# Poster: Leveraging Large Language Models for Zero-Shot Detection and Mitigation of Data Poisoning in Wearable AI Systems

**W.K.M Mithsara**
School of Computing
Southern Illinois University
Carbondale, IL, 62901
malithi.mithsara@siu.edu

**Abdur R. Shahid**
School of Computing
Southern Illinois University
Carbondale, IL, 62901
shahid@cs.siu.edu

**Ning Yang**
School of Computing
Southern Illinois University
Carbondale, IL, 62901
nyang@siu.edu

## Abstract

Wearable AI systems, particularly in Human Activity Recognition (HAR), are becoming integral to applications in healthcare, security, and personal fitness due to the widespread adoption of smart devices and wearable technologies. However, the increasing reliance on machine learning models in HAR introduces significant risks, especially from poisoning attacks that compromise system reliability and data integrity. This paper explores the potential of Large Language Models (LLMs) to detect and sanitize poisoning attacks in wearable AI systems. Building on ongoing research into integrating LLMs within cyber-physical systems, we focus on sensor-based interactions with the physical world. Our case study seeks to answer the following question: How effective are LLMs in detecting and sanitizing poisoning attacks on human activity sensor data? Through zero-shot learning, we evaluate the performance of models such as ChatGPT 3.5, ChatGPT 4, and Gemini, providing insights into the viability of LLMs for real-time defense and data integrity in wearable AI systems.

## 1 Introduction

Wearable Artificial Intelligence (AI) systems, particularly in the domain of Human Activity Recognition (HAR), have become increasingly important in applications such as fitness tracking, healthcare monitoring, and smart environments. HAR systems process time-series data collected from various wearable sensors and convert this data into recognizable human activities (e.g., walking, sitting, standing, jogging). These systems rely on sensors like Inertial Measurement Units (IMUs), gyroscopes, accelerometers, and magnetometers embedded in devices such as smartphones and fitness trackers[14].

Despite their wide-ranging utility, HAR systems face significant vulnerabilities, particularly from adversarial threats like data poisoning attacks. These attacks involve injecting malicious data into the system, compromising the system's reliability and accuracy. For instance, Microsoft's chatbot Tay was compromised by a data poisoning attack, where adversaries manipulated the system's inputs to

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

generate harmful outputs [14]. Such attacks can be particularly detrimental in critical applications, making the detection and mitigation of poisoning attacks a crucial challenge in wearable AI systems.

Traditional methods for detecting data poisoning in HAR systems, such as provenance tracking, have limitations. These approaches typically rely on large volumes of labeled data and retraining, which are often impractical in dynamic, real-time environments where immediate responses are needed. Additionally, the effectiveness of provenance-based detection methods heavily depends on the availability and accuracy of metadata about data origins. In scenarios where this metadata is missing, tampered with, or incomplete, these methods fail to effectively segment the data or detect poisoned samples. Moreover, adversaries can exploit these limitations by developing new strategies that deviate from the expected patterns of data manipulation, further reducing the adaptability of these traditional approaches.

Given these limitations, recent advancements in AI have opened new avenues for addressing the challenges of poisoning attacks. One of the most promising approaches involves leveraging Large Language Models (LLMs), which have demonstrated significant potential in recognizing human activities from wearable sensor data. Building on these findings, LLMs offer an opportunity to develop more effective poisoning detection mechanisms for HAR systems.

LLMs excel in zero-shot learning[6], enabling them to detect novel attack strategies without requiring large labeled datasets or retraining, which is a significant advantage in dynamic IoT ecosystems. Their ability to process both structured and unstructured data, combined with their contextual understanding of relationships between data points, makes LLMs well-suited for identifying anomalies, such as those introduced by data poisoning attacks. Additionally, LLMs' adaptability to emerging attack vectors offers a more robust solution compared to traditional methods, particularly in fast-evolving environments.

This research builds on the recent success of LLMs in recognizing human activities from sensor data [5] to explore their potential for detecting data poisoning attacks in wearable AI systems. Our objective is to design a prompt-based framework that leverages LLMs for the classification of sensor data, such as accelerometer and gyroscope readings, while also identifying instances where action labels have been tampered with. By using a zero-shot learning approach, this method aims to address the shortcomings of traditional techniques and provide a scalable, adaptive solution for real-time poisoning detection in dynamic environments. In a nutshell, contrary to existings works, this paper presents several significant contributions to advancing the detection of data poisoning attacks in wearable AI systems.

- **Innovative Approach to Poisoning Detection:** We propose a novel framework that utilize the zero-shot learning capabilities of large language models (LLMs) for detecting data poisoning attacks in wearable sensor systems. This approach eliminates the need for extensive labeled datasets and retraining, making it particularly effective in dynamic and real-time environments.

- **Advanced Label Sanitization Method:** We introduce a new method for label sanitization by utilizing the contextual understanding of LLMs. This technique identifies and corrects tampered activity labels in sensor data, promising higher integrity and reliability in Human Activity Recognition (HAR) systems.

- **Addressing Critical Gaps in Existing Solutions:** Our research identifies key limitations in current data poisoning detection techniques, such as scalability issues and reliance on complete data provenance. We address these challenges by offering a more adaptable and scalable solution that is resilient to emerging attack strategies in evolving IoT ecosystems.

- **Comprehensive Evaluation of LLM Models:** We rigorously simulate and evaluate our proposed methodologies using cutting-edge models like ChatGPT 3.5, ChatGPT 4, and Gemini, demonstrating the practicality, robustness, and effectiveness of LLMs in real-world poisoning detection scenarios.

In the remainder of this paper, related work is presented in Section 2. the proposed framework is described in Section 3. Section 4 discusses the evaluation and results. Finally, Section 5 concludes the paper.

## 2 Related Work

### 2.1 Data Poisoning Attack

Data poisoning attacks happen when attackers secretly add malicious data samples to the training dataset or modify the training data by label flipping then a machine learning model trained according to adversarial data samples. For instance, an attacker can modify the label in a human activity recognition context based on sensor data. Perdisci et al. [10] create the first poisoning instance in cybersecurity where attacks against worm signatures. Gupta et al. [4] suggest a novel poisoning attack based on federated learning that inverts the loss function of a model by creating malicious gradients at every SGD iteration and they test it using MNIST, Fashion-MNIST, and CIFAR-10 datasets. Shahid et al. [15] propose a label-flipping poisoning attack on wearable sensor data in human activity recognition and test it on a multi-layer perceptron, decision tree, random forest, and XGBoost. Additionally, they propose another method based on context-aware spatiotemporal poisoning attacks in human activity recognition. This attack exploits specific spatiotemporal patterns and conditions to manipulate the labels [13]. Gan et al. [16] proposes a federated learning-based poisoning attack that efficiently derives gradients for poisoned data. In our method, we use a label-flipping poisoning attack on human action sensor data and detect it using large language models with zero-shot learning.

### 2.2 Defence Mechanisms

Various types of defense mechanisms have been introduced, such as data aggregation [8, 17], data augmentation [11], and sanitization [2, 12]. In data aggregation, setting the weights of parameters helps reduce the impact of the poisoned data. An aggregation-based authentication defense technique called Deep Partition Aggregation (DPA) splits the training set into disjoint subgroups directly. Sanitization cleans the data before training, while data augmentation adds regularization to decision boundaries to prevent misclassification of data [3]. Fang et al. [13] suggest two defense strategies such as maximizing the influence of estimation (MIE) and median-of-weighted-average (MWA). Andrea et al. [9] proposes a label sanitization mechanism to label flipping data poisoning attacks. It uses k-nearest Neighbors (k-NN) to ensure that instances close to each other have similar labels, especially in areas far from the decision boundary. For each data point, the algorithm finds its k nearest neighbors. If most of these neighbors HHAR the same label and this majority meets a certain threshold, the data point is relabeled to match.

## 3 The Proposed framework

In this section, we present the design and implementation of our proposed method for detecting poisoning attacks, which leverages large language models (LLMs) with zero-shot learning. This approach allows for detecting maliciously altered data without the need for extensive task-specific training data. By integrating LLMs with zero-shot learning, our method can generalize and identify poisoning attacks across publicly available datasets, such as the MotionSense dataset[7] and the Heterogeneity Activity Recognition dataset[1]. Specifically, we target data poisoning attacks that affect both inter-class similarities and inter-class differences for both datasets.

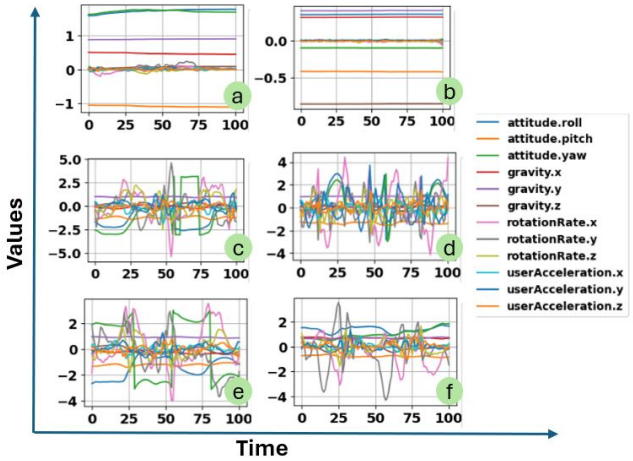


Figure 1: Visualization of activity data in Motionsense Dataset[7]: (a) Standing, (b) Sitting, (c) Walking, (d) Jogging, (e) Upstairs, and (f) Downstairs.

As shown in Figure 1, activities such as 'standing' and 'sitting,' 'upstairs' and 'downstairs,' or 'walking' and 'jogging' are closely related. We group these activities into *inter-class similarity categories* and specifically target label flips

between these closely related activities. For instance, we consider label flips from 'sitting' to 'standing,' 'upstairs' to 'downstairs,' and 'walking' to 'jogging,' as well as from 'walking' or 'jogging' to 'upstairs' or 'downstairs' for the MotionSense dataset. Additionally, we include label flips between 'biking' and 'walking' for the HHAR dataset. We analyze 14 activities with inter-class similarities, considering both directions of these label flips for the MotionSense and HHAR datasets.

For inter-class differences, we consider cases where the label flip involves distinctly different activities, such as flipping the label from 'walking' to 'standing,' 'standing' to 'jogging,' or 'walking' to 'sitting.' We refer to these as *Inter-Class Differences*. In the experiment, we analyze 16 activities with inter-class differences in both datasets.

### 3.1 Threat Model

Building on these inter-class similarities and differences, our threat model assumes the presence of an adversary whose primary objective is to degrade the accuracy of the action recognition model. The adversary could exploit the inter-class similarities by subtly flipping the labels between closely related activities, making it harder for the model to detect such manipulations. Additionally, the adversary may target inter-class differences by randomly introducing malicious labels between distinct activities, further corrupting the training process and dismantling the recognition. We focus on two types of poisoning attacks: the first involves randomly flipping the labels of certain variables, while the second targets specific variables for label flipping. An example of this would be actions like 'sitting' and 'standing,' which are inherently difficult to distinguish from each other. Our assumption is that the adversary has some form of access to the raw data within the dataset and can manipulate the training labels in both random and targeted ways to subvert the human action recognition model.

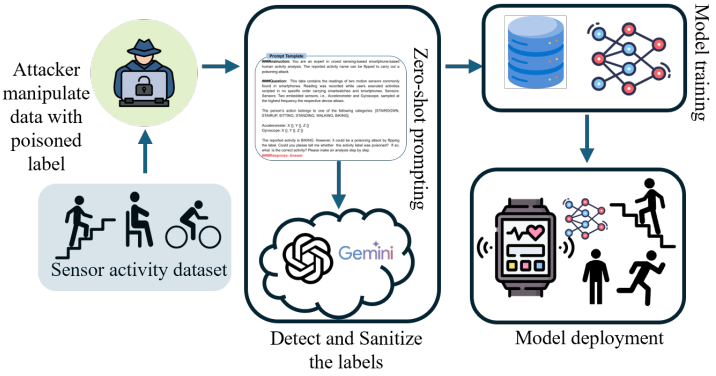


Figure 2: Overall Architecture of the Proposed Framework with LLMs as the Data Poisoning Detection and Sanitization Agent.

### 3.2 LLM-based Poisoned Data Detection and Sanitization

To counter such a threat model, our proposal leverages the benefits of Large Language models (LLM). We aim not only to detect the subtle inconsistencies and malicious poisoning introduced by adversaries but also to correct these data based on contextual understanding. The entire process of this proposed framework is depicted in Figure 2.

In the proposed framework, the poisoned data first goes through an LLM-based module. This module is responsible to detect poisoned data and sanitize them. This approach offers several benefits, including early detection and sanitization of the data, and scalability. As we are dealing with sensor data around activities, a window of the data is fed to the LLM (e.g. 100 continuous data samples) first. Next, we employ a zero-shot prompt template and large language models (LLMs), including ChatGPT-3.5, ChatGPT-4, and Gemini, to identify and sanitize the poisoned labels. Once sanitized, these labels are integrated into the model training phase and are ready for deployment in wearable AI systems, ensuring the integrity and accuracy of the trained models.

In this process, we create a prompt that includes instructions with a question and integrate sensor data to identify the flipped activity using zero-shot prompting. Figure 3 illustrates an example of zero-shot prompting for both the MotionSense dataset [7] and the Heterogeneity Activity Recognition dataset [1]. Those prompt templates target the dataset's features and apply them to the actual action data with the flipped labels. Based on that, large language models can detect the data poisoning attack on that action data.

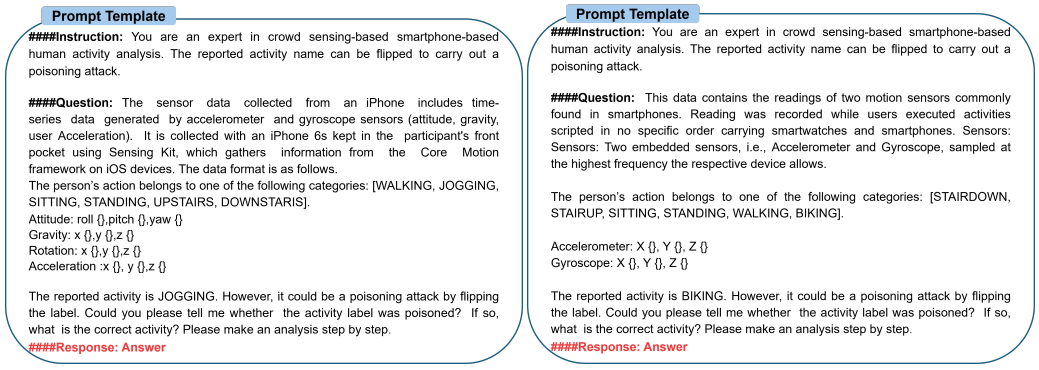


Figure 3: Prompt template for Motion Sense Dataset(left) and Prompt template for HHAR Dataset(right)

Algorithm 1 represents the method for designing a data poisoning attack and detecting and sanitizing the labels on human action sensor data. First, in the DataPoisoningModule (DPM), continuous data samples are randomly selected, and their labels are poisoned. This poisoned data is then sent to the DataPoisoningDetectionModule (DPDM), where the data poisoning attack is identified. Finally, in the LabelSanitizationModule (LSM), the labels that were flipped are sanitized.

**Algorithm 1** Detecting Data Poisoning Attack and Label Sanitization

**Data:** $\mathcal{D} \leftarrow$ Human activity sensor data $(A1, A2, \ldots, Ai)$ (MostionSense, HHAR)
$\mathcal{P} \leftarrow$ Prompt template
$\mathcal{L} \leftarrow$ LLMs with prompt template $\mathcal{P}$ and human activity sensor data$(A1, A2, \ldots, Ai)$ (MostionSense, HHAR)
**Result:** Detected data poisoning attack and Sanitization of Label

**Function** `DataPoisoningModule(DPM)`$(\mathcal{D})$**:**
  $\mathcal{D}_{l_i} \leftarrow$ Poisoned the labels of sensor data $\mathcal{D}$
  **return** $\mathcal{D}_{l_i}$(Send to DataPoisoningDetectionModule(DPDM))
**Function** `DataPoisoningDetectionModule(DPDM)`$(\mathcal{D}, \mathcal{D}_{l_i}, \mathcal{P})$**:**
  $\mathcal{D}_{p_i} \leftarrow$ Detect the data poisoning attack in human action data $\mathcal{D}$
  **return** $\mathcal{D}_{p_i}$(Send to LabelSanitizationModule(LSM))
**Function** `LabelSanitizationModule(LSM)`$(\mathcal{D}_{p_i}, \mathcal{P})$**:**
  $\mathcal{L}_{p_i} \leftarrow$ Sanitize the label in flipped label in human action data $\mathcal{D}$ **return** $\mathcal{L}_{p_i}$

# 4 Results and Discussions

## 4.1 Experimental Setup

### 4.1.1 Datasets

To facilitate this experiment, we use the MotionSense Dataset [7] and the Heterogeneity Activity Recognition dataset [1]. MotionSense Dataset was collected with an iPhone 6s kept in the participant's front pocket using SensingKit[1], which gathers information from the Core Motion framework on iOS devices. In 15 trials, 24 participants of various genders, ages, weights, and heights completed 6 different activities in the same setting: walking, jogging, upstairs, downstairs, sitting, and standing. Each sensor data point in the motion sense dataset consists of attitude, which includes three components: roll, pitch, and yaw, as well as gravity, rotation, and acceleration, all measured along the x, y, and z axes. The heterogeneity Activity Recognition dataset was collected through the smartphone and smartwatch sensors through the 4 smartwatches (2 LG watches, 2 Samsung Galaxy Gears) and 8 smartphones (2 Samsung Galaxy S3 Mini, 2 Samsung Galaxy S3, 2 LG Nexus 4, 2 Samsung Galaxy S++). And it contains 6 different activities 'Biking', 'Sitting', 'Standing', 'Walking', 'Stair Up' and 'Stair down'. Each sensor point in the HHAR dataset consists of accelerometer and gyroscope data, measured along the x, y, and z coordinates.

[1] `https://www.sensingkit.org/`

Table 1: Comparison of Detection and Label Sanitization in ChatGPT 3.5/4 and Gemini for Inter-Class Similarities in MotionSense Dataset

| Actual Label | Poisoned Label | ChatGPT 3.5 | | ChatGPT 4 | | Gemini | |
|---|---|---|---|---|---|---|---|
| | | Detection | Label Sanitization | Detection | Label Sanitization | Detection | Label Sanitization |
| Standing | Sitting | Yes | Standing | Yes | Standing | Yes | Standing |
| Sitting | Standing | Yes | Sitting or Lying Down | Yes | Sitting | Yes | Downstairs or Upstairs |
| Upstairs | Downstairs | No | Downstairs | Yes | Walking | Yes | Jogging or Walking |
| Downstairs | Upstairs | Yes | Walking | Yes | Downstairs | Yes | Downstairs |
| Upstairs | Jogging | Yes | Walking | Yes | Upstairs | No | Upstairs |
| Downstairs | Jogging | Yes | Walking | Yes | Downstairs | Yes | Downstairs or Upstairs |
| Jogging | Upstairs | Yes | Downstairs | Yes | Jogging | Yes | Walking or Standing |
| Jogging | Downstairs | Yes | Walking or Jogging | Yes | Jogging | Yes | Walking or Standing |
| Jogging | Walking | Yes | Downstairs | Yes | Jogging | Yes | Standing |
| Walking | Jogging | Yes | Walking | Yes | Walking | Yes | Walking or Standing |
| Walking | Upstairs | No | Upstairs | Yes | Walking | Yes | Walking or Standing |
| Upstairs | Walking | Yes | Jogging | Yes | Upstairs | Yes | Jogging |
| Walking | Downstairs | No | Downstairs | Yes | Walking | Yes | Walking or Standing |
| Downstairs | Walking | Yes | Jogging | Yes | Downstairs | Yes | Downstairs or Upstairs |

Shaded rows indicate that ChatGPT-3.5, ChatGPT-4, and Gemini can correctly detect and sanitize labels.

#### 4.1.2 Poisoning Attack Simulation, and Large Language Models

We randomly select continuous segments of 100 sensor data samples from both the MotionSense and Heterogeneity activity recognition(HHAR) datasets and systematically flip the labels associated with each action type. The reason for selecting a 100-sample window is that we are following the method by Sijie et al. [5], which targets detecting human actions using zero-shot learning. Our primary focus is on exploring and analyzing the similarities and differences between classes. To detect label poisoning and assess label sanitization for each action, we apply a zero-shot prompt template, as illustrated in Figure 1, using ChatGPT 3.5, ChatGPT 4, and Gemini.

### 4.2 Evaluation using LLMs

For the overall evaluation of ChatGPT-3.5, ChatGPT-4, and Gemini, we employ a comprehensive set of performance metrics, including accuracy and recall. Accuracy is calculated for both datasets based on the correct identification of poisoning attacks out of the total number of cases. Recall measures true positives when the LLM successfully sanitizes the actual label and false negatives when it fails to do so. These metrics provide a detailed assessment of each model's ability to classify and handle poisoning detection, enabling a nuanced comparison of their strengths and weaknesses in terms of prediction accuracy and correct positive classification. Tables I and II present the results for LLM models when data poisoning attacks target inter-class similarities in the MotionSense and HHAR datasets. Tables III and IV represent inter-class differences in the MotionSense and HHAR datasets, while Table V compares the results for detecting poisoning attacks and label sanitization in ChatGPT-3.5, ChatGPT-4, and Gemini using the MotionSense and HHAR datasets.

#### 4.2.1 ChatGPT-3.5

Overall, as shown in Table V, ChatGPT-3.5 achieves an accuracy of 0.9 for detecting poisoning attacks on the MotionSense dataset and 0.83 on the HHAR dataset. However, as indicated in Table I, ChatGPT-3.5 struggles to identify poisoning attacks when the actual label is "upstairs" and the poisoned label is "downstairs," as well as when the actual label is "walking" and the poisoned labels

Table 2: Comparison of Detection and Label Sanitization in ChatGPT 3.5/4 and Gemini for Inter-Class Similarities in HHAR Dataset

| Actual Label | Poisoned Label | ChatGPT 3.5 | | ChatGPT 4 | | Gemini | |
|---|---|---|---|---|---|---|---|
| | | Detection | Label Sanitization | Detection | Label Sanitization | Detection | Label Sanitization |
| Standing | Sitting | No | Sitting | No | Sitting | Yes | Stairsdown |
| Sitting | Standing | No | Standing | Yes | Sitting | No | Standing |
| Stairsup | Stairsdown | Yes | Sitting | Yes | Stairsup | Yes | Sitting or Standing |
| Stairsdown | Stairsup | Yes | Walking | Yes | Stairsdown | Yes | Stairsdown |
| Stairsup | Biking | Yes | Standing | Yes | Standing | Yes | Sitting or Standing |
| Stairsdown | Biking | Yes | Walking | Yes | Stairsdown | Yes | Stairsdown or Upstairs |
| Biking | Stairsup | Yes | Standing | Yes | Biking | Yes | Stairsdown |
| Biking | Stairsdown | Yes | Standing | Yes | Biking | Yes | Standing |
| Biking | Walking | Yes | Stairsdown | Yes | Biking | Yes | Standing |
| Walking | Biking | No | Walking | Yes | Walking | Yes | Standing |
| Walking | Stairsup | Yes | Standing | Yes | Walking | No | Stairsup |
| Stairsup | Walking | No | Walking | Yes | Stairsup | Yes | Stairsup |
| Walking | Stairsdown | Yes | Sitting or Standing | Yes | Walking or Stairsup | Yes | Standing |
| Stairsdown | Walking | Yes | Stairsdown | Yes | Stairsdown | Yes | Stairsdown |

Shaded rows indicate that ChatGPT-3.5, ChatGPT-4, and Gemini can correctly detect and sanitize labels.

are "upstairs" and "downstairs" on the MotionSense dataset. When considering inter-class differences in the MotionSense dataset, as demonstrated in Table III, ChatGPT-3.5 is more effective in detecting data poisoning attacks. Thus, ChatGPT-3.5 has difficulty distinguishing between "upstairs" and "downstairs" activities when these activities are similar within the same class. It also fails to detect poisoning attacks when the actual label is "standing" and the poisoned label is "sitting," when the actual label is "sitting" and the poisoned label is "standing," when the actual label is "walking" and the poisoned label is "biking," and when the actual label is "stairsup" and the poisoned label is "walking" on the HHAR dataset, considering inter-class similarities as in Table II. However, when considering inter-class differences on the HHAR dataset, ChatGPT-3.5 only distinguishes between "biking" and "standing," as shown in Table IV. The recall of ChatGPT-3.5 is 0.20 for the MotionSense dataset and 0.23 for the HHAR dataset. These values are low because ChatGPT-3.5 fails to sanitize the labels correctly. However, ChatGPT-3.5 can successfully sanitize the actual label when the action is "standing" for both inter-class similarities and differences.

### 4.2.2 ChatGPT-4

ChatGPT-4 achieves an accuracy of 1.0 for detecting poisoning attacks on the MotionSense dataset and 0.97 on the HHAR dataset, as shown in Table V. While it successfully detects poisoning attacks on the MotionSense dataset, it fails to identify the actions "standing" and "sitting" on the HHAR dataset when considering inter-class similarities, as indicated in Table II. The recall for ChatGPT-4 is 1.0 on the MotionSense dataset and 0.93 on the HHAR dataset, as shown in Table V. This indicates that while ChatGPT-4 can correctly sanitize labels on the MotionSense dataset, it struggles to do so on the HHAR dataset. This issue arises when ChatGPT-4 is unable to sanitize the label "standing" when the poisoned label is "sitting," or when the actual label is "stairsup" and the poisoned label is "biking," as shown in Table II.

### 4.2.3 Gemini

Gemini achieves an accuracy of 0.9 in detecting poisoning attacks on the MotionSense dataset and 0.93 on the HHAR dataset, as shown in Table V. This occurs when considering inter-class similarities, as indicated in Table I for the MotionSense dataset. Gemini fails to identify poisoning attacks when the actual label is "upstairs" and the poisoned label is "jogging." However, when considering inter-

Table 3: Comparison of Detection and Label Sanitization in ChatGPT 3.5/4 and Gemini for Inter-Class Difference in the MotionSense Dataset

| Actual Label | Poisoned Label | ChatGPT 3.5 | | ChatGPT 4 | | Gemini | |
|---|---|---|---|---|---|---|---|
| | | Detection | Label Sanitization | Detection | Label Sanitization | Detection | Label Sanitization |
| Standing | Walking | Yes | Standing | Yes | Standing | Yes | Standing |
| Standing | Jogging | Yes | Standing | Yes | Standing | Yes | Standing |
| Standing | Upstairs | Yes | Standing | Yes | Standing | Yes | Standing |
| Standing | Downstairs | Yes | Standing | Yes | Standing | Yes | Standing |
| Walking | Standing | Yes | Going Downstairs | Yes | Walking | No | Standing |
| Jogging | Standing | Yes | Walking or Jogging | Yes | Jogging | No | Standing |
| Upstairs | Standing | Yes | Jogging | Yes | Upstairs | Yes | Jogging or Walking |
| Downstairs | Standing | Yes | Walking | Yes | Downstairs | Yes | Downstairs or Upstairs |
| Sitting | Walking | Yes | Sitting or Standing | Yes | Sitting | Yes | Downstairs or Upstairs |
| Sitting | Jogging | Yes | Standing | Yes | Sitting | Yes | Sitting |
| Sitting | Upstairs | Yes | Sitting or Standing | Yes | Sitting | Yes | Walking, Jogging or Going Downstairs |
| Sitting | Downstairs | Yes | Sitting or Standing | Yes | Sitting | Yes | Jogging or Walking |
| Walking | Sitting | Yes | Going Upstairs | Yes | Walking | Yes | Walking or Standing |
| Jogging | Sitting | Yes | Walking or Jogging | Yes | Jogging | Yes | Walking or Standing |
| Upstairs | Sitting | Yes | Jogging | Yes | Upstairs | Yes | Jogging |
| Downstairs | Sitting | Yes | Walking | Yes | Downstairs | Yes | Downstairs or Upstairs |

Shaded rows indicate that ChatGPT-3.5, ChatGPT-4, and Gemini can correctly detect and sanitize labels.

class differences, Gemini successfully identifies poisoning attacks when the actual label is "walking" and the poisoned label is "standing," and when the actual label is "jogging" and the poisoned label is "standing," as shown in Table III. Gemini struggles to distinguish between "standing" and "sitting," as well as between "walking" and "stairsup," on the HHAR dataset when considering inter-class similarities, as shown in Table II. However, when considering inter-class differences, Gemini is able to detect data poisoning attacks for all activities on the HHAR dataset. The recall for Gemini was 0.23 on the MotionSense dataset and 0.30 on the HHAR dataset, as shown in Table V. These low values indicate that Gemini often fails to sanitize the actual label correctly for both inter-class similarities and differences. However, Gemini can correctly sanitize the label when the actual label is "stairsdown" and the poisoned label is "walking" on the HHAR dataset for inter-class similarities, as shown in Table II. Additionally, when considering inter-class similarities on the MotionSense dataset, as shown in Table I, Gemini can sanitize the label "standing" when the poisoned label is "sitting." It also suggests the correct label when the actual label is "walking" and the poisoned label is "jogging," though it suggests "walking or standing" with only 50 percent identification as "standing." When considering inter-class differences, as shown in Tables III and IV for both datasets, Gemini can sanitize the label "standing" when the poisoned labels are other activities on the MotionSense dataset. It can also correctly sanitize the labels when the actual label is "standing" and the poisoned label is "biking," when the actual label is "stairsup" and the poisoned label is "standing," and when the actual label is "sitting" and the poisoned label is "biking" on the HHAR dataset.

Overall, all LLMs, ChatGPT-3.5, ChatGPT-4, and Gemini are able to detect data poisoning attacks and sanitize labels for inter-class similarities on the MotionSense dataset when the actual label is "standing" and the poisoned label is "sitting," or when the actual label is "walking" and the poisoned label is "jogging." They can also do so when the actual label is "stairsdown" and the poisoned label is "walking" on the HHAR dataset. When considering inter-class differences, all LLMs are able to detect and sanitize labels when the actual label is "standing" on the MotionSense dataset, as well as when the actual label is "standing" and the poisoned label is "biking," the actual label is "stairsup"

Table 4: Comparison of Detection and Label Sanitization in ChatGPT 3.5/4 and Gemini for Inter-Class Difference in the HHAR Dataset

| Actual Label | Poisoned Label | ChatGPT 3.5 | | ChatGPT 4 | | Gemini | |
|---|---|---|---|---|---|---|---|
| | | Detection | Label Sanitization | Detection | Label Sanitization | Detection | Label Sanitization |
| Standing | Walking | Yes | Standing | Yes | Standing | Yes | Stairsdown |
| Standing | Biking | Yes | Standing | Yes | Standing | Yes | Standing |
| Standing | Stairsup | Yes | Standing | Yes | Standing | Yes | Stairsdown |
| Standing | Stairsdown | Yes | Sitting | Yes | Standing | Yes | Standing |
| Walking | Standing | Yes | Stairsup or Stairsdown | Yes | Walking | Yes | Stairsup |
| Biking | Standing | No | Standing | Yes | Biking | Yes | Stairsdown |
| Stairsup | Standing | Yes | Stairsup | Yes | Stairsup | Yes | Stairsup |
| Stairsdown | Standing | Yes | Stairsup | Yes | Stairsdown | Yes | Stairsdown |
| Sitting | Walking | Yes | Standing | Yes | Sitting | Yes | Standing |
| Sitting | Biking | Yes | Sitting or Standing | Yes | Sitting | Yes | Sitting |
| Sitting | Stairsup | Yes | Sitting or Standing | Yes | Sitting | Yes | Standing |
| Sitting | Stairsdown | Yes | Sitting | Yes | Sitting | Yes | Standing |
| Walking | Sitting | Yes | Walking | Yes | Walking | Yes | Standing |
| Biking | Sitting | Yes | Walking | Yes | Biking | Yes | Standing |
| Stairsup | Sitting | Yes | Standing | Yes | Stairsup | Yes | Sitting or Standing |
| Stairsdown | Sitting | Yes | Walking | Yes | Stairsdown | Yes | Standing |

Shaded rows indicate that ChatGPT-3.5, ChatGPT-4, and Gemini can correctly detect and sanitize labels.

Table 5: Comparison of Results for detecting poisoning attack and label sanitization in ChatGPT-3.5 ,ChatGPT-4 and Gemini on the MotionSense and HHAR Datasets

| Model | MotionSense | | HHAR | |
|---|---|---|---|---|
| | Accuracy | Recall | Accuracy | Recall |
| ChatGPT-3.5 | 0.90 | 0.20 | 0.83 | 0.23 |
| ChatGPT-4 | 1.00 | 1.00 | 0.97 | 0.93 |
| Gemini | 0.90 | 0.23 | 0.93 | 0.30 |

and the poisoned label is "standing," and the actual label is "sitting" and the poisoned label is "biking" on the HHAR dataset.

## 5 Conclusion

In this paper, we introduced a novel system for detecting data poisoning attacks and sanitizing labels in sensor-based activity data using zero-shot prompting. We created targeted attacks based on inter-class similarities, where activities are similar to each other, and inter-class differences, where activities differ from each other. We developed prompt templates and tested 100 data samples with poisoned data from the MotionSense and HHAR datasets using ChatGPT-3.5, ChatGPT-4, and Gemini. Among the results, ChatGPT-4 performed the best overall in detecting data poisoning attacks and sanitizing labels in a zero-shot manner. Despite these promising results, there are several opportunities for future work to further enhance the system's capabilities by exploring one-shot and few-shot prompting. Additionally, incorporating a broader range of inter-class similarities and differences and comparing the system with traditional techniques for detection and sanitization, while considering computational efficiency, would be valuable areas for improvement.

## References


[1] Bhattacharya Sourav Prentow Thor Kjrgaard Mikkel Blunck, Henrik and Anind Dey. Heterogeneity Activity Recognition. UCI Machine Learning Repository, 2015. DOI: https://doi.org/10.24432/C5689X.

[2] Patrick PK Chan, Zhimin He, Xian Hu, Eric CC Tsang, Daniel S Yeung, and Wing WY Ng. Causative label flip attack detection with data complexity measures. *International Journal of Machine Learning and Cybernetics*, 12:103–116, 2021.

[3] Jiaxin Fan, Qi Yan, Mohan Li, Guanqun Qu, and Yang Xiao. A survey on data poisoning attacks and defenses. In *2022 7th IEEE International Conference on Data Science in Cyberspace (DSC)*, pages 48–55. IEEE, 2022.

[4] Prajjwal Gupta, Krishna Yadav, Brij B. Gupta, Mamoun Alazab, and Thippa Reddy Gadekallu. A novel data poisoning attack in federated learning based on inverted loss function. *Computers Security*, 130:103270, 2023.

[5] Sijie Ji, Xinzhe Zheng, and Chenshu Wu. Hargpt: Are llms zero-shot human activity recognizers? *arXiv preprint arXiv:2403.02727*, 2024.

[6] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.

[7] Mohammad Malekzadeh, Richard G Clegg, Andrea Cavallaro, and Hamed Haddadi. Protecting sensory data against sensitive inferences. In *Proceedings of the 1st Workshop on Privacy by Design in Distributed Systems*, pages 1–6, 2018.

[8] Chenglin Miao, Qi Li, Lu Su, Mengdi Huai, Wenjun Jiang, and Jing Gao. Attack under disguise: An intelligent data poisoning attack mechanism in crowdsourcing. In *Proceedings of the 2018 World Wide Web Conference*, pages 13–22, 2018.

[9] Andrea Paudice, Luis Muñoz-González, and Emil C Lupu. Label sanitization against label flipping poisoning attacks. In *ECML PKDD 2018 Workshops: Nemesis 2018, UrbReas 2018, SoGood 2018, IWAISe 2018, and Green Data Mining 2018, Dublin, Ireland, September 10-14, 2018, Proceedings 18*, pages 5–15. Springer, 2019.

[10] Roberto Perdisci, David Dagon, Wenke Lee, Prahlad Fogla, and Monirul Sharif. Misleading worm signature generators using deliberate noise injection. In *2006 IEEE Symposium on Security and Privacy (S&P'06)*, pages 15–pp. IEEE, 2006.

[11] Avi Schwarzschild, Micah Goldblum, Arjun Gupta, John P Dickerson, and Tom Goldstein. Just how toxic is data poisoning? a unified benchmark for backdoor and data poisoning attacks. In *International Conference on Machine Learning*, pages 9389–9398. PMLR, 2021.

[12] Sanjay Seetharaman, Shubham Malaviya, Rosni Vasu, Manish Shukla, and Sachin Lodha. Influence based defense against data poisoning attacks in online learning. In *2022 14th International Conference on COMmunication Systems & NETworkS (COMSNETS)*, pages 1–6. IEEE, 2022.

[13] Abdur R. Shahid, Syed Mhamudul Hasan, Ahmed Imteaj, and Shahriar Badsha. Context-aware spatiotemporal poisoning attacks on wearable-based activity recognition. In *IEEE INFOCOM 2024 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pages 1–2, 2024.

[14] Abdur R Shahid, Ahmed Imteaj, Shahriar Badsha, and Md Zarif Hossain. Assessing wearable human activity recognition systems against data poisoning attacks in differentially-private federated learning. In *2023 IEEE International Conference on Smart Computing (SMARTCOMP)*, pages 355–360. IEEE, 2023.

[15] Abdur R Shahid, Ahmed Imteaj, Peter Y Wu, Diane A Igoche, and Tauhidul Alam. Label flipping data poisoning attack against wearable human activity recognition system. In *2022 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 908–914. IEEE, 2022.

[16] Gan Sun, Yang Cong, Jiahua Dong, Qiang Wang, Lingjuan Lyu, and Ji Liu. Data poisoning attacks on federated machine learning. *IEEE Internet of Things Journal*, 9(13):11365–11375, 2021.

[17] Yuxi Zhao, Xiaowen Gong, Fuhong Lin, and Xu Chen. Data poisoning attacks and defenses in dynamic crowdsourcing with online data quality learning. *IEEE Transactions on Mobile Computing*, 22(5):2569–2581, 2021.