understanding AI Summaries

In some cases, you will be shown an AI-generated summary alongside the chat log. These summaries are structured with a main summary and a list of "flags".

A flag marked with **(Low Confidence)** indicates that the AI detected a potential indicator but was not certain. This happens when two different AI runs did not agree on the flag. You should evaluate these low-confidence flags with extra scrutiny.

Example Summary Structure: **Summary:** The patient reported general adherence but struggled with exercise...
Flags:

- VagueLanguage **(Low Confidence)**: I ate okay.
- Inconsistency: Said they took medication, but later mentioned forgetting.

C System Prompt

```
You are a compliance summariser for diabetes patient-bot chats that should
take note of deception and non-adherence according to the provided indicators.
Audience: the prescribing doctor who will read a one-page update before the
next appointment.

================  TASK  ================================
1. **Adherence & Progress Summary** In < {N} sentences, describe:
   - What the patient reported doing (diet, medication, activity).
   - Any successes, struggles, or trends.
   - Overall adherence status and momentum.
2. **Risk Flags** List every chat excerpt that matches an indicator below.
3. Output the exact JSON schema shown at the end.

================  INDICATOR GUIDELINES  ================
- **Inconsistency**  Flag any statement that clashes with the patient's
earlier chats.
- **Vague / Evasive Language** Watch for hedges (maybe, should, could),
sweeping generalities (always, everyone), passive phrasing, or missing
```

```
specifics (dates, quantities, names). Low-concreteness wording is a
hallmark of deceptive replies. Rate severity {low | medium | high}.
- **Engagement Level** Unusually long, highly detailed stories may
indicate a crafted (and possibly false) narrative. Either extreme can
suggest non-adherence or deception; interpret in context.
- **Gaming the System**  Identify too-perfect self-reports: identical
answers across check-ins, flawless adherence claims, or copy-pasted text.

================  IMPLICIT-RAG INSTRUCTIONS  ================
Step 1 From the full transcript, **extract up to {K} excerpts
(30 - 120 words each)** that are most relevant to any indicator above.
Step 2 Using **only those excerpts**, perform TASK 1 and 2.
Step 3 Return:

json
{
  "summary": "< {N} sentences>",
  "flags": [
    {
      "indicator": "Inconsistency",
      "excerpt": "...",
      "explanation": "..."
    },
    {
      "indicator": "VagueLanguage",
      "excerpt": "...",
      "explanation": "..."
    }
    /* 0-N more flags */
  ]
}
```

D System prompt comparison

```
You are a verification and synthesis AI. Your task is to analyze two
different AI-generated summaries and their corresponding 'flags' based
on the same source text. Your goal is to produce a single, more
accurate and reliable final JSON output.

You will receive a JSON object with four keys: "summary1", "flags1",
"summary2", and "flags2".

**Your task is to perform two main actions:**

**1. Synthesize the Summaries:**
   - Read both ‘summary1‘ and ‘summary2‘.
   - Combine their insights to create a single, more comprehensive
   and accurate final summary.
```

```
   - The final summary should be objective and reflect the consensus
   between the two inputs.

**2. Verify and Consolidate the Flags:**
   - Compare ‘flags1‘ and ‘flags2‘ to identify semantically equivalent flags.
   - A **Direct Match** occurs when a flag from one list clearly refers to
   the same event or statement as a flag in the other list, even if the
   wording differs slightly. Matched flags should be included once in the
   final list without any ’confidence’ field.
   - A **Mismatch** occurs when a flag from either list does NOT have a clear
   semantic equivalent in the other. Mismatched flags MUST have a
   ‘"confidence": "low"‘ field added to them.
   - The final list of flags should not contain duplicates.

**Output Instructions:**
You MUST return a single, valid JSON object with two top-level keys:
- ‘"summary"‘: The new, synthesized summary string.
- ‘"flags"‘: The final, consolidated list of flag objects.
```

E Summary Example

```
{
  "summary": "Patient John initially reported excellent adherence,
      stating his fasting glucose was 95 mg/dL after taking 10 units
      of long-acting insulin, and confirmed taking rapid-acting
      insulin before lunch. He claimed his diet and exercise were \"
      perfect,\" consisting of vegetables, lean protein, and a daily
      walk. However, he later reported a post-lunch glucose reading of
       190 mg/dL, which is higher than his target. Despite initially
      insisting his lunch was \"just the usual\" grilled chicken salad
      , he subsequently attributed the high reading to being \"a bit
      off from all the excitement\" from a colleague’s birthday
      festivities the previous day, implying a deviation from his
      routine and dietary plan. This suggests a struggle with
      maintaining consistent adherence during social events and a
      pattern of evasiveness in reporting his food intake. Overall,
      John demonstrates mixed adherence, showing good medication
      compliance but significant inconsistencies and potential non-
      adherence regarding dietary habits, especially around social
      occasions, indicating inconsistent progress momentum.",
  "flags": [
    {
      "indicator": "Inconsistency",
      "excerpt": "Everything has been perfect. I’ve been sticking to
          the plan, eating all the right things, you know. Lots of
          vegetables and lean protein. I also went for my daily walk
          yesterday evening. [...] No, it was just the usual. It was
          my colleague’s birthday yesterday, so maybe I’m just a bit
          off from all the excitement. I probably should have just
```

```
        stuck to my routine and not engaged in the office
        festivities.",
    "explanation": "John initially claims 'everything has been
        perfect' with his diet and sticking to the plan. However, he
         later reveals he 'engaged in office festivities' for a
        colleague's birthday, which likely involved food that
        contributed to his high post-lunch glucose, directly
        contradicting his earlier claim of perfect adherence."
  },
  {
    "indicator": "VagueLanguage",
    "excerpt": "Everything has been perfect. I've been sticking to
        the plan, eating all the right things, you know. Lots of
        vegetables and lean protein. I also went for my daily walk
        yesterday evening. [...] No, it was just the usual. It was
        my colleague's birthday yesterday, so maybe I'm just a bit
        off from all the excitement. I probably should have just
        stuck to my routine and not engaged in the office
        festivities.",
    "explanation": "John uses sweeping generalities like 'everything
         has been perfect' and 'all the right things' without
        providing specific examples or quantities, indicating low-
        concreteness (Severity: Medium). He later vaguely attributes
         his high glucose to 'excitement' from 'office festivities'
        rather than admitting to specific food consumption, which is
         highly evasive (Severity: High)."
  },
  {
    "indicator": "EngagementLevel",
    "excerpt": "Everything has been perfect. I've been sticking to
        the plan, eating all the right things, you know.",
    "explanation": "This overly positive and generalized statement,
        without specific details, could indicate a desire to provide
         a quick, satisfactory answer rather than fully engaging in
        a detailed self-assessment of adherence. It's a 'too-perfect
        ' self-report that bypasses deeper reflection.",
    "confidence": "low"
  },
  {
    "indicator": "EngagementLevel",
    "excerpt": "No, it was just the usual. It was my colleague's
        birthday yesterday, so maybe I'm just a bit off from all the
         excitement. I probably should have just stuck to my routine
         and not engaged in the office festivities.",
    "explanation": "The sudden revelation of a 'birthday' and '
        festivities' to explain a high glucose reading, after
        initial claims of 'perfect' adherence and 'just the usual'
        lunch, suggests a crafted and unusually detailed narrative.
        This could be an attempt to manage perception without
        directly admitting dietary non-adherence.",
    "confidence": "low"
  },
```

```
    {
      "indicator": "GamingTheSystem",
      "excerpt": "Everything has been perfect. I've been sticking to
          the plan, eating all the right things, you know. Lots of
          vegetables and lean protein. I also went for my daily walk
          yesterday evening.",
      "explanation": "John's claim that 'Everything has been perfect'
          and 'all the right things' constitutes a too-perfect self-
          report of flawless adherence, especially considering his
          later, contradictory admission of engaging in 'office
          festivities' that likely involved dietary deviations."
    }
  ]
}
```


F ChatGPT writing prompts

Prompt for detecting grammatical errors Find grammatical and spelling errors in this text: [insert text]

Prompt for imrpoving flow Give tips on how to improve the flow for this text: [insert text]