TOWARDS LLM-POWERED AMBIENT SENSOR BASED MULTI-PERSON HUMAN ACTIVITY RECOGNITION

AUTHOR VERSION

Xi Chen[1,2], Julien Cumin[1], Fano Ramparany[1], Dominique Vaufreydaz[2],

[1] Orange Innovation

[2] Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, 38000 Grenoble, France

ABSTRACT

Human Activity Recognition (HAR) is one of the central problems in fields such as healthcare, elderly care, and security at home. However, traditional HAR approaches face challenges including data scarcity, difficulties in model generalization, and the complexity of recognizing activities in multi-person scenarios. This paper proposes a system framework called LAHAR, based on large language models. Utilizing prompt engineering techniques, LAHAR addresses HAR in multi-person scenarios by enabling subject separation and action-level descriptions of events occurring in the environment. We validated our approach on the ARAS dataset, and the results demonstrate that LAHAR achieves comparable accuracy to the state-of-the-art method at higher resolutions and maintains robustness in multi-person scenarios.

Keywords: Human Activity Recognition · Large Language Model · Smart Home · IoT.

1 Introduction

Over the past two decades, **Human Activity Recognition (HAR)** using sensor technology has garnered increasing attention due to its potential applications in healthcare, security surveillance, and smart home environments. While many existing HAR systems employ camera-based technologies [3, 17], these methods often raise substantial privacy concerns, particularly in private settings. As a response, some researchers have explored wearable technologies [21], such as smartwatches and smartphones. However, the requirement for individuals to continuously carry these devices may compromise comfort and convenience. Consequently, ambient sensors have gained back prominence as a key solution in HAR, prized for their non-invasive while avoiding privacy concerns of cameras and microphones.

Ambient sensors can be strategically placed within environments to detect and log changes in the physical state, with each change defined as an event. Common types of ambient sensors include door, presence, temperature, energy consumption sensors, and so on. Given their limited sensing range, multiple sensors are typically installed throughout a space to achieve thorough sensing coverage. The interactions between humans and their surroundings, recorded by these sensors, can be furthermore analyzed to infer individual actions and activities. This technology is known as **Ambient Sensor-Based Human Activity Recognition (AHAR)**.

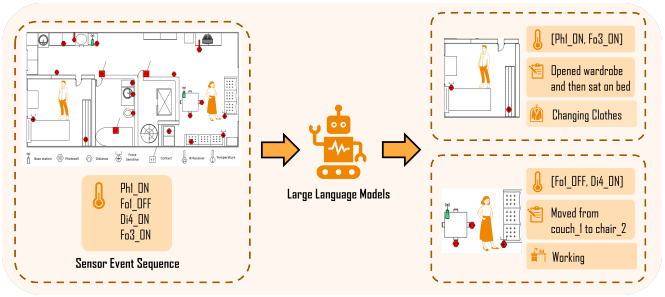


Figure 1: *Illustration of LLM-based multi-person AHAR*

However, AHAR faces the following challenges:

- **Data Collection**: Due to the high cost of setting up experimental environments and the sensitivity of personal daily living data, collecting ambient sensor datasets is often challenging.
- **Model Generalization**: Due to varying sensor setups and activity routines, models trained on specific datasets often struggle to transfer their capabilities to different environments or configurations.
- **Context Integration**: Contextual information like sensor locations, functions, time, environment, and user habits is crucial due to the simplicity of ambient sensor data. However, traditional deep learning methods often fail to efficiently encode this information, making HAR less precise and flexible.
- **Multi-Person Recognition**: In environments where multiple individuals are present, events triggered by different subjects will blend into a single event sequence, complicating the task of activity recognition.
- **Explainability**: Explainable HAR helps increase user trust, enhance user experience, and improve system personalization. However, the inference process of traditional deep learning models is not intuitively understandable and lacks explainability.

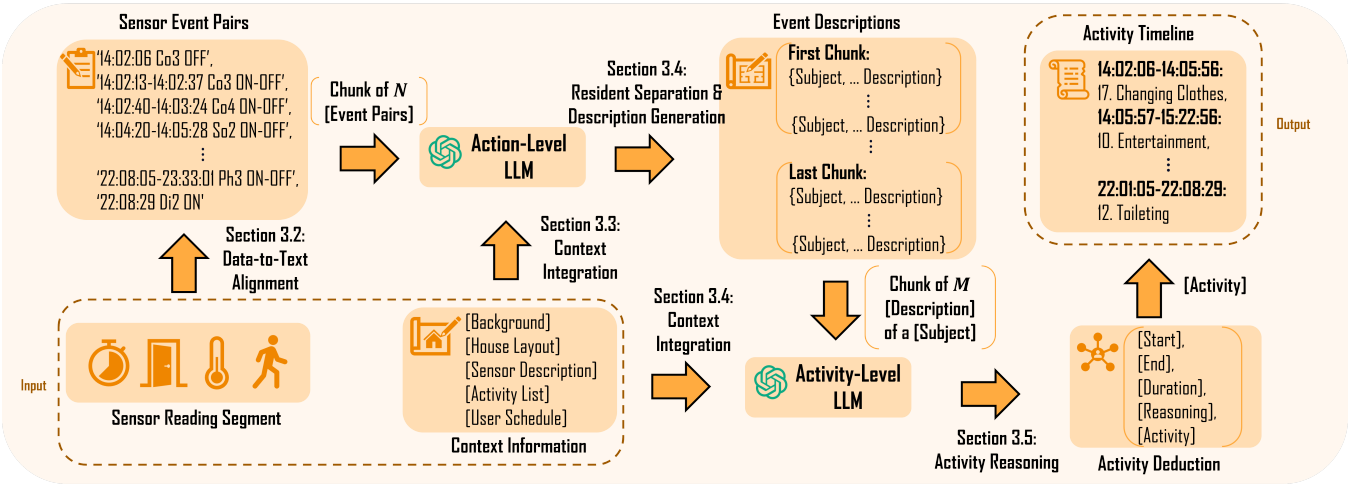


Figure 2: *Workflow for our proposed LLM-based AHAR framework: LAHAR.*

In recent years, significant advancements have been made in **Large Language Models (LLMs)**, with models such as ChatGPT [1] and Llama [15] exhibiting impressive contextual understanding and reasoning capabilities. This endows LLMs with the potential to address the aforementioned five challenges in AHAR: 1) LLMs' in-context learning capability [6] reduces the need for training datasets; 2) by adapting relevant prompts, LLMs can swiftly adapt to novel environments or adjust to new sensor configurations; 3) leveraging the expressiveness of natural language, LLMs can integrate the different types of contextual information; 4) LLMs can connect related events using attention mechanisms, integrating common sense and reasoning to identify meaningful sensor event combinations. Furthermore, LLMs' generation capabilities allow them to generate regrouped coherent sequences. Therefore, the LLMs have the potential to separate mixed event sequences in multi-person scenarios; 5) LLMs possess the ability to explain their reasoning process, thereby enhancing the explainability of the inference.

As illustrated in Figure 1, this study aims to design a multi-person AHAR system that leverages the advanced capabilities of LLMs to distinguish between different subjects' sensor events, describe their atomic actions, and ultimately predict their activities. To achieve this, we propose a two-stage framework, **LAHAR** (**L**LM-powered **AHAR**), designed to process multi-person sensor data from fine to coarse granularity, enabling few-shot learned recognition of activities. In the first stage, LAHAR is fed textualized sensor events at the level of seconds, assigns and describes each subject's actions at a fine granularity using natural language. Based on each subject's action descriptions collected from the first stage, LAHAR then performs reasoning in the second stage to predict a coarse timeline of activities spanning up to tens of hours.

The contributions of this study are listed as follows:

1. We propose an LLM-based AHAR approach which can be applied in multi-person scenarios. To the best of our knowledge, this is the first approach employing LLMs for recognizing multi-person activities from ambient sensor data.
2. We present a fine-to-coarse two-stage prompt engineering method that enables our system to continuously provide precise natural language descriptions of sensor data spanning over hours, and to further integrate these descriptions to infer daily living activities.

2 Related Work

Recently, increasing attention has been given to modeling ambient sensor sequences using natural language models. Bouchabou et al. [4, 5] first introduced the concept of language models into human activity recognition, treating each sensor event as a word (token), and used the word embedding method to learn the correlations between sensor events. Zhao et al. [22] further encoded the sensor environmental location into the embedding vectors, demonstrating the capability of language models to integrate context information. Das et al. [9] elaborated on the importance of explainability in activity recognition and implemented a system capable of explaining activity recognition classifications using natural language. Takeda et al. [14] first used the large language model GPT2 [12] for generative prediction of sensor event sequences, predicting future sensor events based on the labels of the ongoing activity and the sensor events that have already occurred. This work further strengthened the association of human activity recognition with language models and brought the large GPT model [11] into the scope of HAR.

With large language models demonstrating powerful in-context learning [6] and reasoning [20] abilities, Gao et al. [10] first used a large language model to perform unsupervised annotation on single-person activity samples in the ARAS dataset [2], demonstrating the potential of large language models for unsupervised human activity recognition. In this work, Gao et al. used sensor reading data within a 5-minute sliding window as input data. They employed a Chain-of-Thought approach [20] to instruct the LLM to analyze the functions of the

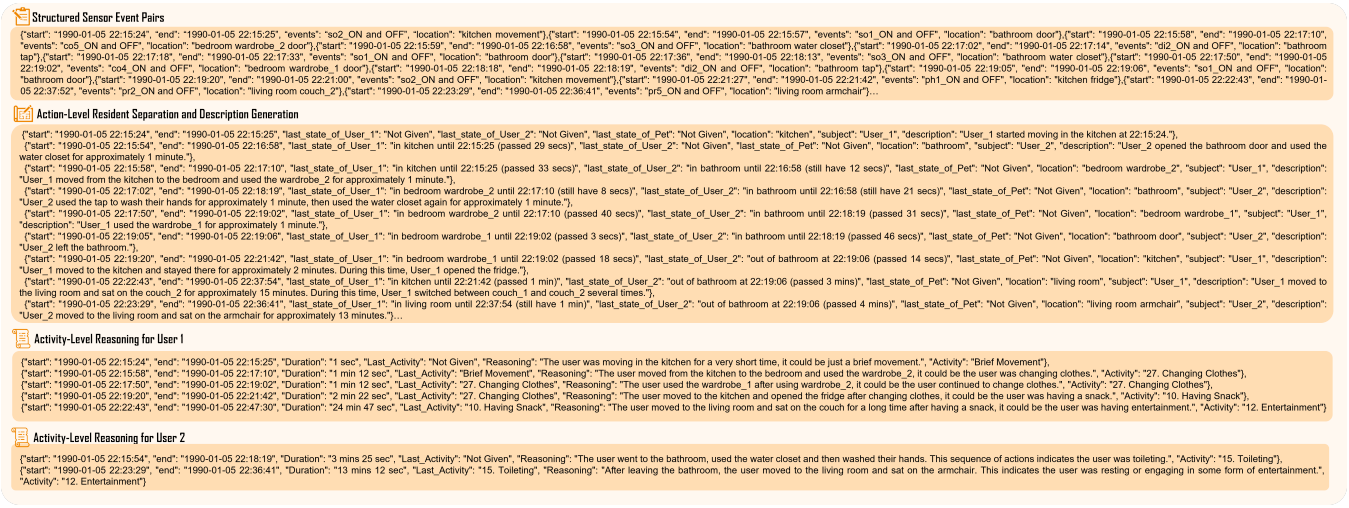


Figure 3: *Example of outputs generated by LAHAR at each stage*

activated sensors. By integrating context information on room layout, time, and the duration of sensor activation, LLM was finally instructed to choose an activity as the recognition results from nine activities selected by the authors. Although the experimental results show comparable accuracy to supervised trained models, this work is limited to nine easily distinguishable activity categories in single-person scenarios, overlooking the recognition of other more challenging categories and failing to provide prompts for reproducibility. Furthermore, using sensor readings in fixed time windows rather than sensor events as input data limits the model's ability to perceive the subject's behavior at a finer granularity.

Although language models are widely used in applications such as sensor representation, event sequence prediction, and activity explanation in single-user scenarios, these methods are often difficult to apply directly in multi-user scenarios. This is because only modeling the correlations of sensor events is insufficient to separate the activity information of different subjects in multi-person scenarios. To separate sensor events from different subjects, Wang and Cook [18, 19] first applied a skip-gram word embedding model to learn sensor correlations and then used a Gaussian Mixture Probability Hypothesis Density (GM-PHD) filter to cluster events into different tracks. Instead of using a probabilistic model, Chen et al. [7] employed a Sequence-to-Sequence model [13], using a machine-translation-like method, further applying language models to multi-person human activity recognition. This method first encoded the mixed event sequence of two subjects and then generatively decoded it into two single-person sequences separated by delimiters, thus achieving the final separation of multi-person event sequences. This research demonstrated the potential of generative language models for separating confounded information.

In this work, we leverage the powerful encoding capabilities of large language models, along with the separation abilities of generative methods, applying a generative LLM to multi-person activity recognition.

3 Methodology

Figure 2 illustrates the workflow of our proposed framework LAHAR. Given a time period T, the collection of all sensor readings within it is referred to as a *sensor reading segment*. LAHAR includes three main steps of information processing: 1) Process the sensor reading segment into a textual form of sensor event pairs (Section 3.2); 2) Integrate the context information (Section 3.3) into the sensor event pair sequence, separate subjects and generate individual action-level descriptions by an LLM (Section 3.4); 3) Based on the action-level descriptions and the context information, an LLM is used to perform activity-level reasoning to predict the timeline of activities for each subject (Section 3.5).

3.1 Problem Formalization

Given an environment $E = \{s_i\}_{1 \le i \le n}$, where s_i is a sensor installed in the environment characterized by its specific setting, we define a sensor event as $e_t =< t, s, c >$, where t represents the time of the event, s represents the sensor, and c represents the change in sensor status. The sequence of events that occur within a time period $T = [t_s, t_e]$ is $S_T = (e_{t_1}, e_{t_2}, ..., e_{t_k})$ where $\forall i \in [1, k], t_i \in [t_s, t_e]$. Given an activity category set of K activities $L_A = \{a_k\}_{1 \le k \le K}$, the activities occurred during T are defined as $A_T = \{a_k^{T_j}\}_{j \in \mathcal{J}}$, where $\mathcal{J} = \{j \in \mathbb{N} | T_j \subseteq T\}$. The objective of this research is to propose a model M such that $A_T = M(S_T | E, L_A)$.

3.2 Data-to-Text Alignment

As LLMs accept text as input, LAHAR first involves data-to-text alignment. This process includes two steps: data preprocessing, and information structuring.

3.2.1 Data Preprocessing

Unlike Gao et al. [10], who extract overall features from all sensor readings within a fixed time window, our method first

preprocesses the sensor readings into sensor events. Specifically, when there is a change in the reading of any sensor, we denote the time of occurrence t, the changed sensor identifier s, and the change of sensor reading c as a sensor event $e =< t, s, c >$. Since a sensor event often corresponds to an action by a subject, analyzing events allows our model to achieve fine-grained, action-level detail. Meanwhile, to reduce redundant information, when a sensor continuously changes at a high frequency between two states without any other sensor events occurring, we retain only the first and the last events.

3.2.2 Information Structuring

To further enhance the information density and quality of the text input to the LLM, we perform information structuring on the sequence of sensor events. For adjacent activation $e_{ON} =< t_s, s, \text{ON} >$ and deactivation events $e_{OFF} =< t_e, s, \text{OFF} >$ of a sensor, we pair them into an event pair p, incorporating the sensor's location information. We then structure them into a JSON format as follows: $p = \{\text{“}start\text{”} :< t_s >, \text{“}end\text{”} :< t_e >, \text{“}event\text{”} :< s >$ ON and OFF, $\text{“}location\text{”} :< l >\}$. In the end, all event pairs are provided to the LLM sorted in ascending order by start time. This structure of event pairs is designed to help the LLM identify residents' occupancy in multi-person scenarios. An example of final structured event pairs is illustrated in the first block of Figure 3.

3.3 Context Integration

Traditional machine learning methods struggle with ambient sensor data due to limited information from sensor readings. However, contextual information like sensor location, type, function, user habits, and environment layout often provide more insight than the sensor data itself. Integrating this contextual information is crucial for understanding the correlations between sensors and for activity recognition.

Given that ambient sensors are usually installed in a relatively stable environment, this contextual information tends to remain constant. Therefore, our method proposes to provide contextual information to LLMs through language prompts, so that LLMs can harness their encoding capabilities to embed this information and align it with relevant sensor events. The contextual information used in this work is listed as follows:

- **Background:** This introduces the role of the LLM, the number of residents, and the fact that ambient sensors are installed to identify activities.
- **House Layout:** This provides the list of rooms contained in the environment, the furniture in each room, and the associated sensors.
- **Sensor Description:** This explains the identifier, type, and location of each sensor in the environment.
- **Activity List:** This offers a list of possible activities within the environment, along with certain behavior patterns or user habits related to these activities.
- **User Schedule:** This emphasizes the intervals during which subjects perform certain activities, such as eating breakfast.

3.4 Action-Level Resident Separation and Description Generation

This section details the design of an LLM-powered module that assigns the sequential event pairs from the Data-to-Text module (Section 3.2) to different subjects, and then provides natural language descriptions with action-level granularity, as illustrated in the second block of Figure 3.

To assign events to the corresponding subjects, LAHAR employs a two-step process. First, it merges related sequential event pairs. For instance, if sensors in a water closet and a bathroom tap are triggered in quick succession, LAHAR treats these events as a single, cohesive event. Second, LAHAR determines the most likely subject responsible for these sensor events by evaluating the subjects' states from the previous time step. For example, a subject who has recently left a chair is more likely to trigger bathroom sensors than a subject who was previously determined to be asleep in bed. This process is based on two primary assumptions: 1) related sensor events are more likely to be triggered by the same person, and 2) a person cannot trigger sensors unrelated to their current state.

To enable reasoning based on these two assumptions, LAHAR employs prompt engineering techniques to endow the LLM with two abilities. The first ability, **inter-sensor relevance estimation**, enables the LLM to merge related sensor events of the same subject. The second ability, **sensor-subject relevance estimation**, allows the LLM to allocate sensor events to the most relevant subject. Both abilities hinge on relevance estimation, which is fundamentally supported by the attention mechanism [16] of the LLM. This mechanism ensures that related sensor events and the states of subjects are encoded with similar semantic representations within the given context. Although the LLM is pre-trained to encode natural language, additional In-Context Learning [6] is necessary to better align sensor events and subject states with the specific context of the task. Therefore, we incorporate the context described in Section 3.3 into the prompt. Additionally, the estimation of sensor-subject relevance requires the LLM to infer and deduce the previous state of each subject, which necessitates the use of Chain-of-Thought (CoT) reasoning [20]. This approach allows the LLM to logically sequence its deductions, thereby enhancing its ability to accurately match sensor events with the appropriate subjects.

Therefore, the prompt contains 4 basic components: 1) *Context*; 2) *Instructions*; 3) *Examples*; 4) *Input*, where *Context* and *Examples* follows the idea of In-Context Learning, and *Instructions* describes the Chain of Thought.

3.4.1 Input

Although the *Input* section appears last in the prompt, we introduce it first for clarity. Given a period T, the sequence of events S_T is formatted into a sequence of event pairs P_T following Data-to-Text alignment. Since P_T can be too long to ensure high-quality generation, we divide P_T into chunks C_i, each containing N event pairs, except for the last chunk, which contains the remaining pairs. We process each chunk sequentially in a loop, concatenating all responses at the end. To en-

able the LLM to infer the users' previous state at the beginning of each new step, we include the final description of each subject from the previous chunk into the input of the subsequent step.

3.4.2 Context

In this part of the prompt, *Background*, *House Layout*, and *Sensor Description* introduced in Section 3.3 are provided to the LLM as the context information. Formally, we have $Context = ([Background], [HouseLayout], [SensorDescription])$.

3.4.3 Instructions

We instruct the LLM to sequentially perform the following steps:

1. Merge related sequential event pairs, and determine the overall start and end times;
2. Summarize the previous action state of each user and determine whether their previous action has ended;
3. Recall the location of the current event pairs;
4. Considering the previous states of users, designate a related user as the subject for the current event pair being processed;
5. Describe the current event pair with natural language.

Ultimately, the prompt ask the LLM to respond in a predefined JSON format, which implicitly formalizes the Chain of Thought while making the generated results easier to post-process and increasing the information density. The keys defined in the JSON format are: {*"start"*, *"end"*, *"last state of User 1"*, ..., *"last state of User i"*, *"location"*, *"subject"*, and *"description"*}.

3.4.4 Examples

To further activate the LLM's ability to use context and follow the chain of thought for reasoning, the prompt provides several examples to the LLMs.

3.5 Activity-Level Reasoning

The objective of this second module is to align descriptions of fine granular action of each subject to each subject's activity timeline A_T, as shown in the last two blocks of Figure 3. For activities that are directly associated with sensors, LLMs can make use of common sense reasoning, such as associating sleeping with the pressure sensor of a bed. On the other hand, recognizing activities that are environment-specific and user-specific relies heavily on in-context learning. Consequently, the design of context and examples is crucial for this module. Similar to the Description Generation module, the prompt contains 4 basic components: 1) Context; 2) Instructions; 3) Examples; 4) Input.

3.5.1 Input

From the output of last module, we separate and reorganize the descriptions for each subject, retaining only 4 key-value pairs: "start", "end", "location", and the "description". After implementing the separation, we input each subject's descriptions independently. Similarly to the previous module, we divide each subject's descriptions into chunks, with each containing M descriptions.

3.5.2 Context

In this part of the prompt, *Sensor Description*, *Activity List*, and *User Schedule* introduced in Section 3.3 are provided to the LLM as the context information. Formally, we have $Context = ([SensorDescription], [ActivityList], [UserSchedule])$.

3.5.3 Instructions

We instruct the LLM to sequentially perform the following steps:

1. Analyse and summarise the descriptions that belong to the same activity, and determine the overall start and end times;
2. Calculate the duration of the acitivty;
3. Recall the last activity predicted;
4. Considering the previous activities of the subject and the duration of current actions, reason the subject's current activity;
5. Choose an activity with ID from the activity list.

Ultimately, we instruct the LLM to respond in a predefined JSON format, in which the keys defined are: {*"start"*, *"end"*, *"Duration"*, *"Last_Activity"*, *"Reasoning"*, and *"Activity"*}.

3.5.4 Examples

For activities that cannot be directly detected by sensors, they are often described by multiple groups of sensor events and typically exhibit certain patterns. We filter out the corresponding descriptions for these activities, then provide correct reasoning results to explain why these descriptions correspond to the given activity.

4 Experiments

4.1 Dataset

To evaluate LAHAR, we require an ambient-based multi-person HAR dataset that provides sufficient contextual information for all sensors, enabling them to be described with language. To the best of our knowledge, the ARAS dataset [2] best meets this requirement. It is a publicly available dataset that includes two real houses (named House A and B), each equipped with 20 ambient sensors. Within each house, a maximum of two subjects can concurrently be observed engaging

Table 1: *Activity ID and labels of ARAS Datasets after regrouping.*

ID	Activity	ID	Activity	ID	Activity	ID	Activity
0	Other	1	Preparing Breakfast	2	Having Breakfast	3	Preparing Lunch
4	Having Lunch	5	Preparing Dinner	6	Having Dinner	7	Washing Dishes
8	Having Snack	9	Sleeping	10	Entertainment	11	Having Shower
12	Toileting	13	Working	14	Shaving	15	Brushing Teeth
16	Talking on the Phone	17	Changing Clothes				

in any of 27 different daily activities. Each house dataset contains 30 files, representing 30 days. Each daily file contains sensor reading data and multi-person activity annotations, both at the level of seconds, thus including 86400 annotated data instances.

4.2 Experiment Settings

4.2.1 Data Segmentation

Although our method can reason coherently without prior data segmentation, we performed necessary segmentation. We noted that House A's data is daily independent, while House B's data spans 30 consecutive days. Thus, we concatenated House B's 30 days of data but treated each day in House A as an independent segment. For evaluation, we further divided the data into single-person and multi-person scenarios based on the "Leaving House" activity. Consequently, House A had 59 single-person and 61 multi-person segments, while House B had 10 single-person and 24 multi-person segments.

4.2.2 Error Preprocessing

We assessed sensor error levels in the houses by examining the number of events that occurred when both residents were leaving the house. According to our observation, House A exhibited significant noise, especially from the hall motion sensor, kitchen motion sensor, and kitchen temperature sensor. To address this, we removed the hall motion sensor events and deactivated the kitchen motion and temperature sensors, except during kitchen activities.

4.2.3 Class Selection and Regrouping

Due to similar activities in the ARAS dataset that sensors don't distinguish, we merged certain activities: *Napping* and *Sleeping* into *Sleeping*, and *Watching TV*, *Reading books*, and *Listening to music* into *Entertainment*. In House A, *Using Internet* and *Studying* were merged into *Working*; in House B, *Using Internet* was merged into *Entertainment*, and *Studying* was renamed to *Working*. We removed infrequent activities like *Laundry*, *Cleaning*, *Having conversations*, and *Having guests*, as recognizing these activities is beyond the capability of our method. This resulted in the list of activities shown in Table 1.

4.2.4 Parameters

Queries to the large language model are based on API calls to the gpt-4-32k-0613 model provided by the Azure OpenAI Service, with the Temperature parameter set to 0 to reduce the randomness of the model's output, while keeping other parameters at their default settings. For the two hyperparameters in LAHAR—the chunk size N for the Action-Level Resident Separation and Description Generation module, and the chunk size M for the Activity-Level Reasoning module—we used $N = 20$ and $M = 15$, respectively. This setup was determined based on preliminary experiments, taking into account two factors: on one hand, we need each chunk to contain as much context as possible, and on the other hand, chunks that are too long can impair the LLM's ability to follow instructions within the prompt.

4.3 Evaluation Metric

For a data segment whose time period is T, the activities occurred is denoted as $A_T = \{a_k^{T_j}\}_{j \in \mathcal{J}}$, where a_k is k-th activity class in K classes and $\mathcal{J} = \{j \in \mathbb{N} | T_j \subseteq T\}$. We perform one-hot encoding for all the activities $\{a_k\}$ present at each second of T and apply the union operation. For instance, if the i-th and j-th activities are ongoing at the second t, the encoding is a length-K vector with ones at positions i and j and zeros elsewhere. By stacking these vectors for all seconds in T, we obtain a two-dimensional matrix $M_{T\times K}$. To compare our prediction $\hat{M}_{T\times K}$ with the ground truth $M_{T\times K}$, we define the following:

$$S_{T\times K} = \hat{M}_{T\times K} \cdot M_{T\times K},$$

where $\cdot$ denotes element-wise multiplication. Using this, we calculate:

$$TP = \left[\sum_{t\in T} S_{t,k}\right]_{1\times K},$$

$$FP = \left[\sum_{t\in T} (\hat{M}_{t,k} - S_{t,k})\right]_{1\times K},$$

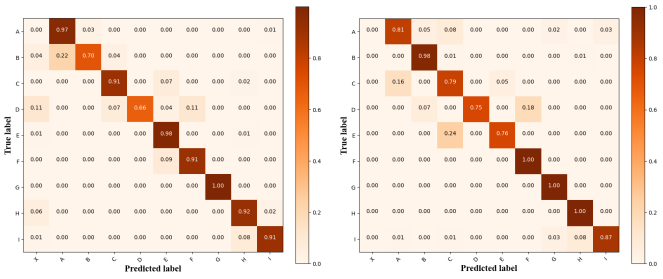
$$FN = \left[\sum_{t\in T} (M_{t,k} - S_{t,k})\right]_{1\times K}.$$

Based on TP, FP, and FN given above, we can then calculate the precision, recall, and F1-score of each class in our prediction. To evaluate models' performance globally, we compute both the macro-average, which is the simple average of the metrics across all classes, and the weighted average, which accounts for the time occupied by each class.

4.4 Single-Subject Activity Recognition

4.4.1 Comparison with the State-of-the-art

To validate the activity recognition capability of LAHAR, a comparison is performed against the research of Gao et al. [10]. The experimental setup for this comparison is consistent with the research of Gao el al., focusing solely on the recognition of selected nine activity categories in single-person scenarios. We extract the longest data segments from the original data where Resident 2 was leaving home and Resident 1's activities are in these nine activities. These segments



(a) *LAHAR: at the level of seconds* ***(b)*** *Gao et al. [10]: at the level of 5 minutes*

Figure 4: *Comparison of confusion matrices between our method and the Gao et al. method [10]. The categories are: X) Unknown, A) Preparing Breakfast, B) Having Breakfast, C) Preparing Lunch, D) Having Lunch, E) Preparing Dinner, F) Having Dinner, G) Sleeping, H) Having Shower, I) Toileting.*

Table 2: *Comparison of performance between our method and the state of the art*

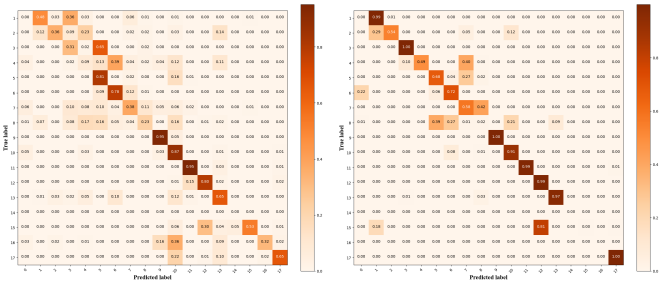
Method	Resolution	Precision	Recall	F1-score
Gao et al. [10]	5 minutes	96.00	95.56	95.60
LAHAR	Second	88.54	92.31	90.39

vary in length and can contain more than one activity. Without additional segmentation, our method can generate action-level descriptions and achieve activity sequence prediction at the resolution of a second. In contrast, the method of Gao et al, which did not use events as the smallest divisible units but instead used a fixed 5-minute time window, has thus a resolution of 5 minutes. Despite our higher resolution, our results are comparable to the results of Gao et al. in terms of the confusion matrix in Figure 4 and of precision, recall, and F1 score as shown in Table 2.

4.4.2 Extended Validation

We further validated our method in more realistic and complex scenarios. As described in the experiment settings, our single-person scenario data include 17 activities across both houses, without class-based segmenting. Consequently, each segment is longer and contains more actions, making it more complex compared to the setting of Gao et al. [10].

Figure 5 presents the confusion matrices of both houses. Furthermore, Table 3 gives the precision, recall, and F1 scores of each class. From the confusion matrix of House A, it can be observed that compared to extracting data segments for 9 activities individually in Gao's setting, having more categories and longer data segments results in a lag in predicting activities related to breakfast and lunch. The primary reason is that the users in House A have their meals 2-3 hours later than typical meal times. Despite explicitly highlighting this discrepancy in the prompt, the LLMs still tend to classify meals based on conventional timing norms. Another reason is the issue of activity alternation. For example, if a subject briefly watches TV



(a) *House A* ***(b)*** *House B*

Figure 5: *Confusion matrices of single-user activity recognition.*

while preparing breakfast and then resumes breakfast preparation, the LLM might interpret this as the subject having already prepared breakfast earlier and now preparing lunch. In House B, the activity of brushing teeth is difficult to accurately recognize because the subjects usually use the toilet after brushing their teeth. This leads the LLM to merge and predict brushing teeth and using the toilet as a single activity of toileting.

4.5 Multi-Subject Activity Recogntion

Since the ARAS dataset does not label events with the IDs of their subjects, we cannot perform a one-to-one comparison of event assignments. To validate our method's activity recognition capability in multi-person scenarios, we qualitatively present an example of our results and indirectly demonstrate our method's ability to separate residents by comparing its performance with that in single-person scenarios.

4.5.1 Qualitative Results

Figure 3 provides an example to better illustrate our results. We excerpted the experimental output of about 21 minutes of sensor data from 22:15:24 to 22:36:41 of the 5th day in House B in the multi-person scenario. It can be seen that these sensor events were integrated into 9 descriptions by the LLM, in which each description is assigned to a subject. By separating these descriptions by subjects, timestamped activities are finally predicted respectively for each resident.

4.5.2 Quantitative Results

In Table 3, we present the recognition results for each activity class in multi-person scenarios. These results demonstrate that even when extended to multiple people, our method's performance in activity recognition remains comparable to its performance in single-person scenarios. Despite differences in time and activity distributions between single-person and multi-person scenarios, this comparison highlights the scalability of LAHAR in multi-person contexts. Furthermore, we observed that performance in the multi-person scenario in House A is higher than in the single-person scenario. This difference is primarily because mealtimes in multi-person scenarios in House A are closer to conventional meal times compared to

Table 3: *Results of activity recognition of each class in different scenarios*

Metric		Precision (%)				Recall (%)				F1-score (%)			
Scenario		Single_A	Multi_A	Single_B	Multi_B	Single_A	Multi_A	Single_B	Multi_B	Single_A	Multi_A	Single_B	Multi_B
Class	Preparing Breakfast	59.95	66.54	67.41	80.07	50.45	62.17	98.12	84.38	54.79	64.28	79.92	82.16
	Having Breakfast	65.05	97.70	97.29	96.20	39.21	74.18	53.93	82.05	48.92	84.33	69.40	88.56
	Preparing Lunch	25.08	27.37	67.31	73.46	30.02	17.84	100.	68.74	27.33	21.60	80.46	71.02
	Having Lunch	5.19	47.87	100.	47.78	9.00	33.33	62.22	66.23	6.58	39.30	76.71	55.51
	Preparing Dinner	21.53	64.90	66.22	54.56	73.84	84.78	67.59	76.56	33.33	73.52	66.90	63.71
	Having Dinner	7.59	58.92	11.67	31.39	82.47	65.67	70.04	29.91	13.89	62.11	20.01	30.63
	Washing Dishes	46.22	41.06	3.85	10.37	33.02	12.57	57.55	7.97	38.52	19.25	7.21	9.00
	Having Snack	40.53	48.07	0.58	12.84	22.04	26.27	1.54	44.42	28.56	33.97	0.85	19.92
	Sleeping	87.84	95.53	93.80	97.84	93.99	88.20	100	97.82	90.81	91.72	96.80	97.83
	Entertainment	70.44	69.31	98.58	89.45	88.65	90.66	90.85	90.28	78.50	78.56	94.55	89.86
	Having Shower	72.95	59.39	100	60.27	91.82	81.05	84.99	86.57	81.30	68.55	91.89	71.06
	Toileting	82.67	53.74	82.40	91.37	76.86	81.43	96.76	84.27	79.66	64.75	89.00	87.68
	Working	76.60	81.50	99.51	98.39	61.08	60.49	98.81	85.63	67.96	69.44	99.16	91.57
	Shaving	/	84.16	/	0.	/	53.17	/	0.	/	65.17	/	0.
	Brushing Teeth	75.24	46.63	1.09	84.32	43.86	34.30	0.40	51.15	55.42	39.53	0.58	63.68
	Talking on the Phone	98.67	42.83	/	0.	29.56	33.96	/	0.	45.49	37.88	/	0.
	Changing Clothes	49.93	64.54	83.07	95.19	52.56	59.45	92.06	82.95	51.22	61.89	87.34	88.65
Macro-Average		55.34	61.77	64.85	60.20	54.90	56.44	71.66	61.11	50.14	57.40	64.05	59.46
Weighted-Average		66.00	77.67	92.37	93.60	67.56	76.23	93.79	90.28	76.95	86.49	93.08	91.91

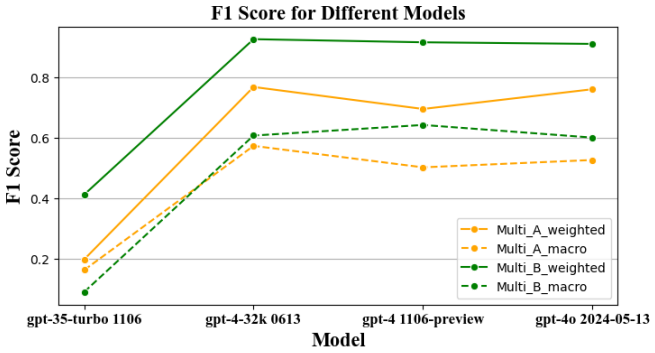


Figure 6: *A comparison of the impact of different LLM models on LAHAR*

single-person scenarios, making the prediction of meal activities more accurate.

4.5.3 Impact of LLM model

To investigate the impact of the capabilities of large language models (LLMs) on the results of LAHAR, we compared the macro and weighted F1 scores obtained in House A and House B by applying our method to four different LLM models, as shown in Figure 6. The models under comparison, listed in ascending order of their capabilities [8], are: *gpt-35-turbo 1106*, *gpt-4-32k 0613*, *gpt-4 1106-preview*, and *gpt-4o-2024-05-13*. The weakest model, *gpt-35-turbo 1106*, exhibits a significant decline in performance compared to the other three models in the GPT-4 series. Our observations suggest that this decline is primarily due to its inability to perform fine-grained reasoning, often excessively merging and omitting events, which results in a loss of critical details for activity recognition. We speculate that the underlying cause is GPT-3.5's insufficient capability to retrieve information from long contexts. Furthermore, performances of the GPT-4 series models have not shown improvement with increased model capacity, indicating that our methods may not yet fully exploit the potential of LLMs.

5 Conclusion

In this paper, we propose LAHAR, a framework using LLMs for multi-person HAR with ambient sensors. Our prompts enable LLMs to assign sensor events to individuals based on their states, generating detailed descriptions and reasoning about their activities. This method extends LLM application to multi-person HAR, achieving time resolutions matching sensor timestamps. LAHAR's explicit descriptions and activity reasoning offer promising perspectives to address explainability challenges. Experimental validation shows performance comparable to the state-of-the-art in single-person and multi-person scenarios. Future plans include validation with different LLMs, model fine-tuning, and further evaluation of conversational explainability.

References

[1] Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)

[2] Alemdar, H., Ertan, H., Incel, O.D., Ersoy, C.: Aras human activity datasets in multiple homes with multiple residents. In: 2013 7th International Conference on Pervasive Computing Technologies for Healthcare and Workshops. pp. 232–235. IEEE (2013)

[3] Arshad, M.H., Bilal, M., Gani, A.: Human activity recognition: Review, taxonomy and open challenges. Sensors **22**(17), 6463 (2022)

[4] Bouchabou, D., Nguyen, S.M., Lohr, C., Leduc, B., Kanellos, I.: Fully convolutional network bootstrapped by word encoding and embedding for activity recognition in smart homes. In: Deep Learning for Human Activity Recognition: Second International Workshop, DL-HAR 2020, Held in Conjunction with IJCAI-PRICAI 2020, Kyoto, Japan, January 8, 2021, Proceedings 2. pp. 111–125. Springer (2021)

[5] Bouchabou, D., Nguyen, S.M., Lohr, C., LeDuc, B., Kanellos, I.: Using language model to bootstrap human activity recognition ambient sensors based in smart homes. Electronics **10**(20), 2498 (2021)

[6] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020)

[7] Chen, X., Cumin, J., Ramparany, F., Vaufreydaz, D.: Generative resident separation and multi-label classification for multi-person activity recognition. In: 2024 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops). pp. 1–6. IEEE Computer Society, Los Alamitos, CA, USA (mar 2024)

[8] Chiang, W.L., Zheng, L., Sheng, Y., Angelopoulos, A.N., Li, T., Li, D., Zhang, H., Zhu, B., Jordan, M., Gonzalez, J.E., et al.: Chatbot arena: An open platform for evaluating llms by human preference. arXiv preprint arXiv:2403.04132 (2024)

[9] Das, D., Nishimura, Y., Vivek, R.P., Takeda, N., Fish, S.T., Ploetz, T., Chernova, S.: Explainable activity recognition for smart home systems. ACM Transactions on Interactive Intelligent Systems **13**(2), 1–39 (2023)

[10] Gao, J., Zhang, Y., Chen, Y., Zhang, T., Tang, B., Wang, X.: Unsupervised human activity recognition via large language models and iterative evolution. In: ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 91–95. IEEE (2024)

[11] Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al.: Improving language understanding by generative pre-training. OpenAI (2018)

[12] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al.: Language models are unsupervised multitask learners. OpenAI blog **1**(8), 9 (2019)

[13] Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. Advances in neural information processing systems **27** (2014)

[14] Takeda, N., Legaspi, R., Nishimura, Y., Ikeda, K., Minamikawa, A., Plötz, T., Chernova, S.: Sensor event sequence prediction for proactive smart home support using autoregressive language model. In: 2023 19th International Conference on Intelligent Environments (IE). pp. 1–8. IEEE (2023)

[15] Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023)

[16] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)

[17] Vrigkas, M., Nikou, C., Kakadiaris, I.A.: A review of human activity recognition methods. Frontiers in Robotics and AI **2**, 28 (2015)

[18] Wang, T., Cook, D.J.: smrt: Multi-resident tracking in smart homes with sensor vectorization. IEEE transactions on pattern analysis and machine intelligence **43**(8), 2809–2821 (2020)

[19] Wang, T., Cook, D.J.: Multi-person activity recognition in continuously monitored smart homes. IEEE transactions on emerging topics in computing **10**(2), 1130–1141 (2021)

[20] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al.: Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems **35**, 24824–24837 (2022)

[21] Zhang, S., Li, Y., Zhang, S., Shahabi, F., Xia, S., Deng, Y., Alshurafa, N.: Deep learning in human activity recognition with wearable sensors: A review on advances. Sensors **22**(4), 1476 (2022)

[22] Zhao, J., Suleiman, B., Alibasa, M.J.: Feature encoding by location-enhanced word2vec embedding for human activity recognition in smart homes. In: International Conference on Mobile and Ubiquitous Systems: Computing, Networking, and Services. pp. 191–202. Springer (2022)