Collaborative Lightweight LLM Agents for Daily Activity Summarization on Edge Devices

Kentaro Inohara
The University of Osaka, Japan
k-inohara@ist.osaka-u.ac.jp

Tatsuya Amano
The University of Osaka / RIKEN R-CCS, Japan
t-amano@ist.osaka-u.ac.jp

Hamada Rizk
The University of Osaka / RIKEN R-CCS, Japan
hamada_rizk@ist.osaka-u.ac.jp

Hirozumi Yamaguchi
The University of Osaka / RIKEN R-CCS, Japan
h-yamagu@ist.osaka-u.ac.jp

***Abstract*—This paper presents a privacy-preserving monitoring system for elderly individuals living alone, utilizing collaborative lightweight Large Language Model (LLM) agents deployed on edge devices. The system integrates non-invasive sensors with a novel three-agent architecture: an activity recognition agent processes sensor data, an hourly summarization agent generates intermediate reports in 3-hour segments, and a daily summarization agent produces comprehensive summaries. Our evaluation on real-world data from two elderly households demonstrates that the system achieves 95.8% accuracy in activity recognition and generates natural language summaries comparable to GPT-4o, while maintaining privacy through local processing on affordable Raspberry Pi hardware. The results indicate that our approach effectively balances monitoring accuracy, summary quality, and practical deployment constraints, making it suitable for widespread adoption in elderly care applications.**

***Index Terms*—Generative AI, Large Language Model (LLM), Human Activity Recognition, Elderly Monitoring**

I. Introduction

The proportion of elderly individuals living alone is rapidly increasing, with approximately 15.7% of Japanese households consisting of an elderly person living alone [1]. This demographic shift, combined with changing family structures and increasing urbanization, has led to a growing need for remote monitoring systems to ensure the safety and well-being of elderly residents.

While numerous monitoring systems using various sensors (e.g., wearable sensors [2], power consumption meters [3], and small-LiDAR sensors [4]) have been developed [5]–[7], family caregivers without technical expertise struggle to interpret the resulting sensor data and activity logs. Even when these systems successfully detect activities, family caregivers struggle to make sense of raw sensor readings or basic activity logs. They need more intuitive ways to understand their elderly relatives' daily living conditions and potential concerns.

To bridge this gap between technical data and human understanding, natural language descriptions provide an ideal solution, as they can convey complex information in an accessible format. Recent advancements in large language models (LLMs) have made it possible to automatically transform structured data into natural language summaries [8]. More importantly, LLMs can generate personalized summaries by incorporating user-defined preferences in natural language. By simply describing "an active senior who enjoys cooking" or "a primary caregiver focused on health patterns," the system can tailor its summaries to both elderly individuals' lifestyle preferences and family members' specific information needs. This intuitive personalization capability makes LLMs particularly suitable for generating accessible activity reports.

However, implementing such a LLM-based system for elderly monitoring presents two key challenges. First, we must protect privacy through local processing rather than cloud services. Second, we need to ensure wide adoption by operating on affordable edge devices like Raspberry Pi. Additionally, deploying LLMs on such resource-constrained devices faces significant technical hurdles, particularly regarding limited context windows and memory capacity.

To address these constraints, we propose a framework of collaborative LLM agents that break down the daily monitoring task into manageable segments. The framework distributes the complex workload across multiple agents working together in sequence. An activity recognition agent first interprets raw sensor data to identify daily activities, while the hourly summarization agent processes these activity predictions by generating intermediate summaries for every three-hour period throughout the day. The daily summarization agent then synthesizes them into a comprehensive daily report. This collaborative approach of lightweight specialized agents enables privacy-preserving operation even on affordable edge devices.

To validate our approach, we collected real-world data from two elderly households over six days. The dataset includes readings from infrared motion sensors and power meters that track appliance usage, providing insights into daily living patterns. Our evaluation results demonstrate that the proposed system running on Raspberry Pi 4 achieved 95.8% accuracy in correctly associating activities with their corresponding time periods, approaching the performance of the large-scale GPT-4o model (98.1%). Moreover, in human evaluation of summary quality and accuracy, our system outperformed GPT-4o (average score difference: +0.03 points) while maintaining privacy and cost efficiency through local processing.

II. Related Work

Prior research in elderly monitoring can be categorized into sensor-based activity recognition and monitoring systems for daily living support. Activity recognition approaches have utilized various sensors including wearable devices [5], [9], depth

cameras [10], and appliance power meters [3]. While these systems achieve high accuracy, they often face challenges in privacy protection or require elderly individuals to consistently wear devices. Monitoring systems have focused on either anomaly detection using physiological data [2] or long-term observation through ambient sensors like water usage meters [6]. However, these systems typically provide raw sensor data or basic metrics that family members find difficult to interpret.

Recent advances in large language models have enabled more natural data-to-text generation [8], suggesting their potential for generating intuitive activity reports. However, deploying LLMs for elderly monitoring presents unique challenges, particularly in terms of privacy preservation and computational constraints on edge devices. Our work bridges this gap by proposing a collaborative framework of lightweight agents that can operate locally while providing accessible natural language summaries for family members.

III. Proposed Methods

A. Overview

Figure 1 illustrates the operational flow of our monitoring system. The system integrates non-invasive sensors for data collection, an activity recognition agent that processes the sensor data, and a summarization agent composed of two specialized subagents. Each subagent handles specific aspects of the summarization process, with the first focusing on short-term activity patterns and the second synthesizing comprehensive daily reports based on predefined personas that capture the elderly individual's lifestyle and preferences. By breaking down the complex summarization task into smaller segments processed by specialized lightweight LLM agents, this modular design enables efficient operation even on resource-constrained edge devices.

B. Sensor Data Collection

Our system employs two types of non-invasive sensors: appliance power sensors and motion sensors. Appliance power sensors monitor the use of household devices, providing indirect indicators of specific activities (e.g., microwave usage for cooking), while maintaining privacy compared to cameras or microphones. Motion sensors complement this by capturing physical movement across different rooms, helping distinguish between activities with similar power consumption patterns such as watching TV versus doing housework. The system collects sensor data at 15-minute intervals, with ground truth labels obtained through direct feedback from elderly participants.

C. Human Activity Recognition Agent

The Random Forest model serves as the activity recognition agent, interpreting sensor data from 17 sensors to identify specific activities. The model considers both the target time slot and the preceding four time slots to incorporate temporal context, while employing the Synthetic Minority Oversampling Technique (SMOTE) [11] to address class imbalances in the training data.

The choice of Random Forest was motivated by its computational efficiency and interpretability compared to complex models such as deep neural networks, enabling local processing on edge devices without relying on external cloud infrastructure. This design ensures privacy preservation while maintaining sufficient accuracy for daily activity recognition.

D. Summarization Agent with Subagents

The summarization agent operates as a collaborative entity comprising two specialized subagents, the **Hourly-Summarizing Agent (HS-LLM)** and the **Daily-Summarizing LLM Agent (DS-LLM)**, which transform structured activity data into natural language summaries personalized for family members. This agent's design reflects a dual-step process, balancing fine-grained detail with overarching coherence to meet diverse user preferences.

The HS-LLM subagent segments the day ' s data into eight discrete parts, each representing a 3-hour period (12 timeslots). By summarizing these segments into concise text comprising 1–3 sentences, the Hourly-Summarizing LLM ensures the data remains manageable and interpretable for users. Fine-tuning with QLoRA optimizes its ability to handle complex inputs efficiently while operating on resource-constrained edge devices [12]. The training dataset for this subagent includes 15-minute, 12-slot activity data annotated with summaries generated by a larger LLM, such as GPT-4o, ensuring the output aligns with the expected quality and content relevance. Specifically, the model was fine-tuned using 469 samples to focus on maintaining accuracy and consistency in generating hourly summaries.

The DS-LLM subagent consolidates the segmented summaries into a comprehensive daily narrative, while incorporating predefined personas to generate personalized reports. These personas capture both the elderly individual's lifestyle preferences (e.g., "an active senior who enjoys cooking") and family members' information needs (e.g., "primary caregiver focused on health patterns"). To enable the model to efficiently focus on the necessary information in the segmented summaries and its temporal relationships, the day is divided into four key periods: *Early Morning (00:00–05:59)*, *Morning (06:00–11:59)*, *Afternoon (12:00–17:59)*, and *Evening (18:00–23:59)*. The DS-LLM combines these temporal segments with persona information to generate appropriate summaries. For example, when generating a report for a primary caregiver, the model emphasizes health-related patterns, while for family members interested in social engagement, it focuses more on daily activities and interactions. Fine-tuning of the DS-LLM, performed with QLoRA, utilized a dataset of 100 samples that included various persona combinations and their corresponding summary styles, enabling the system to generate distinctly different summaries from the same activity data [13].

The interaction between the HS-LLM and the DS-LLM ensures a balance between granularity and readability. The HS-LLM ' s segmented outputs allow detailed examination of specific time intervals, while the DS-LLM ' s synthesis creates a holistic narrative that emphasizes key activities and

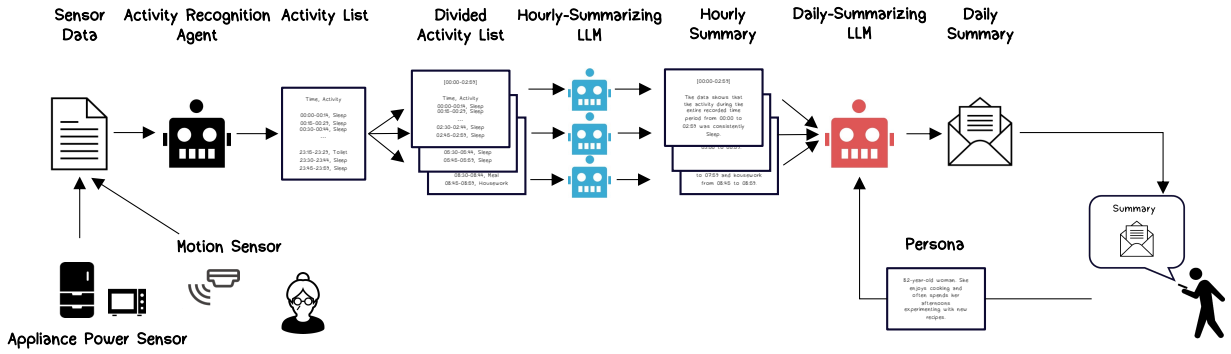

Fig. 1. Overview of Proposed Monitoring System.

insights. This division of labor between the subagents enables the summarization agent to cater to varied user requirements, such as providing detailed logs for some family members and concise overviews for others.

E. Collaborative Framework

The activity recognition agent provides structured activity labels to the summarization agent, whose two subagents (HS-LLM and DS-LLM) work together to generate personalized reports. This modular design enables privacy-preserving local processing on edge devices while maintaining adaptability to different user needs through lightweight models and fine-tuning techniques.

IV. Evaluation

A. Experimental Setup

The dataset was collected from elderly participants in their homes in Japan for six days each. The setup included motion sensors placed in key living areas (living room, kitchen, bedroom, bathroom, and corridor) to track movement patterns, and appliance power sensors monitoring daily-use appliances such as TV, microwave, washing machine, and refrigerator. This sensor combination enabled comprehensive monitoring of daily activities including sleeping, eating, cooking, housework, and going out.

Data was recorded at 15-minute intervals, with ground truth activity labels derived from participants' self-reported logs. This setup in actual living environments enabled evaluation of the proposed system under realistic daily life conditions.

B. Evaluation Metrics

We evaluated our system using three distinct metrics. For activity recognition, we employed both standard accuracy and a time-flexible custom accuracy metric with a 30-minute tolerance window (including two slots before and after the target slot) to better reflect real-world monitoring needs where precise timing is less critical than activity identification.

For summary evaluation, we first adopted the widely-used Multidimensional Quality Metric (MQM) [14], which assesses summaries across eight dimensions: unnecessary information (ADD), missing information (OMI), irrelevant details (IE), time order (TO), time-activity relationships (TAR), word order (WO), grammar (WF), and duplication (DUP).

To specifically evaluate the temporal accuracy of activity descriptions, we developed an additional Time-Activity Level Precision-Source (TAL $prec_s$) metric, inspired by entity-level evaluations [15]:

$$\text{Time-Activity Level } prec_s = \frac{N(h \cap s)}{N(h)} \tag{1}$$

where $N(h)$ represents total activities and $N(h \cap s)$ represents correctly timed activities. This metric eliminates the need for reference summaries, enabling direct evaluation of temporal accuracy in the generated descriptions.

C. Results

1) Human Activity Recognition: The proposed system demonstrates strong performance in activity recognition using the Random Forest model (Table I). The model achieved an standard accuracy of **0.7604** and a time-flexible accuracy of **0.9167**, showcasing its ability to reliably detect activities in real-world settings with temporal flexibility. The model excelled in identifying key activities such as "Sleep" and "Go Out" with perfect precision and recall. These results confirm the system's reliability for monitoring stable and distinct activities. High performance in categories like "Other" and "Housework" further demonstrates its sensitivity to nuanced sensor data variations, ensuring actionable insights in elderly monitoring scenarios. While moderate prediction scores were observed for "Cook" and "Meal," these results reflect the inherent challenge of overlapping sensor patterns in these activities. The system's ability to accurately detect sporadic activities such as "Bath" highlights its adaptability. Misclassifications occurred primarily in low-frequency activities, which can be addressed by optimizing sensor placement and expanding the dataset.

2) Summary Generation: Thirteen participants evaluated the generated summaries using the MQM criteria on a five-point scale. Tables II and III show the MQM evaluation results and TAL $prec_s$ scores, respectively.

Our fine-tuned gemma2 model achieved comparable TAL $prec_s$ scores to GPT-4o when using estimated activity lists (0.87 vs 0.80), while maintaining more concise summaries ($\Sigma N(h)$ = 54 vs 62). When using actual activity lists, both models showed high accuracy, with GPT-4o slightly outperforming (0.98 vs 0.95). Fine-tuning improved gemma2's

TABLE I
ACTIVITY RECOGNITION ACCURACY OF RANDOM FOREST MODEL

Activity Label	Precision	Recall	F1-Score	Support
Sleep	1.00	1.00	1.00	31
Meal	0.50	0.25	0.33	4
Cook	0.20	0.33	0.25	3
Toilet	0.00	0.00	0.00	6
Bath	1.00	0.50	0.67	4
Housework	0.89	0.73	0.80	11
Go Out	1.00	1.00	1.00	2
Other	0.70	0.80	0.75	35

TABLE II
RESULTS OF MQM-BASED HUMAN EVALUATION

Configuration	ADD	OMI	IE	TO	TAR	WO	WF	DUP	Average
EA + FT gemma2-2b	3.077	4.000	3.769	4.385	4.308	**4.769**	4.538	4.154	4.125
AA + FT gemma2-2b	3.615	**4.308**	4.000	**4.538**	**4.462**	4.615	**4.615**	4.385	**4.317**
EA + GPT-4o	2.923	3.846	3.000	3.769	3.615	4.538	4.385	4.154	3.779
AA + GPT-4o	3.846	4.231	**4.154**	**4.538**	**4.462**	4.462	4.231	4.385	4.288
EA + gemma2-2b	3.846	3.308	3.462	4.154	4.154	4.462	4.308	**4.538**	4.029
AA + gemma2-2b	**4.308**	3.154	3.923	4.077	4.000	4.538	4.154	4.385	4.067

EA: Estimated Activity list, AA: Actual Activity list, FT: Fine-Tuned

TABLE III
EVALUATION RESULTS OF TIME-ACTIVITY LEVEL PRECISION-SOURCE

Configuration	$\Sigma N(h)$	$\Sigma N(h \cap s)$	TAL $prec_s$
EA + FT gemma2	54	47	0.8704
AA + FT gemma2	48	46	0.9583
EA + GPT-4o	62	50	0.8065
AA + GPT-4o	52	51	**0.9808**
EA + gemma2	45	36	0.8000
AA + gemma2	37	30	0.8108

EA: Estimated Activity list, AA: Actual Activity list, FT: Fine-Tuned

performance across all configurations, particularly in summary content quality metrics (OMI, IE, TO, and TAR) as shown in Table II. This improvement can be attributed to our task segmentation approach, which allows the model to focus on key information more effectively. While all models demonstrated strong linguistic capabilities (WO, WF, DUP), the proposed method achieved the highest scores in time-related aspects (TO, TAR) and information completeness (OMI), suggesting that our approach effectively balances accuracy and conciseness in activity summarization.

3) Edge Device Performance: To evaluate the feasibility of local deployment, we assessed the computational performance using a Raspberry Pi 4 (8GB) running Gemma2 2b model via Ollama. The system generated hourly summaries (223 characters) in 1.5 minutes and daily summaries (1,098 characters) in 6.5 minutes, with average inference speeds of 0.6-0.7 tokens/s. Given that elderly monitoring reports are typically reviewed the following day, these processing times are practical for the intended use case, demonstrating the system's viability on cost-effective edge devices.

V. CONCLUSION

We proposed and evaluated a privacy-preserving monitoring system for elderly individuals living alone, integrating LLMs with non-invasive sensors and machine learning models. Our experimental results demonstrated that the proposed method, using small-scale fine-tuned models, achieved comparable or superior performance to GPT-4o while maintaining privacy through local processing on edge devices. The system's effectiveness in generating understandable summaries and its practical implementation on affordable hardware suggests its potential for wide adoption in elderly care applications.

ACKNOWLEDGEMENT

This work was partially funded by JST CREST Grant JPMJCR21M5, JST PRESTO Grant JPMJPR2361 and JSPS KAKENHI 23H03384.

REFERENCES

[1] Ministry of Health, Labour and Welfare, Japan, "Summary report of comprehensive survey of living conditions 2023," https://www.mhlw.go.jp/toukei/saikin/hw/k-tyosa/k-tyosa23/index.html, 2023, [Accessed 23-Jan.-2024].

[2] M. Al-khafajiy, T. Baker, C. Chalmers *et al.*, "Remote health monitoring of elderly through wearable sensors," *Multimedia Tools and Applications*, vol. 78, no. 17, pp. 24 681–24 706, Sep 2019.

[3] K. Ishizu, T. Mizumoto, H. Yamaguchi *et al.*, "Home activity pattern estimation using aggregated electricity consumption data," *Sensors and Materials*, vol. 33, no. 1, pp. 69–88, 2021.

[4] S. Yamada, H. Rizk, T. Amano *et al.*, "Fall detection and assessment using multitask learning and micro-sized lidar in elderly care," in *Mobile and Ubiquitous Systems: Computing, Networking and Services*, A. Zaslavsky, Z. Ning, V. Kalogeraki *et al.*, Eds. Cham: Springer Nature Switzerland, 2024, pp. 280–293.

[5] A. Hayat, F. Morgado-Dias, B. P. Bhuyan *et al.*, "Human activity recognition for elderly people using machine and deep learning approaches," *Information*, vol. 13, no. 6, 2022.

[6] T. Koketsu, Y. Ohno, T. Ishihara *et al.*, "Monitoring living activities of the elderly living alone using a lifeline," *Japanese Journal of Applied IT Healthcare*, vol. 13, no. 2, pp. 12–19, 2018.

[7] X. Ji, X. Li, A. Yuh *et al.*, "Short: Integrated sensing platform for detecting social isolation and loneliness in the elderly community," in *2023 IEEE/ACM Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)*, 2023, pp. 148–152.

[8] P. Laban, W. Kryscinski, D. Agarwal *et al.*, "SummEdits: Measuring LLM ability at factual reasoning through the lens of summarization," in *Proc. of the 2023 Conference on Empirical Methods in Natural Language Processing*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 9662–9676.

[9] M. Z. Uddin and A. Soylu, "Human activity recognition using wearable sensors, discriminant analysis, and long short-term memory-based neural structured learning," *Scientific Reports*, vol. 11, no. 1, p. 16455, Aug 2021.

[10] S. Park, J. Park, M. Al-masni *et al.*, "A depth camera-based human activity recognition via deep learning recurrent neural network for health and social care services," *Procedia Computer Science*, vol. 100, pp. 78–84, 2016.

[11] N. V. Chawla, K. W. Bowyer, L. O. Hall *et al.*, "Smote: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, p. 321–357, Jun. 2002.

[12] T. Dettmers, A. Pagnoni, A. Holtzman *et al.*, "Qlora: Efficient finetuning of quantized llms," 2023.

[13] G. Team, M. Riviere, S. Pathak *et al.*, "Gemma 2: Improving open language models at a practical size," *arXiv preprint arXiv:2408.00118*, 2024.

[14] D. Huang, L. Cui, S. Yang *et al.*, "What have we achieved on text summarization?" in *Proc. of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, B. Webber, T. Cohn, Y. He *et al.*, Eds., 2020, pp. 446–469.

[15] F. Nan, R. Nallapati, Z. Wang *et al.*, "Entity-level factual consistency of abstractive text summarization," in *Proc. of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, P. Merlo, J. Tiedemann, and R. Tsarfaty, Eds. Online: Association for Computational Linguistics, Apr. 2021, pp. 2727–2733.