Vital Insight: Assisting Experts' Sensemaking of Multi-modal Personal Tracking Data Using Visualization and LLMs

Jiachen Li
li.jiachen4@northeastern.edu
Northeastern University
Boston, Massachusetts, USA

Xiwen Li
li.xiwe@northeastern.edu
Northeastern University
Boston, Massachusetts, USA

Akshat Choube
choube.a@northeastern.edu
Northeastern University
Boston, Massachusetts, USA

Justin Steinberg
steinberg.ju@northeastern.edu
Northeastern University
Boston, Massachusetts, USA

Bingsheng Yao
b.yao@northeastern.edu
Northeastern University
Boston, Massachusetts, USA

Xuhai Xu
xx2489@columbia.edu
Columbia University
New York, New York, USA

Dakuo Wang
d.wang@northeastern.edu
Northeastern University
Boston, Massachusetts, USA

Elizabeth Mynatt
e.mynatt@northeastern.edu
Northeastern University
Boston, Massachusetts, USA

Varun Mishra
v.mishra@northeastern.edu
Northeastern University
Boston, Massachusetts, USA

Abstract

In modern ubiquitous computing studies, there remains a significant gap in translating sensing streams from passive tracking devices into meaningful, high-level, context-aware insights that are required for various applications. In this study, we design and develop Vital Insight (VI), a novel, LLM-assisted, prototype system to enable inference and visualizations of personal tracking data. VI aims to provide high-level summary and insights to assist researchers in understanding the data collected from multi-modal passive sensing data such as smartphones and wearables. We conducted a preliminary user study with 13 experts to assess general usability and identify opportunities for improvement.

ACM Reference Format:
Jiachen Li, Xiwen Li, Akshat Choube, Justin Steinberg, Bingsheng Yao, Xuhai Xu, Dakuo Wang, Elizabeth Mynatt, and Varun Mishra. 2025. Vital Insight: Assisting Experts' Sensemaking of Multi-modal Personal Tracking Data Using Visualization and LLMs. In *Proceedings of (CHI '25 Workshop on Envisioning the Future of Interactive Health).* ACM, New York, NY, USA, 6 pages.

1 Background

The ubiquitous presence of sensor-rich smartphones and wearables has prompted researchers to use data from these devices to track living activities for various outcomes including health tracking, context-aware applications, and digital health interventions [4, 15, 17, 21, 27, 28, 30, 35, 37, 40, 43]. However, it is extremely challenging to generate valuable high-level insights such as "is it a normal day" or a summary of someone's day that is often more valuable for various stakeholders [1, 36].

CHI '25 Workshop on Envisioning the Future of Interactive Health, Yokohama, Japan

Realizing the potentials of Large Language Models (LLMs) to memorize and use 'common sense' and world knowledge to directly generate insights from sensor data, such as identifying activities, making diagnostic predictions, and more [13, 14, 16, 23, 26, 45], we explore the possibility of using LLMs and visualization to assist researchers in understanding personal tracking data. We develop Vital Insight that leveraged both visualization and LLM augmentation to provide experts with high-level insights, and invited 13 experts to provide preliminary feedback.

2 Formative Interview

We first conducted IRB-approved formative interviews with 12 experts[1] to understand the needs in inferring complex real-world personal tracking data. Two researchers reviewed the interview transcripts and employed an inductive open coding approach to conduct a thematic analysis following Grounded Theory [6, 8, 19].

From the interviews, we identify two primary needs based on these paths: **direct representation** of data and **indirect inference**. Experts require a direct representation of sensing data to manage the vast volumes of information, as sensor data often does not come in a human-readable format. They commonly relied on visualizations for this purpose, however, most visualizations lack systematic design, are overly generic, and are often created ad-hoc for specific tasks. Experts mentioned techniques like stacking modalities or creating correlation graphs, but these efforts are sporadic and have not been standardized even within the same research group. Beyond direct data representation, experts also expressed a need for assistance with automatic inference generation. Even with effective visualizations, the transformation of data into actionable insights is still predominantly a manual process that the experts conduct. P9 highlighted a challenge,

> *"We just have so much data, and it'd be pretty overwhelming to just kinda put that all in there." (P9)*

[1]We define "experts" as people who are experienced researchers in ubiquitous computing who have utilized passive sensing data for health-related applications.

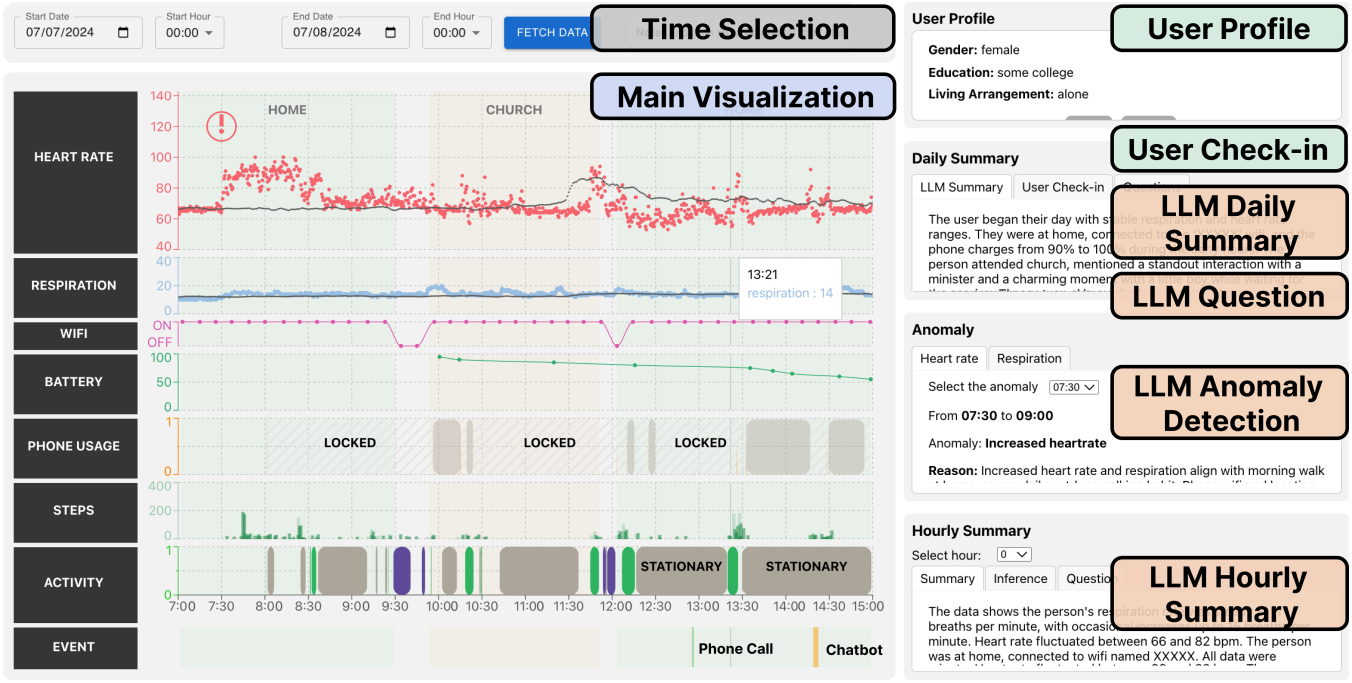

Figure 1: The Design of Vital Insight: Eight Modules.

In longitudinal studies that span months or even years, assistance in quickly navigating to the desired information seemed crucial. Experts noted a significant lack of tools for this purpose, as the sensemaking process is complex and often beyond the capabilities of current algorithms and tools. Some experts who used machine learning in their studies mentioned that they only viewed benchmarks as "results to be compared" against ground truth data, rather than as useful assistance for them. They expressed interest in getting some extracted data insights that go beyond simple labels. Based on these insights, we designed and developed our prototype.

3 Design of Vital Insight

In this section, we present the proposed design of our prototype, **Vital Insight** (VI), based on the insights from the interviews.

The input for VI could be various data types commonly used in real-world passive sensing deployments from smartphones, wearables, and voice assistants/chatbots. These data types include time-series data (e.g., physiological signals), discrete data (e.g., phone unlock states), and self-reported data (e.g., conversation with chatbots) – all commonly used data elements in ubiquitous computing studies that involve passive sensing [18, 32]. We built the system using React, with a MongoDB backend hosted on a local server.

3.1 Interface design

Our initial prototype has two major parts: (1) visualizations that provide direct representation of sensor data, and (2) varying granularities of summaries generated by LLM as indirect inference.

The **Time Selection** panel on the top left allows users to select the start and end date/time to zoom in/out. The **Main Visualization** panel serves as the central component to represent five categories of information gathered from the phone, smartwatch, and chatbot: location, health, phone usage, activity, and events. We used a small multiple-style time series plot arrangement to display each category of information along a unified timeline [3, 29, 42], a method particularly effective for tasks that involve direct visual comparisons of time series data, aiding in comparing, exploring, and analyzing trends [2, 5, 20, 22, 34]. The heart rate and respiration data are positioned at the top, followed by phone usage details and activity levels. Each modality is plotted in distinct colors and formats corresponding to its data type. Time series data are presented using scatter plots. Wi-Fi connection (0/1) is displayed using a purple line chart. Battery levels are displayed using a line chart, ranging from 0 to 100. Screen unlock periods are highlighted with rectangular overlays to indicate active phone usage. Step counts are depicted using a bar chart, alongside activities detected by the phone represented by rectangles('stationary' in gray, 'walking' in green, 'automotive' in purple). Location labels derived from frequently visited addresses are color-coded as the background. Finally, the Event section includes significant events throughout the day, including phone calls (green) and interactions with the Alexa chatbot (orange), and their respective durations. Interactive features include hovering to view data points and selecting different time ranges for detailed examination. The **User Profile** provides demographic information of the person. **LLM Daily Summary** panel, provides a high-level overview of the participant's day, including LLM Summary - an LLM-generated summary of the entire day in paragraphs and bullet points, User Check-in - conversation between the chatbot and the participant, and Question - a list of questions and missing information that the system identifies as potentially helpful for data interpretation. The **LLM Anomaly** panel provides insights

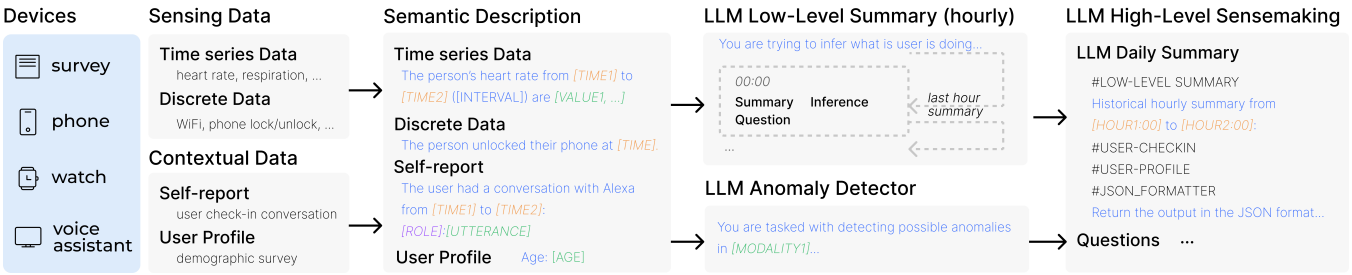

Figure 2: Overview of the structure of LLM augmentation for Vital Insight.

about unusual data points, and consists of specific time ranges, descriptions, and possible reasons of each anomaly generated by LLM. Similar to the daily summary, **LLM Hourly Summary** provides interpretations for each hour of data in detail.

3.2 LLM Augmentation

We employ a multi-level LLM augmentation framework to generate high-level insights for Vital Insight. For **Sensing Data**, we process the data into **semantic description**. Discrete data, such as WiFi and phone battery levels, are directly converted into sentences with values and timestamps (*e.g. "The battery level of the person's phone is [BATTERY] at [TIME]."*). For time-series data like heart rate, encoding each data point individually would overwhelm our system. Instead, since these data are typically sampled at a fixed rate, we group them and send them in an array format like *"The person's respiration from [TIME] to [TIME] (10-second interval) is [VALUE1, VALUE2, ...]."*, and sort this narrative in chronological order. Another data input is **Contextual data** which includes User Profile - demographic and routine information, and User Check-in - which we process conversation in a time*(From [TIME] to [TIME])* and utterance *([ROLE]:[UTTERANCE])* format.

After transferring raw data into semantic descriptions, the prototype generates **LLM low-level summary**, which processes the most granular level of data and provides an hourly summary. The prompts consist of the following components: Goal, Data Interpretation Guidance, Data, User Profile, User Check-in, Historical Summary from GPT, and a JSON Formatter. The prototype then processes the low-level summary into **LLM High-Level sensemaking**. Recognizing the daily patterns in activity, it aggregates the outputs from 24 hours of the hourly summary and other contextual information, to generate summary and inference on a daily basis, and format the results in a human-readable way. Other than summary from the hourly data, **LLM anomaly detector** identifies potential anomalies in different modalities directly using the semantic descriptions of each sensing modality. After identifying potential anomalies, it then provides explanations of these anomalies by combining contextual information.

4 Exploratory User Study

We conducted preliminary user testing with 13 experts using VI.

4.1 Method

4.1.1 Data collection. To generate the dashboard, we used real-world data from a deployment with participants, where we collected: 1) pre-study surveys, 2) passive mobile and wearable sensing data, and 3) voice assistant check-ins. Participants completed surveys on demographics and regular routines before deployment. We provided the participants with a Garmin smartwatch and installed the study app on their phones to collect physical activity, location, call logs, app usage, Wi-Fi/Bluetooth connections, IBI, heart rate, accelerometer, step counts and more. Additionally, participants can initiate a chit-chat conversation with an LLM-powered Amazon Alexa to check in on their day. The system runs in the background continuously, and we used a secure, HIPAA-compliant GPT-4 model and deleted identifiable data to ensure that neither GPT-4 nor the participants could identify the subject of data collection. The individuals who provided data consented to share it with external researchers through an IRB-approved study.

4.1.2 Study Design. At the beginning of the session, experts watched a brief tutorial on VI and shared their screens. Experts then looked at the data and wrote a short summary of the person's day based on the data, and provided general feedback on the prototype. Experts rated each module based on its impact, usage, trustworthiness, and clarity.

4.2 Results

4.2.1 Survey results. Regarding the **impact** on sensemaking, Visualization was the most crucial component, while Hourly Summary was the least important. The other modules (User Check-in, Daily Summary, and Anomaly) were similarly important. For **trust** (1 = Distrust to 5 = Trust), experts trusted User Check-ins (Mean 4.37, SD=0.47) and Visualization (4.32, SD=0.50) the most, followed by Daily Summary (3.91, SD=0.75), Hourly Summary (3.73, SD=0.69), and Anomaly (3.69, SD=1.03). The system's overall trustworthiness score was 4.09 (SD=0.70), indicating general trust, with slightly lower trust in the LLM-based inference modules. The lower trust in the Anomaly module may be due to differing definitions and vague guidance, which will be discussed later in section 4.2.2. Next, we examined the **usage** and **clarity** of each data view in the visualization. Experts used an average of 6.23 data views (SD=1.01). Heart rate, steps, and activity were used by all experts, followed by respiration, WiFi, battery, and phone unlock usage. The average clarity score across all modalities was 4.29 (SD=0.37), indicating that most data views were clear and actively used during sensemaking.

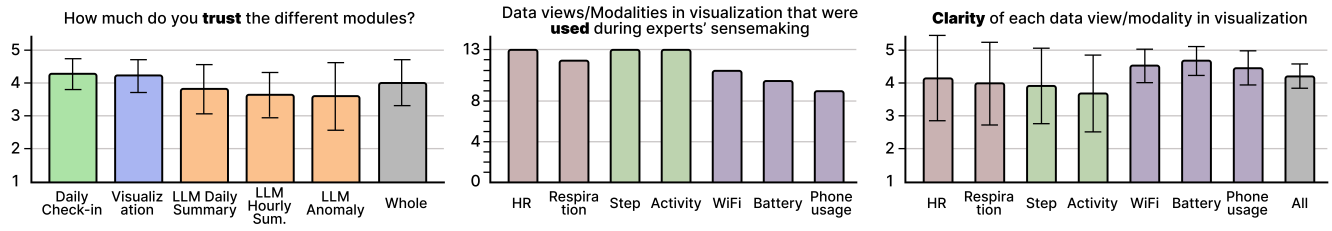

Figure 3: Survey results from the user testing.

4.2.2 General usability and qualitative feedback. Experts generally liked the prototype: *"I wish I had this for my study." (P4)*

Experts loved the visualization aspect of the prototype. For example, P3 specifically mentioned that they appreciated how the steps from both the phone and watch were plotted together. Further, experts found the User Profile module helpful (*"The user profile, that's very powerful(P1)"*). Similarly, experts loved the user check-in module and wanted it to be more prominent on the dashboard. Some experts like P1 suggested extracting key information from the user check-in panel and displaying it on the visualization according to the time. For example, P1 suggested it would be helpful *"if the dashboard said, Oh, this is when the little boy sat. This is when the brunch happened."* as an addition to the original paragraph.

Experts actively used the LLM features. Many experts found the LLM-generated summaries to be *"good"*, *"beyond expectations."*, which aligns with the survey results. They appreciated the rich details in LLM-generated summaries, but still value a simple summary, like *"this is a normal Sunday,"* as a starting point. Experts enjoyed reading the different possibilities of events provided by the LLM, and liked the current ways where evidence is generally separated with inference: *"I don't mind making some guess, but I wanted to know where that info (from LLM) is coming from." (P5).*

Another interesting insight emerged around the differing **definitions of** what should be considered an **anomaly**. Many experts noted that the current anomalies are often more like *"standout events"* rather than true anomalies. This highlights a potential need for the prototype to distinguish more clearly between special events or highlights and genuine anomalies that experts should be concerned about. In summary, both the qualitative and quantitative results indicate that experts had a positive experience using VI.

5 Discussion

5.1 Biases in LLM-generated inferences

During user testing with experts, we observed them gradually building trust in the LLM-generated inferences and summaries while exploring the various system components. However, this process also increased the potential risk of bias introduced by inaccurate LLM results. Several prior works have discussed concerns with bias in LLM-generated outcomes: highlighting biases related to gender [44], age [12], geography [33], and politics [41]. These biases could influence the interpretation of contextual information about individuals. For example, LLM may offer a stereotyped interpretation of the data based on the user's certain demographics. LLMs also exhibit bias in their information retrieval processes, such as giving preference to items from specific input positions like the beginning or end of the list [10, 11, 31], which could affect the retrieval of raw sensing data. Experts in out study generally expressed a highly skeptical view of the LLM-generated results, often validating the inferences before accepting them. However, as their trust in the system grew, there is a risk that they might be increasingly influenced by the LLM's outputs, even for experts [7, 24, 39]. During our internal evaluation of the LLM results, we also noted a potential concern with over-interpreting sensor data, such as noticing trivial fluctuations in heart rate and forcing itself to make an inference. Without explicit instructions, LLMs tend to not question the data reliability and always offer a potential explanation regardless. Future research must recognize the importance of examining both the LLM's biases in interpreting personal tracking data and the biases experts might develop when integrating those results.

5.2 Understanding tracking data using LLMs

Some recent works also trying to interpret passive tracking data using LLM. HARGPT employs simple role-playing prompts without expert guidance to identify activities from raw IMU data [23]. Health-LLM evaluated zero/few-shot learning and fine-tuning with simple prompts on sensor data [25]. LLMSense and PhysioLLM use text-formatted sensor data and prompts to derive high-level inferences from sensor traces [14, 38]; Cosetino et al. incorporated domain knowledge into prompts and fine-tuned models using expert responses [9]. We build on previous works to provide an iterative and evolving approach to continuously include domain expertise to guide LLM in generating meaningful insights compared to brute-force RAG techniques. We aim to encourage future studies to conduct a more thorough comparison of the accuracy and usefulness of various methods through human-in-the-loop LLMs.

6 Conclusion

In this study, we conducted interviews with experts and designed and developed Vital Insight, a prototype that provides both direct data representation and indirect inference through visualization and Large Language Models (LLMs). We conducted preliminary user testing sessions with 13 experts using Vital Insight and found good usability in both the visualization and LLM components.

Acknowledgements

This research is supported by the NSF AI-Caring Institute. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the sponsors.

References

[1] Daniel A Adler, Yuewen Yang, Thalia Viranda, Xuhai Xu, David C Mohr, Anna R Van Meter, Julia C Tartaglia, Nicholas C Jacobson, Fei Wang, Deborah Estrin, et al. Beyond detection: Towards actionable sensing research in clinical mental healthcare. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies*, 8(4):1–33, 2024.

[2] Benjamin Bach, Nathalie Henry-Riche, Tim Dwyer, Tara Madhyastha, J-D Fekete, and Thomas Grabowski. Small multipiles: Piling time to explore temporal patterns in dynamic networks. In *Comput. Graph. Forum*, volume 34, pages 31–40. Wiley Online Library, 2015.

[3] Louis Bavoil, Steven P Callahan, Patricia J Crossno, Juliana Freire, Carlos E Scheidegger, Cláudio T Silva, and Huy T Vo. Vistrails: Enabling interactive multiple-view visualizations. In *VIS 05. IEEE Visualization, 2005.*, pages 135–142. IEEE, 2005.

[4] George Boateng, Vivian Genaro Motti, Varun Mishra, John A Batsis, Josiah Hester, and David Kotz. Experience: Design, development and evaluation of a wearable device for mhealth applications. In *The 25th Annual International Conference on Mobile Computing and Networking*, pages 1–14, 2019.

[5] Ilya Boyandin, Enrico Bertini, and Denis Lalanne. A qualitative study on the exploration of temporal changes in flow maps with animation and small-multiples. In *Comput. Graph. Forum*, volume 31, pages 1005–1014. Wiley Online Library, 2012.

[6] Virginia Braun and Victoria Clarke. Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2):77, 2006.

[7] Avishek Choudhury and Zaira Chaudhry. Large language models and user trust: Consequence of self-referential learning loop and the deskilling of health care professionals. *Journal of Medical Internet Research*, 26:e56764, 2024.

[8] Juliet M Corbin and Anselm Strauss. Grounded theory research: Procedures, canons, and evaluative criteria. *Qualitative sociology*, 13(1):3–21, 1990.

[9] Justin Cosentino, Anastasiya Belyaeva, Xin Liu, Nicholas A Furlotte, Zhun Yang, Chace Lee, Erik Schenck, Yojan Patel, Jian Cui, Logan Douglas Schneider, et al. Towards a personal health large language model. *arXiv preprint arXiv:2406.06474*, 2024.

[10] Sunhao Dai, Chen Xu, Shicheng Xu, Liang Pang, Zhenhua Dong, and Jun Xu. Bias and unfairness in information retrieval systems: New challenges in the llm era. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 6437–6447, 2024.

[11] Sunhao Dai, Yuqi Zhou, Liang Pang, Weihao Liu, Xiaolin Hu, Yong Liu, Xiao Zhang, Gang Wang, and Jun Xu. Neural retrievers are biased towards llm-generated content. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 526–537, 2024.

[12] Yucong Duan. The large language model (llm) bias evaluation (age bias). *DIKWP Research Group International Standard Evaluation. DOI*, 10, 2024.

[13] Zachary Englhardt, Chengqian Ma, Margaret E Morris, Chun-Cheng Chang, Xuhai" Orson" Xu, Lianhui Qin, Daniel McDuff, Xin Liu, Shwetak Patel, and Vikram Iyer. From classification to clinical insights: Towards analyzing and reasoning about mobile and behavioral health data with large language models. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(2):1–25, 2024.

[14] Cathy Mengying Fang, Valdemar Danry, Nathan Whitmore, Andria Bao, Andrew Hutchison, Cayden Pierce, and Pattie Maes. Physiollm: Supporting personalized health insights with wearables and large language models. *arXiv preprint arXiv:2406.19283*, 2024.

[15] Nicholas Farber, Douglas Shinkle, Jana Lynott, Wendy Fox-Grage, and Rodney Harrell. Aging in place: A state survey of livability policies and practices. 2011.

[16] Emilio Ferrara. Large language models for wearable sensor-based human activity recognition, health monitoring, and behavioral modeling: A survey of early trends, datasets, and challenges. *arXiv preprint arXiv:2407.07196*, 2024.

[17] Michael Friedewald and Oliver Raabe. Ubiquitous computing: An overview of technology impacts. *Telematics and Informatics*, 28(2):55–65, 2011.

[18] Thomas Fritz, Elaine May Huang, Gail C. Murphy, and Thomas Zimmermann. Persuasive technology in the real world: a study of long-term use of activity sensing devices for fitness. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2014.

[19] Barney G Glaser, Anselm L Strauss, and Elizabeth Strutzel. The discovery of grounded theory; strategies for qualitative research. *Nursing research*, 17(4):364, 1968.

[20] Jeffrey Heer, Nicholas Kong, and Maneesh Agrawala. Sizing the horizon: the effects of chart size and layering on the graphical perception of time series visualizations. In *Proc. SIGCHI Conf. Human Factors*, pages 1303–1312, 2009.

[21] Joyce Ho and Stephen S Intille. Using context-aware computing to reduce the perceived burden of interruptions from mobile devices. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 909–918, 2005.

[22] Waqas Javed, Bryan McDonnel, and Niklas Elmqvist. Graphical perception of multiple time series. *IEEE Trans. Vis. Comput. Graphics*, 16(6):927–934, Nov.-Dec. 2010.

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

[24] Charalampia Xaroula Kerasidou, Angeliki Kerasidou, Monika Buscher, and Stephen Wilkinson. Before and beyond trust: reliance in medical ai. *Journal of medical ethics*, 48(11):852–856, 2022.

[25] Yubin Kim, Xuhai Xu, Daniel McDuff, Cynthia Breazeal, and Hae Won Park. Health-llm: Large language models for health prediction via wearable sensor data. *arXiv preprint arXiv:2401.06866*, 2024.

[26] Evan King, Haoxiang Yu, Sangsu Lee, and Christine Julien. Sasha: creative goal-oriented reasoning in smart homes with large language models. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(1):1–38, 2024.

[27] Anastasia Kononova, Lin Li, Kendra Kamp, Marie Bowen, RV Rikard, Shelia Cotten, Wei Peng, et al. The use of wearable activity trackers among older adults: focus group study of tracker perceptions, motivators, and barriers in the maintenance stage of behavior change. *JMIR mHealth and uHealth*, 7(4):e9832, 2019.

[28] Florian Künzler, Varun Mishra, Jan-Niklas Kramer, David Kotz, Elgar Fleisch, and Tobias Kowatsch. Exploring the state-of-receptivity for mhealth interventions. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 3(4):1–27, 2019.

[29] Fritz Lekschas, Benjamin Bach, Peter Kerpedjiev, Nils Gehlenborg, and Hanspeter Pfister. Hipiler: visual exploration of large genome interaction matrices with interactive small multiples. *IEEE Trans. Vis. Comput. Graphics*, 24(1):522–531, Jan. 2018.

[30] Jiachen Li, Bingrui Zong, Tingyu Cheng, Yunzhi Li, Elizabeth D Mynatt, and Ashutosh Dhekne. Privacy vs. awareness: Relieving the tension between older adults and adult children when sharing in-home activity data. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW2):1–30, 2023.

[31] Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. Halueval: A large-scale hallucination evaluation benchmark for large language models. *arXiv preprint arXiv:2305.11747*, 2023.

[32] Antonio Alfredo Ferreira Loureiro. Sensing, tracking and contextualizing entities in ubiquitous computing. In *International Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems*, 2012.

[33] Rohin Manvi, Samar Khanna, Marshall Burke, David Lobell, and Stefano Ermon. Large language models are geographically biased. *arXiv preprint arXiv:2402.02680*, 2024.

[34] Miriah Meyer, Bang Wong, Mark Styczynski, Tamara Munzner, and Hanspeter Pfister. Pathline: A tool for comparative functional genomics. In *Comput. Graph. Forum*, volume 29, pages 1043–1052. Wiley Online Library, 2010.

[35] Varun Mishra, Tian Hao, Si Sun, Kimberly N Walter, Marion J Ball, Ching-Hua Chen, and Xinxin Zhu. Investigating the role of context in perceived stress detection in the wild. In *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*, pages 1708–1716, 2018.

[36] Elizabeth D Mynatt, Jim Rowan, Sarah Craighill, and Annie Jacobs. Digital family portraits: supporting peace of mind for extended family members. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 333–340, 2001.

[37] Carsten Orwat, Andreas Graefe, and Timm Faulwasser. Towards pervasive computing in health care–a literature review. *BMC medical informatics and decision making*, 8:1–18, 2008.

[38] Xiaomin Ouyang and Mani Srivastava. Llmsense: Harnessing llms for high-level reasoning over spatiotemporal sensor traces. *arXiv preprint arXiv:2403.19857*, 2024.

[39] Mark Ryan. In ai we trust: ethics, artificial intelligence, and reliability. *Science and Engineering Ethics*, 26(5):2749–2767, 2020.

[40] Alexander Seifert, Anna Schlomann, Christian Rietz, and Hans Rudolf Schelling. The use of mobile devices for physical activity tracking in older adults' everyday life. *Digital health*, 3:2055207617740088, 2017.

[41] Aleksandra Urman and Mykola Makhortykh. The silence of the llms: Cross-lingual analysis of political bias and false information prevalence in chatgpt, google bard, and bing chat. 2023.

[42] Stef van den Elzen and Jarke J van Wijk. Small multiples, large singles: A new approach for visual data exploration. In *Comput. Graph. Forum*, volume 32, pages 191–200. Wiley Online Library, 2013.

[43] Dimitri Vargemidis, Kathrin Gerling, Katta Spiel, Vero Vanden Abeele, and Luc Geurts. Wearable physical activity tracking systems for older adults—a systematic review. *ACM Transactions on Computing for Healthcare*, 1(4):1–37, 2020.

[44] Yixin Wan, George Pu, Jiao Sun, Aparna Garimella, Kai-Wei Chang, and Nanyun Peng. " kelly is a warm person, joseph is a role model": Gender biases in llm-generated reference letters. *arXiv preprint arXiv:2310.09219*, 2023.

[45] Bufang Yang, Siyang Jiang, Lilin Xu, Kaiwei Liu, Hai Li, Guoliang Xing, Hongkai Chen, Xiaofan Jiang, and Zhenyu Yan. Drhouse: An llm-empowered diagnostic reasoning system through harnessing outcomes from sensor data and expert knowledge. *arXiv preprint arXiv:2405.12541*, 2024.

A Appendix

A.1 Participants' demographic

Table 1: Demographics of experts participating in the user studies.

Participant	Gender	Education	Years of Experience
P1	M	Doctorate	8 years
P2	F	Doctorate	1 year
P3	F	Master	3 years
P4	F	Bachelor	4 years
P5	F	Master	4 years
P6	M	Doctorate	18 years
P7	M	Master	2 years
P8	M	Master	8 years
P9	M	Master	1 year
P10	F	Doctorate	30 years
P11	F	Master	2 years
P12	M	Doctorate	7 years
P13	M	Doctorate	5 years