Entropy-Based Modeling For Detecting Behavioral Anomalies in Users of a Diabetes Lifestyle Management Support System

Identifying non-adherence indicators in a chatbot-based diabetes support system

Sorin - Andrei Ciuntu[1]

Supervisor(s): Prof. C.M. Jonker[1]**, J.D. Top, MSc**[2]
Prof. L.C. Verbrugge[2]**, Dr. H.A. de Weerd**[2]

[1]**EEMCS, Delft University of Technology, The Netherlands**
[2]**Bernoulli 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: Sorin - Andrei Ciuntu
Final project course: CSE3000 Research Project
Thesis committee: Prof. C.M. Jonker, J.D. Top, Msc, Dr. A. Anand

Abstract

Individuals with diabetes face rigorous demands when it comes to managing their health, yet patients sometimes struggle to stay adherent to treatment. CHIP is an AI-based conversational platform that allows patients to report lifestyle factors and receive personalized support for making healthy lifestyle changes. However, detecting patient non-adherence remains a significant challenge in this system, as this can hinder treatment and complicate decision-making for healthcare providers.

This study presents an anomaly detection system designed to identify behavioral changes in diabetes patients through their chatbot interactions. Such shifts have previously been shown to correlate with non-adherence. The approach extracts temporal, frequency, and content features from patient-chatbot conversations and quantifies behavioral variability using entropy to detect deviations from individual baseline patterns.

The approach was evaluated using synthetic patient-chatbot conversations generated by a locally-hosted large language model, with behavioral shifts manually introduced in the simulated users. The system detected these irregularities with an accuracy of approximately 76% and a recall of around 35%. However, the false positive rate remained high, at around 15%, primarily due to over-flagging in users with naturally high variability. Future improvements could involve machine learning-based personalization to better distinguish between true anomalies and normal variability. With refined detection thresholds, integration into CHIP may enable timely support for patients at risk of non-adherence.

1 Introduction

Regular glucose monitoring remains a persistent challenge for individuals with diabetes, a condition that affects approximately 10% of the world population [1]. In a study conducted in Brazil, more than 88% of patients with diabetes reported performing glucose self-monitoring to guide insulin therapy [2]. However, research indicates that despite the associated health risks, patients may misreport their glucose levels when using self-reporting monitoring tools [3; 4]. Additionally, another study found that only around 56% of participants with type 2 diabetes reported full adherence to their prescribed medication regimen [5]. This lack of consistent adherence and accuracy in self-reporting creates challenges for both patients and healthcare providers, who rely on accurate data to guide treatment decisions.

To support lifestyle changes, the Hybrid Intelligence Centre (HI) [6] and the Netherlands Organisation for Applied Scientific Research (TNO) have collaborated to develop CHIP, an AI-based diabetes lifestyle management support system [7]. Patients can interact with an AI chatbot to report glucose levels, diet, and lifestyle habits, and receive personalized feedback and advice. A key challenge in this system is identifying non-adherence, where patients fail to follow their prescribed medication or lifestyle regimen, deliberately or not.

Research on patients of severe mental illness indicates that behavioral shifts, defined as deviations from an individual's observed patterns, may be correlated with non-adherence [8]. In this study, we propose a method to model and analyze patient behavior over time to detect and flag such shifts, which could be implemented into CHIP to flag potential issues and offer timely support.

Anomaly detection is a data analysis method used to flag data points that deviate from what is normally expected in a dataset. In this study, we focus on contextual anomalies, which are unusual only when interpreted in relation to nearby data points. For example, if a highly organized patient typically interacts with the chatbot each morning and evening, a sudden change to logging during the night would be considered a contextual anomaly because it deviates from their established routine. By comparing behavior against each individual's own baseline, the system adapts to personal habits and avoids relying on predefined ground truths about what is considered a normal pattern.

There is growing interest in modeling behavioral patterns in healthcare contexts to predict mental state, treatment disengagement, and treatment non-adherence. Recent studies have successfully applied entropy-based behavioral modeling to predict depression from smartphone screen-time data [9] and variability in locations visited [10], the latter also being applied to predict the emergence of symptoms in schizophrenia patients [11]. Furthermore, a study on patients with severe mental illnesses such as schizophrenia, major depressive disorder, and bipolar 1 disorder, used an entropy-based anomaly detection system to detect behavioral shifts, and found they were associated with medication non-adherence [8].

There is currently a gap in the literature regarding the application of entropy-based behavioral shift detection methods to conversational data from diabetes patients. This study seeks to address this by presenting a system that identifies behavioral changes through entropy-based anomaly detection and evaluating its performance using synthetically generated conversations.

The research question for this study is: "How effective is entropy-based behavioral modeling in detecting anomalies that indicate non-adherence among users of a chatbot-based diabetes lifestyle management support system?"

Early detection of behavioral shifts, correlated with non-adherence, could enable timely intervention from diabetes lifestyle management support systems. Ultimately, this approach could be integrated into platforms like CHIP to enhance patient monitoring and support more proactive care.

The structure of this study is as follows: Chapter 2 outlines the methodology, including a summary of the system design and the rationale behind key choices. Chapter 3 presents the experimental results, which are then interpreted and discussed in detail in Chapter 4. Chapter 5 summarizes the main findings, addresses the study's limitations, and outlines recommendations for future research. Chapter 6 reflects on the ethical considerations of the study and provides instructions to support the reproducibility of the experiment.

2 Methodology

2.1 Entropy

To quantify behavioral variability in this study, we used Shannon Entropy [12]. Entropy is a measure of uncertainty in a probability distribution, and in this context, it reflects how predictable or irregular a user's behavior is within a given time window.

Given a discrete distribution $P = \{p_1, p_2, \ldots, p_n\}$, representing the relative frequency of observed behavioral states (e.g., time bins or log types), entropy is calculated as

$$H(P) = -\sum_{i=1}^{n} p_i \log_2 p_i$$

For example, consider a feature such as the time of day when a patient logs their glucose readings, divided into discrete bins such as morning, afternoon, and evening. Here, each x_i corresponds to a time bin (e.g., morning), and p_i is the proportion of logs recorded in that bin during the time window.

If a user almost always logs in the morning, then $p_{morning} \approx 1$, and $p_{afternoon} \approx p_{evening} \approx 0$, therefore the entropy will be approximately 0, reflecting predictable behavior. Conversely, if logging times are spread evenly across morning, afternoon, and evening, the p_i values will all be roughly equal to $\frac{1}{3}$, resulting in a higher entropy of about 1.6. This formula yields higher values when behavior is more varied and lower values when it is more repetitive or focused.

By tracking entropy over time across multiple features, we aim to identify shifts that signal behavioral irregularities relevant to treatment adherence. For example, a user with consistent logging at the same time each day has highly predictable behavior and therefore should have low baseline entropy. If the logging of this patient suddenly became erratic, the entropy would increase greatly above baseline, resulting in the latest log being marked as an anomaly, prompting an intervention by the diabetes lifestyle management support system.

Note that each probability p_i is a real number in the interval $[0, 1]$. For all $p_i > 0$, the logarithm $\log_2 p_i$ is non-positive, which means each term $-p_i \log_2 p_i$ in the entropy sum is non-negative. Additionally, when $p_i = 0$, the expression $p_i \log_2 p_i$ is conventionally defined as 0 in the context of information theory, and is therefore omitted from the sum.

2.2 System design

Before diving into each component in detail, we briefly summarize the system's overall architecture. The system is organized into modular components to support the generation, analysis, and visualization of synthetic user interaction logs for behavioral research. It consists of four main modules: data generation, feature extraction, entropy calculation, and anomaly detection.

While these components are intended for eventual integration with the CHIP system, the current implementation is developed as a standalone prototype to enable faster development, easier testing, and flexible experimentation.

2.3 Synthetic data generation

To evaluate the proposed method in a pilot experiment, we generated synthetic patient conversations consisting of status updates about glucose, diet, mood, activity, insulin, other medication, sleep, and weight using a large language model. These were stored in a structured JSON format to support easy integration into other system modules. An example of a prompt and its output can be found in Appendix A.

Due to time and resource constraints, conducting a longitudinal user study with sufficient participants and responses to reliably capture behavioral patterns was not feasible. Additionally, access to real diabetes patient-chatbot conversations was unavailable, leading to synthetic data being used instead. Synthetic data also avoids privacy concerns associated with real patient logs and allows for controlled insertion of behavioral anomalies, which is essential for validating detection performance. From these conversations, we extracted a set of features correlated with engagement and treatment adherence, described in more detail below.

Conversations were generated using a locally hosted large language model, specifically the Mistral-7B-Instruct-v0.2.Q4_K_M model [13]. This model was chosen for its fast inference, strong performance on conversational tasks, and low resource demands, with weights quantized to 4 bits. The model was run locally using llama.cpp [14], an open-source, high-performance C/C++ implementation for running large language models on consumer hardware, together with the llama-cpp-python binding [15] that allows llama.cpp models to be called from Python code. This also facilitates future integration into the CHIP system, implemented in Python.

2.4 User persona traits

To support controlled testing of different user behavior patterns, we define a set of simulated user personas characterized by three key behavioral traits, each framed as a binary option. These traits were deliberately chosen to manipulate the behavioral features described below in Section 2.6, allowing us to test whether the prototype can differentiate between distinct engagement styles.

In a real-world study, such predefined personas would not be necessary, as patients naturally exhibit a wide range of behavioral variation. The matrix only provides a structured way to simulate diverse user profiles during early-stage evaluation.

Each simulated user is defined by selecting one value from each of the three behavioral traits. All possible combinations are used, resulting in eight distinct personas. These users are non-anomalous by design: they follow consistent patterns that match their assigned traits, and no artificial behavioral anomalies are introduced. As such, the system should not flag any of their behavior as anomalous, making them useful for validating that the anomaly detection logic does not produce false positives in response to expected behavioral variability.

The traits and their corresponding options are:

- **Timing Consistency (Consistent – Erratic):**
  Reflects temporal regularity in user interaction patterns. Consistent users engage at steady times, while erratic users exhibit irregular, unpredictable timing.

- **Interaction Frequency (Frequent – Infrequent):** Indicates overall interaction volume. Frequent users interact often; infrequent users engage rarely.
- **Content Diversity (Varied – Similar):** Captures the diversity of content produced. Varied users generate a wide range of topics or types of messages; similar users tend to repeat the same content.

By combining these traits, we define a comprehensive trait matrix, allowing us to represent and analyze nuanced user behavior patterns. Each combination corresponds to a distinct persona reflecting a unique engagement style.

Table 1 lists all trait combinations and their corresponding persona identifiers, which are used in the graphs below.

Table 1: User personas defined by the trait matrix combining Consistency, Frequency, and Content Diversity.

ID	Consistency	Frequency	Diversity
CFV	Consistent	Frequent	Varied
CFS	Consistent	Frequent	Similar
CIV	Consistent	Infrequent	Varied
CIS	Consistent	Infrequent	Similar
EFV	Erratic	Frequent	Varied
EFS	Erratic	Frequent	Similar
EIV	Erratic	Infrequent	Varied
EIS	Erratic	Infrequent	Similar

2.5 Anomaly injection

To evaluate the system's ability to detect irregular behavioral patterns, we introduced a set of users with controlled behavioral anomalies. In addition to the eight baseline personas described earlier, five additional transitional personas were generated. These users maintain stable behavior until day 15, after which they undergo a predefined behavioral shift that continues through the remainder of the simulation at day 30.

These anomalies were not random but were deliberately crafted to reflect meaningful transitions in user engagement, simulating realistic disruptions or improvements in adherence. Each transitional persona begins with a behavioral pattern and then switches to a contrasting one, either indicating a breakdown, improvement, or fluctuation in routine.

The five transitional personas are defined as follows:

- **Adherence Breakdown**: Switches from Consistent, Frequent, Varied to Erratic, Infrequent, Similar, simulating a decline in adherence.
- **Gradual Improvement**: Switches from Erratic, Infrequent, Similar to Consistent, Frequent, Varied, modeling improved engagement over time.
- **Selective Adherence**: Switches from Consistent, Frequent, Varied to Consistent, Infrequent, Similar, reflecting reduced breadth and frequency of interaction while retaining temporal regularity.
- **Erratic Behavior**: Switches from Consistent, Infrequent, Similar to Erratic, Frequent, Varied, simulating an abrupt shift to chaotic behavior.
- **Minimal-To-Detailed**: Switches from Erratic, Infrequent, Similar to Consistent, Frequent, Varied, modeling a user who becomes more structured and engaged over time.

Each of these transitions was implemented by modifying one or more of the user's behavioral traits (i.e., logging consistency, interaction frequency, and content variety) after the midpoint of the simulation. This design enables a structured test of whether the entropy-based anomaly detection method can distinguish between stable and transitional users.

The anomalies were selected to produce measurable effects in entropy across multiple features, such as time of day, frequency, and semantic diversity. By comparing these transitional users with the non-anomalous personas, we aim to assess the sensitivity and specificity of the detection pipeline.

Due to time constraints, only these five transitional users were included in this study. Nevertheless, this approach provides a concrete foundation for evaluating the detection of temporal behavioral changes, and it opens the door for more extensive simulation-based evaluation in future work.

2.6 Feature modeling

To analyze patient conversations, we selected behavioral features that can be reliably extracted from patient-chatbot interactions. These features have also shown relevance in digital health monitoring, as supported by prior research described later in this section.

Entropy requires a probability distribution over discrete, mutually exclusive categories. Therefore, continuous or diverse inputs from user logs must be converted into such categories through a process called discretization. Discretization involves grouping data into predefined bins or categories, for example, dividing times of day into morning, afternoon, evening, and night. This transformation enables us to represent the features as frequency distributions of observed behaviors within a given time window, which serve as the basis for entropy calculation.

The choice of bin size is critical: having fewer, meaningful bins is generally preferable to using very fine-grained bins (e.g., one per minute), which can lead to sparse data, high-variance probability estimates, and increased noise.

Selecting appropriate bin numbers involves balancing interpretability with statistical robustness and depends on the specific use case. In our current system, this process is mainly based on domain knowledge and informed assumptions, such as natural divisions like times of day or response categories. These assumptions are consistent with those used in our synthetic data generation, maintaining methodological alignment. However, because we are responsible for both generating the synthetic data and designing the system, this interplay may confound the results and limit the generalizability of our findings. We anticipate that more thorough tuning and validation will be required before a fully operational system can be deployed.

Future work could incorporate data-driven methods such as equal-frequency binning or clustering, the latter having been applied in user profiling to identify customer personas [16]. Our current system is a proof of concept and does not yet implement these data-driven or iterative refinement techniques.

However, the initial design allows for future fine-tuning as additional data and insights become available. Validating bin stability through entropy estimates and anomaly detection, alongside iterative refinement, will be important next steps to ensure bins effectively capture meaningful behavioral variation without sacrificing reliability.

Table 2: Selected behavioral features and their controlling persona traits

Feature	Controlling Persona Trait
Time of day	Consistency
Logging frequency	Frequency
Log type	Diversity
Semantic similarity	Diversity

Table 2 illustrates the behavioral features selected for analysis and the persona traits designed to control them. Below, we describe these features and the discretization schemes applied in this study.

Time of day

This feature captures the time at which each log is made, providing insight into the patient's daily routine. To enable entropy calculation, log times are discretized into four time-of-day bins: morning from 5:00 to 11:59, afternoon from 12:00 to 16:59, evening from 17:00 to 21:59, and night from 22:00 to 4:59. These bins were selected as exploratory, experimental values intended to correspond roughly to natural and commonly recognized periods of daily activity, reflecting typical human circadian rhythms and social behaviors (e.g., waking hours, work periods, evening leisure, and nighttime rest). However, these bin ranges remain tentative and require further adjustment and validation in future work to confirm their effectiveness in capturing meaningful behavioral variation in this context.

Consistent logging within a specific time window, such as always in the morning, results in low entropy, indicating predictable behavior. If the patient logs at varying times, entropy increases, signaling a potential behavioral shift. While our measure focuses on temporal patterns of logging behavior rather than physical movement, prior research using mobile phone sensor data has shown that variations in the entropy of circadian movement, defined as regularity in daily location patterns, and daily activity rhythms are strongly associated with depressive symptom severity [10]. This suggests that temporal behavioral features may be informative for patient monitoring, though further validation in the context of patient-chatbot interactions is needed.

Logging frequency

Logging frequency can be binned into states based on the number of logs per day (chosen here as 0–3, 4–6, 7–10, and 10 or more) to capture meaningful differences between low, moderate, and high levels of engagement. These bins were selected to provide a balance between granularity and interpretability, while reflecting natural groupings of user activity observed in preliminary data. In actual patient data, this parameter is expected to be highly volatile; therefore, adapting the binning strategy to better reflect personalized patterns is an important goal for future work.

To the best of our knowledge, entropy-based modeling of logging frequency in the context of patient-chatbot interactions has not been studied. However, prior work has successfully used entropy to model variability in medication frequency as an indicator of non-adherence among patients with severe mental illness [8]. This suggests a promising direction for our approach.

Log type

Each log message is categorized into one of several predefined types: glucose, diet, mood, activity, insulin, medication, sleep, weight, notes, or other. These categories were selected to represent important areas of diabetes management and to capture varying degrees of diversity in what users report during their interactions, following their persona trait instructions. Entropy calculated over these categories reflects how broadly a user engages with different aspects of diabetes management. High entropy suggests a varied and well-rounded logging pattern, whereas low entropy may indicate a narrow or repetitive focus.

Although not specific to diabetes, entropy has been successfully applied to analyze food category distributions in personal diets, helping to model adherence to dietary guidelines [17]. This suggests that this type of categorical diversity may offer insight into behavioral variability.

Semantic similarity

This feature measures how similar each log message is to the one immediately before it, capturing the degree of repetition or variation in a user's reporting style. High semantic similarity between consecutive messages may indicate routine, formulaic, or less engaged reporting, while lower similarity suggests more varied, expressive, or thoughtful input. To compute semantic similarity, each message is first converted into a vector based on the frequency of each word it contains (a bag-of-words representation).

The similarity between two consecutive messages is then quantified using cosine similarity, a standard metric in natural language processing. Cosine similarity calculates the cosine of the angle between two word-frequency vectors, resulting in a score between 0 and 1: a score of 1 means the messages are nearly identical in word usage, while a score of 0 means they share no common words. For analysis, these similarity scores are grouped into the following categories:

- **Identical** (0.7–1.0): Messages are highly similar or nearly the same.
- **Similar** (0.4–0.7): Messages share substantial overlap but are not identical.
- **Moderately different** (0.2–0.4): Messages have moderate differences in content.
- **Different** (0.0–0.2): Messages are largely distinct in their wording and content.

A user whose messages fall mostly into the "identical" or "similar" categories may be providing repetitive or habitual responses, possibly indicating lower engagement.

In contrast, a broader spread across categories, especially with more "somewhat different" and "different" scores, suggests the user is varying their reports and potentially putting more thought and effort into their entries.

2.7 Entropy calculation

To quantify the predictability and stability of patient behavior over time, we computed Shannon entropy over sliding windows of 10 days. Within each window, we considered only the data points that occurred during those 10 days to estimate the probability distribution of the behavioral categories for that specific synthetic user.

For example, if a 10-day window contained 30 log messages, and 27 were sent in the morning while 3 were sent in the evening, the empirical probabilities would have values $p_{morning} = 0.9; p_{evening} = 0.1$. Categories not observed in that window (e.g., afternoon, night) would have a probability of zero. These probabilities define a discrete distribution, which is then used to compute the Shannon entropy as described in Section 2.1. The resulting value quantifies the behavioral regularity within that time period: lower entropy indicates more predictable patterns, while higher entropy reflects greater variability.

The choice of a 10-day window is inspired by related work in behavioral anomaly detection using entropy [8], where a 10-day observation window was used to capture temporal patterns in patient behavior. However, this choice is somewhat arbitrary and may not be optimal for all contexts. Future work should explore different window lengths and validate their effectiveness in capturing meaningful behavioral variations in patient-chatbot interactions.

2.8 Anomaly detection

To identify unusual changes in user behavior, entropy values computed over sliding 10-day windows were analyzed using a moving average anomaly detection method. For each entropy value starting after an initial set of observations, the mean and standard deviation of the previous five entropy values were calculated. An entropy value was flagged as an anomaly if it deviated from this moving average by more than two standard deviations. Therefore, for each window, the system returns a binary result that marks the window as either normal or anomalous.

This approach allows for adaptive detection of significant deviations relative to recent behavioral patterns, capturing abrupt increases or decreases in entropy that may indicate shifts in adherence or engagement. The output consists of flagged anomaly points indicating potential behavioral changes that warrant further investigation.

While this method provides a simple and interpretable baseline, more sophisticated, personalized anomaly detection techniques tailored to individual user patterns could substantially enhance detection accuracy. Incorporating user-specific models or machine learning approaches represents a promising direction for future improvement.

3 Results

This section presents the findings from evaluating the entropy-based anomaly detection system using synthetic patient-chatbot conversations across 13 distinct user personas. The analysis covers behavioral pattern differentiation, entropy modeling effectiveness, anomaly detection performance, and system behavior characteristics.

3.1 Behavioral pattern analysis

Baseline entropy characterization

The entropy-based system produced distinct baseline entropy values across different user personas. Table 3 shows the baseline entropy distributions across all 13 personas for all the features.

Taking as an example the "time of day" feature, consistent personas (CFV, CFS, CIV, CIS) exhibited baseline entropy values around 1.9, while erratic personas (EFV, EFS, EIV, EIS) showed baseline entropy closer to approximately 1.5. Persona CIV was an outlier in the consistent group, having a mean entropy of only around 0.9.

Figures 1, 2, and 3 present entropy values over time for the "time of day" feature across three simulated users. Each plot covers 21 sliding windows, representing 30 days of messages with a window size of 10.

Figure 3 depicts a transitional user, with the anomalous period following day 15 highlighted in orange to indicate the windows that could catch the behavioral shift. The anomaly detection system flagged 4 windows in Figure 1, 3 in Figure 2, and 4 in Figure 3, marked by red indicators on each timeline. Out of these, only three windows, the first 3 in the third graph, were true anomalies, with the rest being false positives.

Table 4 shows the relevant metrics for anomaly detection. Note that recall and F1 score are undefined for datasets with no anomalies, and are therefore left blank.

From the table, we can also extract the false positive rate as $FPR = \frac{FP}{FP+TN} \approx 15\%$, which is relatively high for an anomaly detection system.

4 Discussion

This study demonstrates the feasibility of using entropy-based behavioral modeling to detect anomalies that may indicate non-adherence among users of chatbot-based diabetes lifestyle management support systems. The proposed system addresses a gap in digital health monitoring by providing an automated approach to identify behavioral shifts that correlate with treatment disengagement.

The entropy-based anomaly detection system achieved approximately 76% accuracy and 35% recall in detecting behavioral irregularities in synthetic patient-chatbot conversations. These results demonstrate that on the test data, Shannon entropy can effectively quantify behavioral variability across multiple dimensions of patient engagement, including temporal patterns, interaction frequency, content diversity, and semantic similarity of communications.

The system's ability to establish individualized behavioral baselines proves particularly valuable, as it adapts to personal habits rather than relying on predefined universal patterns. This personalized approach addresses the inherent variability in patient behaviors and preferences, making the system more suitable for diverse patient populations. The modular architecture of the system, consisting of data generation,

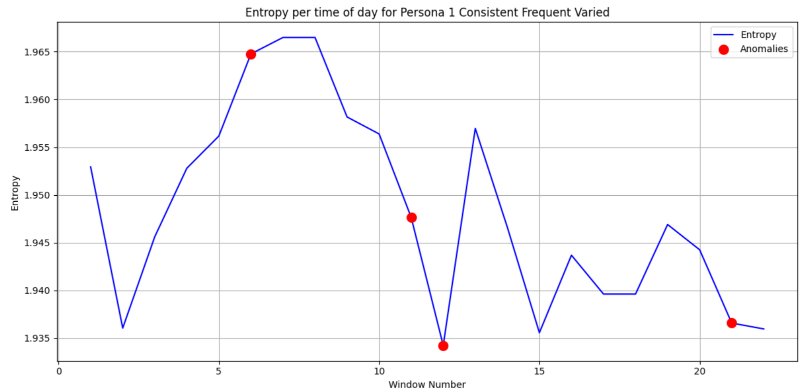

Figure 1: Entropy evolution graph calculated by the system for conversations of the synthetic user with traits Consistent, Frequent, Varied. Anomalies are marked by red dots.

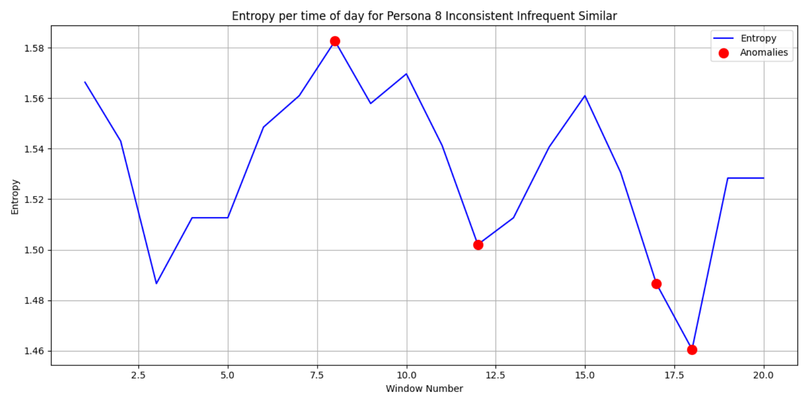

Figure 2: Entropy evolution graph calculated by the system for conversations of the synthetic user with traits Erratic, Infrequent, Similar. Anomalies are marked by red dots.

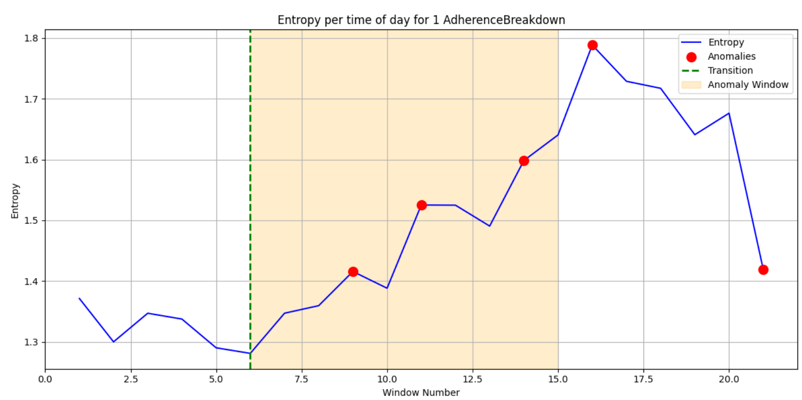

Figure 3: Entropy evolution graph calculated by the system for conversations of the synthetic user with transitional profile ”Adherence Breakdown” (shifting from traits Consistent, Frequent, Varied to Erratic, Infrequent, Similar). Anomalies are marked by red dots. Anomalous windows (windows that could catch the injected behavioral shift) are highlighted in orange.

Table 3: Baseline entropy values (mean ± standard deviation) by persona and feature

Persona	Time of Day	Log Frequency	Log Type	Semantic Similarity
CFV	1.95 ± 0.01	0.98 ± 0.01	3.01 ± 0.03	0.82 ± 0.05
CFS	1.98 ± 0.01	0.89 ± 0.09	2.99 ± 0.11	0.89 ± 0.14
CIV	0.88 ± 0.00	0.00 ± 0.00	2.62 ± 0.17	0.72 ± 0.00
CIS	1.91 ± 0.01	0.47 ± 0.00	2.59 ± 0.07	0.42 ± 0.10
IFV	1.47 ± 0.06	0.97 ± 0.05	3.06 ± 0.08	1.11 ± 0.16
IFS	1.55 ± 0.02	0.99 ± 0.01	2.87 ± 0.08	1.14 ± 0.05
IIV	1.52 ± 0.06	0.00 ± 0.00	2.79 ± 0.15	0.69 ± 0.09
IIS	1.52 ± 0.03	0.67 ± 0.10	2.69 ± 0.07	0.81 ± 0.09
(Transitional) Adherence Breakdown (CFV to EIS)	1.33 ± 0.03	0.85 ± 0.06	3.10 ± 0.01	0.64 ± 0.08
(Transitional) Gradual Improvement (EIS to CFV)	1.61 ± 0.06	0.00 ± 0.00	2.87 ± 0.02	0.88 ± 0.01
(Transitional) Selective Adherence (CFV to CIS)	1.41 ± 0.04	0.77 ± 0.16	3.10 ± 0.03	0.69 ± 0.07
(Transitional) Erratic Behavior (CIS to EFV)	1.39 ± 0.03	0.00 ± 0.00	2.67 ± 0.14	0.72 ± 0.08
(Transitional) MinimalToDetailed (EIS to CFV)	1.59 ± 0.10	0.00 ± 0.00	2.54 ± 0.14	0.65 ± 0.02

Table 4: Anomaly detection metrics (true positives, false positives, false negatives, true negatives, accuracy, precision, recall, F1 score) for all personas and features

Persona	TP (total)	FP (total)	FN (total)	TN (total)	Acc (total)	Prec (total)	Rec (total)	F1 (total)
CFV	0	13	0	74	0.85	0.00	0.00	
CFS	0	15	0	69	0.82	0.00	0.00	
CIV	0	6	0	77	0.93	0.00	0.00	
CIS	0	11	0	73	0.87	0.00	0.00	
IFV	0	10	0	78	0.89	0.00	0.00	
IFS	0	19	0	69	0.78	0.00	0.00	
IIV	0	7	0	63	0.90	0.00	0.00	
IIS	0	13	0	68	0.84	0.00	0.00	
Adherence Breakdown	12	9	28	35	0.56	0.57	0.30	0.39
Gradual Improvement	17	5	23	39	0.67	0.77	0.42	0.55
Selective Adherence	19	6	21	38	0.68	0.76	0.47	0.58
Erratic Behavior	10	10	30	34	0.52	0.50	0.25	0.33
Minimal To Detailed	13	8	27	36	0.58	0.62	0.33	0.43
Total	71	132	129	753	0.76	0.35	0.35	0.35

feature extraction, entropy calculation, and anomaly detection components, provides a flexible framework that can be adapted for integration with existing digital health platforms like CHIP.

The findings suggest that automated behavioral monitoring through entropy analysis could enhance patient care through:

- **Early detection:** Identifying potential non-adherence before clinical outcomes deteriorate
- **Personalized monitoring:** Adapting to individual patient patterns rather than using one-size-fits-all approaches
- **Timely interventions:** Enabling healthcare providers to offer targeted support when behavioral changes are detected
- **Reduced clinical burden:** Automating the monitoring process to free up healthcare resources for direct patient care

4.1 Limitations

While the system demonstrated promising accuracy, the high false positive rate of approximately 15% indicates room for improvement. This over-flagging primarily occurred in users with naturally high behavioral variability, suggesting that more sophisticated personalization techniques are needed to distinguish between genuine anomalies and normal individual variation.

Another key limitation lies in the discretization process. The bin sizes used to convert continuous behavioral data into categorical values—such as fixed divisions of the day into morning, afternoon, evening, and night—were selected heuristically and not empirically validated. These manually defined categories may not optimally reflect the nuances of individual user behavior, potentially contributing to inaccurate entropy calculations and misclassification of normal patterns as anomalies. Future work should explore data-driven or adaptive binning strategies that better capture the structure of user-specific behavior.

Additionally, the system was evaluated entirely on synthetic data. While synthetic conversations allow for controlled insertion of behavioral anomalies and eliminate privacy concerns, they inherently lack the complexity, noise, and variability found in real-world clinical interactions. This limits the generalizability of the findings and may lead to an overestimation of system performance in practical deployments. Nevertheless, the use of synthetic data enabled systematic testing and provided a valuable first step in evaluating the feasibility of entropy-based anomaly detection in this domain.

5 Future Work

5.1 Integration with CHIP

Integrating the entropy-based anomaly detection system into the CHIP conversational support platform could enhance its ability to identify early signs of patient disengagement or non-adherence. This would enable more timely and proactive interventions by healthcare providers.

The system's modular Python-based architecture supports seamless integration with CHIP's infrastructure. Its sliding-window analysis enables near real-time monitoring, while the entropy-based metrics provide interpretable signals clinicians can act on. By continuously tracking entropy across behavioral features, the system can help maintain engagement and support responsive diabetes care.

5.2 Advanced Anomaly Detection Algorithms

The current threshold-based method serves as a proof of concept but lacks adaptability. Future work should explore more advanced techniques to improve accuracy and personalization.

Dynamic Binning. Static bin sizes may not capture individual behavioral variation. Adaptive methods like equal-frequency binning and clustering [16] could better reflect natural behavioral groupings. Cross-validation and performance metrics could guide optimal bin selection.

Personalized Thresholds. Global thresholds overlook individual baselines. User-specific thresholds informed by historical data, performance feedback, and clinical context could reduce false positives.

Machine Learning Approaches. Algorithms such as Isolation Forests offer robust alternatives, identifying anomalies based on data separability. Isolation Forest is a widely used anomaly detection technique that functions by recursively partitioning the dataset using random splits. The underlying intuition is that anomalies are easier to isolate and therefore require fewer splits on average during the partitioning process. Other candidates include One-Class SVMs, autoencoders, and ensemble models.

Feature Expansion. Incorporating linguistic sentiment, interaction timing, or other patterns could improve detection. Temporal and emotional patterns may offer additional behavioral insight.

Feature Selection. As features grow, dimensionality reduction methods like PCA and correlation analysis will be key to maintaining model efficiency and interpretability.

5.3 Clinical Validation

Real-world validation is essential to assess the system's practical utility in healthcare settings.

Data Collection. Future studies should involve longitudinal interaction data from actual diabetes patients using platforms like CHIP. Collaborations with healthcare providers and adherence to ethical data governance will be crucial for responsible data use.

Outcome Correlation. Detected anomalies should be evaluated against real adherence lapses, clinical metrics (e.g., HbA1c, hospital visits), and patient outcomes. Additionally, the effectiveness of system-triggered interventions should be measured to assess clinical relevance.

5.4 Final Remarks

This study presents an initial implementation of entropy-based behavioral anomaly detection using synthetic diabetes patient data. While useful for controlled testing, clinical validation is necessary to assess real-world effectiveness.

The methods proposed—particularly sliding-window entropy and simple anomaly flagging—offer a baseline for integrating behavioral monitoring into digital care platforms. With further refinement and real-world testing, such systems could support earlier detection of disengagement and improve chronic disease management through more adaptive and personalized care.

6 Acknowledgements

I would like to express my sincere gratitude to Prof. C.M. Jonker, J.D. Top, MSc., Dr. H.A. de Weerd, and Prof. L.C. Verbrugge for their guidance and support throughout this project. Their insights, feedback, and mentorship were essential in shaping the direction of this research. This work would not have been possible without their continued help and expertise.

7 Responsible Research

7.1 Ethical concerns

While the ability to detect behavioral anomalies can support timely interventions and improve patient outcomes, it also introduces ethical risks if misused. In particular, identifying moments of vulnerability—such as disengagement, stress, or emotional instability—could be exploited to manipulate behavior rather than support it. For instance, commercial actors might use such insights to target individuals with persuasive messaging when they are most susceptible, or insurers could discriminate based on inferred patterns of non-adherence. These risks underscore the importance of implementing strong safeguards, including transparency about data use, strict access controls, and clear boundaries that prevent behavioral data from being repurposed for non-clinical or coercive ends.

7.2 EU AI Act

Under the European Union's AI Act, systems that monitor health-related behaviors—particularly those used in medical contexts—may fall under the category of high-risk AI systems. The anomaly detection system proposed in this project, when integrated into a platform like CHIP and used to influence clinical decision-making or trigger interventions, would likely be subject to these high-risk requirements. This includes obligations around transparency, robustness, human oversight, and risk management. Although the current implementation uses synthetic data and is not deployed in a clinical setting, future work involving real patient data and integration into healthcare workflows must comply with the AI Act's provisions. This entails conducting impact assessments, ensuring data quality and security, documenting model behavior, and maintaining human control over final decisions. Proactively aligning the system with the AI Act will be essential to ensure legal compliance and to uphold patient trust in data-driven healthcare tools.

7.3 Reproducibility

Due to constraints related to time and available resources, this study utilizes a large language model (LLM) to generate synthetic patient-chatbot conversations instead of collecting real clinical data. While synthetic data enables the exploration and validation of analytic methods in a controlled setting, it does not fully capture the complexity and variability of real-world patient interactions. Therefore, findings based on these simulated conversations should be interpreted with caution and validated with clinical studies in the future to ensure generalizability and clinical relevance. This approach reflects a pragmatic balance between rigorous research and ethical considerations during early-stage development.

In order to ensure the reproducibility of this study, the codebase and prompts used for the project can be found inside a code repository that is publicly accessible [18].

In order to ensure the synthetic conversations generated of the large language model are reproducible, we set the temperature value of the local model to 0, and the model seed number was fixed to "2025".

A Data generation prompt example

A.1 LLM prompt

As an example of the prompts given to the locally-hosted large language models, the following is the full prompt for generating the conversations for the user persona with traits Consistent, Frequent, Varied. This was generated programatically by combining multiple parts, each corresponding to a different trait or feature:

"You are a diabetic patient using a chatbot to log health data.

Your behavioral characteristics:

- Maintains strict adherence to logging schedule
- Uses precise timestamps and regular intervals
- Shows high attention to timing details
- Rarely deviates from established patterns
- Logs multiple times throughout the day
- Provides frequent updates on status
- Maintains detailed daily records
- Shows high engagement with logging
- Uses diverse log types
- Provides varied and detailed content
- Shows interest in multiple aspects of health
- Maintains comprehensive records

Based on these characteristics, you should:

- Adjust your language and tone accordingly
- Match the expected logging frequency
- Use appropriate level of detail
- Maintain consistency with your trait patterns

It is 08:00 (morning), and you want to log a glucose update.

Guidelines for glucose logging:

Report your current blood glucose level:

- Include the exact value in mg/dL
- Specify if it's fasting, pre-meal, post-meal (how many hours after), or random
- Mention any factors that might have affected the reading
- Note if the reading is within your target range

Example: 'My fasting glucose this morning was 95 mg/dL, which is within my target range.'

Generate a natural, conversational message as if you're talking to your healthcare provider. Be specific and include all relevant details. Keep your message concise but informative. Use appropriate medical terminology where relevant.

Your message:"

A.2 LLM response

The following are three example messages generated by the large language model based on the previously stated prompt:

"Good morning, doc! It's 11:32 AM, and I've just finished my breakfast. I had a bowl of oatmeal with 1 banana, 1 tbsp of honey, and a handful of almonds. I estimate the total carbohydrate intake to be around 75g. I'm sticking to my meal plan as usual.",

"Hi there, it's 12:04 PM, and I just took my metformin 1000mg with my lunch. No side effects to report today. I've been maintaining my schedule consistently, and my blood sugar levels have been stable. I'll keep you updated on any changes. Let me know if you need any additional information."

"Hi there, just wanted to log my sleep from last night. I managed to get a full 8 hours of sleep, but the quality wasn't great, I'd rate it a 5/10. I woke up twice during the night, but overall I feel okay this morning. My morning glucose reading was 135 mg/dL."

References

[1] H. Sun, P. Saeedi, S. Karuranga, M. Pinkepank, K. Ogurtsova, B. B. Duncan, C. Stein, A. Basit, J. C. Chan, J. C. Mbanya, M. E. Pavkov, A. Ramachandaran, S. H. Wild, S. James, W. H. Herman, P. Zhang, C. Bommer, S. Kuo, E. J. Boyko, and D. J. Magliano, "IDF diabetes atlas: Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045," *Diabetes Research and Clinical Practice*, vol. 183, 2022.

[2] M. B. Gomes, L. R. M. Tannus, R. A. Cobas, A. S. M. Matheus, P. Dualib, A. T. Zucatti, C. Cani, A. D. Guedes, F. M. Santos, J. Sepulveda, M. Tolentino, M. C. Façanha, A. C. R. A. Faria, S. Lavigne, A. P. Montenegro, M. Rodacki, M. de Fatima Guedes, R. Szundy, M. M. Cordeiro, P. T. S. Santos, C. A. Negrato, and BrazDiab1SG, "Determinants of self-monitoring of

blood glucose in patients with type 1 diabetes: A multicentre study in Brazil," *Diabetic Medicine*, vol. 30, no. 10, pp. 1255–1262, 2013.

[3] M. Blackwell and B. Wheeler, "Clinical review: The misreporting of logbook, download, and verbal self-measured blood glucose in adults and children with type I diabetes," *Acta Diabetologica*, vol. 54, pp. 1–8, 2017.

[4] 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.

[5] A. Karymsakov, F. Foster, L. Toleubekova, N. Yegembayeva, A. Gaipov, and M. Aljofan, "The nonadherence to prescriptions among type 2 diabetes patients, and its determining factors," *Electronic Journal of General Medicine*, vol. 21, no. 1, p. em565, 2024.

[6] 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.

[7] B. J. W. Dudzik, J. S. van der Waa, P. Y. Chen, R. Dobbe, 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," *Journal of Artificial Intelligence Research*, vol. 80, pp. 919–929, 2024.

[8] J. Knights, Z. Heidary, and J. M. Cochran, "Detection of behavioral anomalies in medication adherence patterns among patients with serious mental illness engaged with a digital medicine system," *Journal of Medical Internet Research Mental Health*, vol. 7, no. 9, 2020.

[9] K. Opoku Asare, Y. Terhorst, J. Vega, E. Peltonen, E. Lagerspetz, and D. Ferreira, "Predicting depression from smartphone behavioral markers using machine learning methods, hyperparameter optimization, and feature importance analysis: Exploratory study," *Journal of Medical Internet Research Mhealth and Uhealth*, vol. 9, no. 7, 2021.

[10] S. Saeb, M. Zhang, C. J. Karr, S. M. Schueller, M. E. Corden, K. P. Kording, and D. C. Mohr, "Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: An exploratory study," *Journal of Medical Internet Research*, vol. 17, no. 7, 2015.

[11] R. Wang, E. A. Scherer, M. Walsh, W. Wang, M. H. Aung, D. Ben-Zeev, R. Brian, A. T. Campbell, T. Choudhury, M. Hauser, and J. Kane, "Predicting symptom trajectories of schizophrenia using mobile sensing," *GetMobile: Mobile Computing and Communications*, vol. 22, no. 2, pp. 32–37, 2018.

[12] C. E. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948.

[13] Mistral AI, "Mistral 7B Instruct v0.2." https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2, 2023. Accessed May 29 2025.

[14] G. Gerganov, "llama.cpp." https://github.com/ggerganov/llama.cpp, 2023. Accessed May 29 2025.

[15] G. Gerganov and contributors, "llamacpp-python." https://github.com/abetlen/llama-cpp-python, 2023. Accessed May 29 2025.

[16] N. Tu, X. Dong, P.-L. P. Rau, and T. Zhang, "Using cluster analysis in persona development," in *2010 8th International Conference on Supply Chain Management and Information*, pp. 1–5, 2010.

[17] C. Huffman, A. G. Ortega-Avila, and H. Nájera, "An entropy-based approach to measuring diet adherence," *Entropy*, vol. 25, no. 9, 2023.

[18] S. Ciuntu, "Entropy-Based Modeling For Detecting Behavioral Anomalies in Users of a Diabetes Lifestyle Management Support System." https://github.com/sorin-a-c/diabetes-anomaly-detection, 2025. Accessed: 2025-06-22.