Personalised Explainable Robots Using LLMs

Ferran Gebellí
PAL Robotics
Barcelona, Spain
ferran.gebelli@pal-robotics.com

Lavinia Hriscu
IRI (UPC-CSIC)
Barcelona, Spain
lavinia.beatrice.hriscu@upc.edu

Raquel Ros
PAL Robotics
Barcelona, Spain
raquel.ros@pal-robotics.com

Séverin Lemaignan
PAL Robotics
Barcelona, Spain
severin.lemaignan@pal-robotics.com

Alberto Sanfeliu
IRI (UPC-CSIC)
Barcelona, Spain
alberto.sanfeliu@upc.edu

Anaís Garrell
IRI (UPC-CSIC)
Barcelona, Spain
anais.garrell@upc.edu

***Abstract*—In the field of Human-Robot Interaction (HRI), a key challenge lies in enabling humans to comprehend the decisions and behaviours of robots. One promising approach involves leveraging Theory of Mind (ToM) frameworks, wherein a robot estimates the mental model that a user holds about its functioning and compares this with the representation of its internal mental model. This comparison allows the robot to identify potential mismatches and generate communicative actions to bridge such gaps. Effective communication requires the robot to maintain unique mental models for each user and personalise explanations based on past interactions. To address this, we propose an architecture grounded in Large Language Models (LLMs) that operationalises this theoretical framework. We demonstrate the feasibility of this approach through qualitative examples, showcasing responses provided by a robot patrolling a geriatric hospital.**

***Index Terms*—XHRI, explainability, personalisation, LLM**

I. Introduction

In the domain of Human-Robot Interaction (HRI), explainability is frequently regarded as a crucial mechanism for enhancing human understanding of robotic behaviours and decision-making processes [1], [2].

From the perspective of *Theory of Mind (ToM)*, eXplainable HRI (XHRI) has been conceptualised as a model reconciliation problem [3], [4], [5]. Within these theoretical frameworks, a human is assumed to possess a mental model of the robot, expressed as $M_{\text{human}}(M_{\text{robot}})$. Similarly, the robot's state-of-mind model, represented as M_{robot}, encompasses an estimate of the human's mental model of the robot, captured as $M_{\text{robot}}(M_{\text{human}}(M_{\text{robot}}))$. In this context, the objective of explainability is to minimise the discrepancy between the robot's mental model, M_{robot}, and the robot's estimate of the human's perception of the robot's mental model, $M_{\text{robot}}(M_{\text{human}}(M_{\text{robot}}))$, as depicted in Figure 1.

Because robots should estimate separate models for each human, personalisation becomes crucial [6], [7]. However, the review conducted by [8] highlights that few studies have addressed the issue of personalisation in XHRI. In response to this gap, we propose a novel methodology for delivering personalised explanations in HRI, which leverages the capabilities of Large Language Models (LLMs).

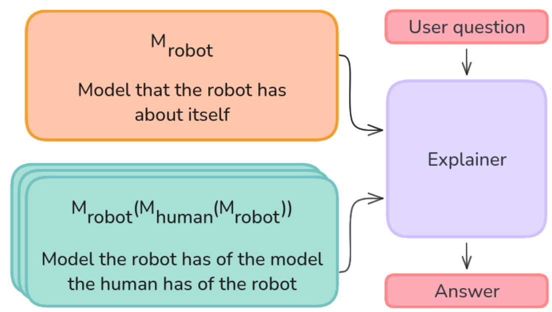

Figure 1. Explanations as reconciliation between the robot internal model and the estimate of the human model about the robot.

II. Related Work

This section is organised as follows: first, we review the concept of personalisation within the field of robotics. Next, we examine personalised, LLM-based chatbots in non-robotic domains. Subsequently, we explore approaches to personalised explainability in non-robotic systems. Finally, we review the application of LLM-based explanations in robotics, highlighting the absence of personalisation in these implementations.

a) Personalised robots: In HRI, the modelling of users and personalisation have been widely treated [9], but mostly as a way to adapt the decision-making and behaviours, either with rather classical approaches [10], [11] or recently with LLMs [12], [13], [14]. However, personalisation in HRI has not been studied from the perspective of adapting dialogues to explain the reasons behind those decisions and behaviours.

b) Personalised LLM conversations: LLMs are a promising tool for personalising conversations. Approaches exploiting user context through prompting mechanisms have been proposed in [12], [15]. To address the limitation of managing extensive historical conversations, alternative strategies retrieve relevant excerpts from prior conversations [16], [17]. Another widely adopted method involves fine-tuning Large Language Models (LLMs) based on past interactions [18], [19], [20] or constructing user-specific embeddings [21]. Additionally, external memory systems have been explored, including dictionary-based approaches [22] and human-supervised frameworks [23].

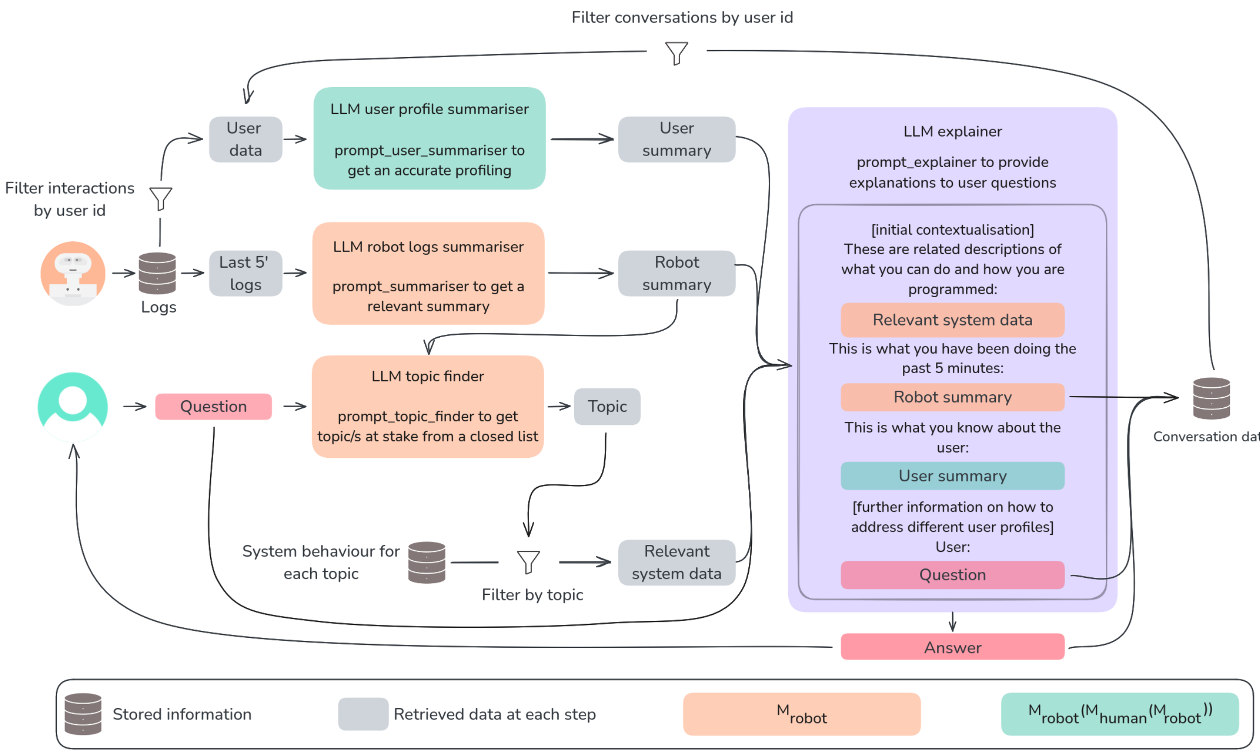

Figure 2. Complete architecture for processing the user's question to provide an answer using the combination of Large Language Models.

c) Personalised explainability: In [24] explanations are personalised to clustered types of users. Other works do not personalise to user types, but to specific individuals [25], [26], according to user preferences [27], or combining preferences with other metrics, such as performance [28]. However, in these works there is no interaction with robots, so they do not address challenges that are specific only to XHRI: (1) The M_{robot} is especially complex due to the extended capabilities coming from the embodiment; (2) The robot's anthropomorphism involves that the $M_{human}(M_{robot})$ often starts from wrong and high expectations; (3) The robot's sensors allow a more accurate $M_{robot}(M_{human}(M_{robot}))$ that goes beyond the knowledge acquired from conversations only, expanding on situated information such as humans' positions, gestures or expressions; and (4) The robots' multi-modal communicative capabilities can extend conversations with gestures, displayed information or actions.

d) LLM-based explanations for robots: In [29], prompts are enhanced by integrating the robot's behaviour trees. Some other studies employ retrieval techniques that leverage the robot's logs [30] to address the limitations imposed by excessively long contexts. This approach is further refined in [31] through the incorporation of a Visual Language Model (VLM). Additionally, other works propose the generation of episodic memories based on the automatic detection of key events, capturing both the robot's internal states and relevant world information [32], [33].

Motivated by the literature on LLM-based explanations for robots, this work will include the personalisation component inspired by the other subsections above.

III. LLM-BASED FRAMEWORK

The architecture developed in this work aims to enhance interpretability by using structured knowledge, with the purpose of explaining decision-making and actions. Given that LLMs inherently lack the capability to maintain persistent world models [34], our approach incorporates multiple specialised models, as illustrated in Figure 2. By providing such models with the internal state of the robot, the environment and the user data, one can obtain a representation of M_{robot} and $M_{robot}(M_{human}(M_{robot}))$. The system eventually answers the user by reconciling both representations. The model used is *gpt-4o-mini* from OpenAI.

A. M_{robot} representation

The robot's representation of itself combines two types of information: a summary of past actions and a set of predefined behaviour rules classified by different topics or situations. Two LLMs retrieve this data.

1) Past actions summary: The robot activity logs are stored in a database, with each log entry timestamped to record the precise time of each action. Under the assumption that users will ask about the very recent past, the logs from a 5-minute time window are selected and summarised by an LLM for every new user input. While the logs are comprehensive, not all information is useful nor relevant. Therefore, only high-level logs related to the robot's state, failures and user interactions are used as input for the LLM. Specific instructions are given to include certain actions in the summary, if present in the logs. This ensures the output focuses on relevant actions for the user, while ignoring less insightful data.

2) Topic finder: Conversations between the user and the robot assistant can cover a variety of topics. To provide more relevant and accurate responses, it is crucial to first identify the specific topic the user is addressing. This topic is identified using an LLM, which processes the summary of events from the preceding 5 minutes along with the user question. It is important to remark that multiple topics may be associated with both the conversation logs and the query, allowing for more nuanced understanding and response generation.

3) Relevant system data: The robot is programmed with specific behaviour guidelines for each situation it encounters. These protocols are encoded in natural language by the system developers, classified by conversation topics, and stored in a database. Once the LLM identifies the topic(s) of conversation, the corresponding predefined behaviour rules are retrieved.

B. $M_{robot}(M_{human}(M_{robot})$ representation

The robot's understanding of how it is perceived by the user is modelled using an LLM, which profiles the user based on their past interactions.

1) User data: All interactions between the user and the robot, specifically questions and answers, are stored in a database along with their corresponding timestamps. In addition to the dialogue, the summaries of log entries from the 5-minute window prior to every interaction are also retained.

Furthermore, task-based interactions, such as user engagement with robot graphical interfaces, are recorded and timestamped, as these are also pertinent for user profiling.

2) User profiling: An LLM processes all the data of the user, taking the current date as a temporal reference and the relevant topics for the ongoing conversation. The model then generates a concise summary of the user knowledge level, noting any potential gaps—such as a period of inactivity lasting at least two weeks—that may affect the depth and clarity of the responses.

C. LLM explainer

To answer the user questions, the system combines *static knowledge*, which contains the behaviour rules according to the topic(s) of the conversation, and *dynamic knowledge*, a summary of the logs from the last 5 minutes, along with guidelines for clear responses. To enable personalisation, the user profile is also included, which reflects the knowledge level of the user and the time elapsed since its last engagement.

IV. Use Case Instances

We evaluate our approach using a robot that autonomously patrols the geriatric unit of a hospital, as shown in Figure 3. The robot's primary task is to monitor the patients' rooms and alert the nursing staff via a mobile app in the event of potentially hazardous situations. The available alerts include instances where a person has fallen on the floor, a vulnerable individual is standing alone in a room, or a door is closed when it should remain open. Additionally, the robot can trigger alarms in case of system failures, such as being blocked. The nursing staff can schedule patrols for specific subsets of patient rooms within desired timeframes.

Figure 3. Use case of a robot patrolling a hospital.

During two months, we recorded the high-level logs of the robot, which included information on schedule changes, behaviour state transitions, and alarm triggers along with their corresponding reasons. Interactions through the mobile app from nursing staff were also logged, with each staff member using a unique account, enabling us to track who responded to each alarm.

We utilise these logs to simulate interactions where different users pose specific explainability questions at designated time points throughout the two months. Those questions are related to alarms that the robot has triggered or missed. The logs are fast-forwarded to these time points, assuming that users always inquire about events that have just occurred. In the subsections below, we present the generated personalised explanations for 3 different user types.

A. Non-active user

We first simulate that a user who has not yet been actively interacting or conversing with the system asks a question about a missed "standing patient" alarm. The user's summary correctly detects that the user has a lack of knowledge about the topic, and therefore the user receives a detailed answer that explains extensively the reasons behind not triggering the alarm. Figure 4 shows the user summary and response generated by the LLM framework.

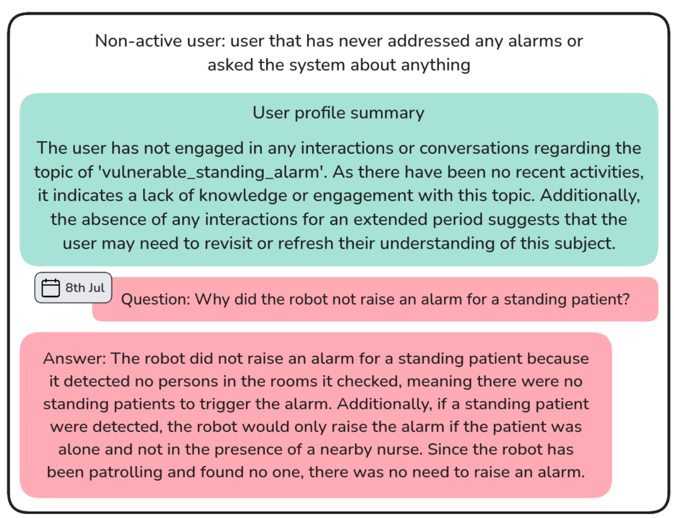

Figure 4. User summary and response for a non-active user.

B. Active user

For an active user, we first simulate that the user makes a question related to the "closed door" alarm, and then we fast-forward the logs to one week later and simulate that the same user, who has interacted with the robot during the week, asks the same question. In Figure 5, we present the response of the system for the second time the question is asked. The user summary accurately detects that the user is very engaged with the system, thus providing a concise answer.

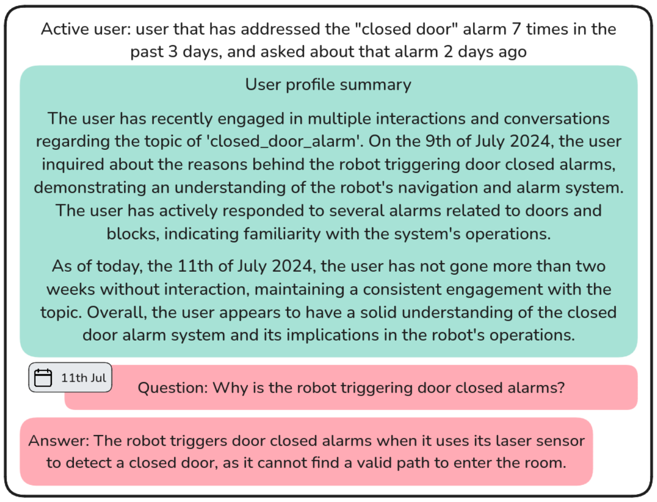

Figure 5. User summary and response for an active user.

C. Active user with time gap

Figure 6 includes the user summary and responses for a user that had been active with the "robot blocked" alarm, addressing it 4 times and asking the system about that topic, but who has not interacted at all with the robot for more than one month due to a holiday period. The user summary not only acknowledges the past activity but also detects the time elapsed, thus providing a detailed answer referencing earlier discussions on the topic.

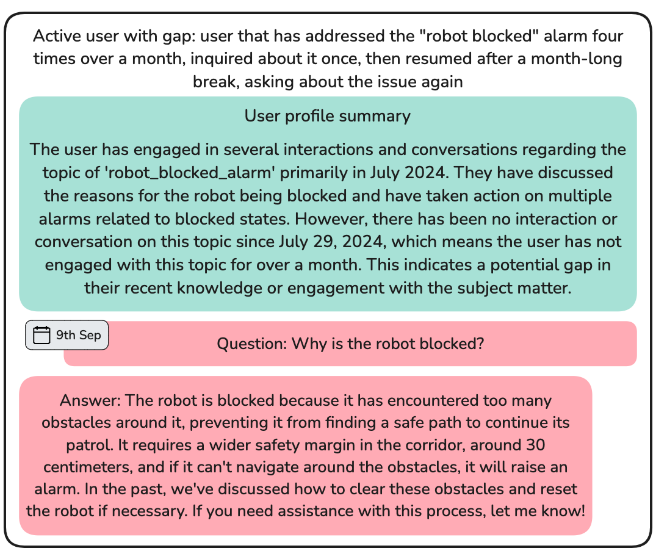

Figure 6. User summary and response for an active user with a time gap.

V. Future work

In a future user study, we will compare ablations of the representaition, updating and retrieval of M_{robot} and $M_{robot}(M_{human}(M_{robot})$, while integrating multi-modal inputs. We also foresee exploring Retrieval Augmented Generation (RAG), which would allow the 5-minute window's suppression. Moreover, we have used a dataset to run a simulation, so future work should ensure real-time requirements by parallelising some steps and comparing LLM models from different providers. The current pipeline takes 4 to 5 seconds between a question and an answer. Below we detail further directions that add new features.

A. Reinforcement learning from human feedback

Since our goal is to provide users with better explanations about the robot, it is essential to gather user feedback with respect to such explanations. One potential approach to achieve a deeper personalisation is to fine-tune LLMs using reinforcement learning from human feedback.

B. Uncertainty awareness

In future research, we would like to explore how to incorporate in M_{robot} and $M_{robot}(M_{human}(M_{robot})$ an estimate of the uncertainty in the models and instigate the LLM to respond accordingly. This way answers such as: "I am not sure if you already know this, but [...]" could let the user know that the system is not completely confident.

C. Explanations to heterogenous and changing groups

We would like to address the challenge of explanations to changing groups of people. Suppose the robot first explains something to Bob, and then Alice comes and the robot explains something else to Alice, while Bob is still there. In those cases, the robot should extend $M_{robot}(M_{human}(M_{robot})$ of a given listener with information exposed to other users.

VI. Conclusion

In this work, we introduced an architecture based on LLMs that delivers personalised explanations regarding the robot's decisions and behaviours. By utilising summaries from previous conversations and other interactions with the robot, the system provides tailored explanations. The user summary is integrated with a summary of the robot's high-level logs and relevant system data about the specific topic. We implemented our framework in the context of a robot patrolling a geriatric hospital and discussed potential directions for future research.

Ethics statement

The ethical committee from the hospital where the data was collected allowed the execution of the study and participants gave their consent.

Acknowledgment

This work was supported by Horizon Europe Marie Skłodowska-Curie TRAIL project (grant 101072488), the Horizon Europe CoreSense project (grant 10107025), the JST Moonshot R&D project (grant JPMJMS2011-85) and the LENA Spanish national project (PID2022-142039NA-I00).

References

[1] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artificial intelligence*, vol. 267, pp. 1–38, 2019.

[2] R. S. Verhagen, M. A. Neerincx, and M. L. Tielman, "A two-dimensional explanation framework to classify AI as incomprehensible, interpretable, or understandable," in *International Workshop on Explainable, Transparent Autonomous Agents and Multi-Agent Systems*, pp. 119–138, Springer, 2021.

[3] T. Hellström and S. Bensch, "Understandable robots-what, why, and how," *Paladyn, Journal of Behavioral Robotics*, vol. 9, no. 1, pp. 110–123, 2018.

[4] S. Sreedharan, T. Chakraborti, and S. Kambhampati, "Foundations of explanations as model reconciliation," *Artificial Intelligence*, vol. 301, p. 103558, 2021.

[5] T. Sakai and T. Nagai, "Explainable autonomous robots: A survey and perspective," *Advanced Robotics*, vol. 36, no. 5-6, pp. 219–238, 2022.

[6] H. Felzmann, E. Fosch-Villaronga, C. Lutz, and A. Tamo-Larrieux, "Robots and transparency: The multiple dimensions of transparency in the context of robot technologies," *IEEE Robotics & Automation Magazine*, vol. 26, no. 2, pp. 71–78, 2019.

[7] R. Setchi, M. B. Dehkordi, and J. S. Khan, "Explainable robotics in human-robot interactions," *Procedia Computer Science*, vol. 176, pp. 3057–3066, 2020.

[8] S. Anjomshoae, A. Najjar, D. Calvaresi, and K. Främling, "Explainable agents and robots: Results from a systematic literature review," in *18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), Montreal, Canada, May 13–17, 2019*, pp. 1078–1088, International Foundation for Autonomous Agents and Multiagent Systems, 2019.

[9] S. Rossi, F. Ferland, and A. Tapus, "User profiling and behavioral adaptation for HRI: A survey," *Pattern Recognition Letters*, vol. 99, pp. 3–12, 2017.

[10] E. B.-N. Sanders and P. J. Stappers, "Co-creation and the new landscapes of design," *Co-design*, vol. 4, no. 1, pp. 5–18, 2008.

[11] A. Andriella, C. Torras, and G. Alenyà, "Learning robot policies using a high-level abstraction persona-behaviour simulator," in *2019 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pp. 1–8, IEEE, 2019.

[12] J. Wu, R. Antonova, A. Kan, M. Lepert, A. Zeng, S. Song, J. Bohg, S. Rusinkiewicz, and T. Funkhouser, "Tidybot: Personalized robot assistance with large language models," *Autonomous Robots*, vol. 47, no. 8, pp. 1087–1102, 2023.

[13] Y. Dai, R. Peng, S. Li, and J. Chai, "Think, act, and ask: Open-world interactive personalized robot navigation," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3296–3303, IEEE, 2024.

[14] D. Han, T. McInroe, A. Jelley, S. V. Albrecht, P. Bell, and A. Storkey, "Llm-personalize: Aligning llm planners with human preferences via reinforced self-training for housekeeping robots," *arXiv preprint arXiv:2404.14285*, 2024.

[15] H. Wang, R. Wang, F. Mi, Y. Deng, Z. Wang, B. Liang, R. Xu, and K.-F. Wong, "Cue-cot: Chain-of-thought prompting for responding to in-depth dialogue questions with llms," *arXiv preprint arXiv:2305.11792*, 2023.

[16] A. Salemi, S. Kallumadi, and H. Zamani, "Optimization methods for personalizing large language models through retrieval augmentation," in *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 752–762, 2024.

[17] C. Richardson, Y. Zhang, K. Gillespie, S. Kar, A. Singh, Z. Raeesy, O. Z. Khan, and A. Sethy, "Integrating summarization and retrieval for enhanced personalization via large language models," *arXiv preprint arXiv:2310.20081*, 2023.

[18] S. Woźniak, B. Koptyra, A. Janz, P. Kazienko, and J. Kocoń, "Personalized large language models," *arXiv preprint arXiv:2402.09269*, 2024.

[19] A. Salemi, S. Mysore, M. Bendersky, and H. Zamani, "Lamp: When large language models meet personalization," *arXiv preprint arXiv:2304.11406*, 2023.

[20] Z. Tan, Z. Liu, and M. Jiang, "Personalized pieces: Efficient personalized large language models through collaborative efforts," *arXiv preprint arXiv:2406.10471*, 2024.

[21] J. Liu, Y. Zhu, S. Wang, X. Wei, E. Min, Y. Lu, S. Wang, D. Yin, and Z. Dou, "Llms+ persona-plug= personalized llms," *arXiv preprint arXiv:2409.11901*, 2024.

[22] A. Madaan, N. Tandon, P. Clark, and Y. Yang, "Memory-assisted prompt editing to improve gpt-3 after deployment," *arXiv preprint arXiv:2201.06009*, 2022.

[23] H. Joko, S. Chatterjee, A. Ramsay, A. P. De Vries, J. Dalton, and F. Hasibi, "Doing personal laps: Llm-augmented dialogue construction for personalized multi-session conversational search," in *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 796–806, 2024.

[24] U. Soni, S. Sreedharan, and S. Kambhampati, "Not all users are the same: Providing personalized explanations for sequential decision making problems," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 6240–6247, IEEE, 2021.

[25] L. Li, Y. Zhang, and L. Chen, "Personalized prompt learning for explainable recommendation," *ACM Transactions on Information Systems*, vol. 41, no. 4, pp. 1–26, 2023.

[26] R. S. Verhagen, M. A. Neerincx, C. Parlar, M. Vogel, and M. L. Tielman, "Personalized agent explanations for human-agent teamwork: Adapting explanations to user trust, workload, and performance," in *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, pp. 2316–2318, 2023.

[27] K. Boggess, S. Chen, and L. Feng, "Towards personalized explanation of robot path planning via user feedback," *arXiv preprint arXiv:2011.00524*, 2020.

[28] A. Silva, P. Tambwekar, M. Schrum, and M. Gombolay, "Towards balancing preference and performance through adaptive personalized explainability," in *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, pp. 658–668, 2024.

[29] G. LeMasurier, C. Tagliamonte, J. Breen, D. Maccaline, and H. A. Yanco, "Templated vs. generative: Explaining robot failures," in *2024 33rd IEEE International Conference on Robot and Human Interactive Communication (ROMAN)*, pp. 1346–1353, IEEE, 2024.

[30] D. Sobrín-Hidalgo, M. A. González-Santamarta, Á. M. Guerrero-Higueras, F. J. Rodríguez-Lera, and V. Matellán-Olivera, "Explaining autonomy: Enhancing human-robot interaction through explanation generation with large language models," *arXiv preprint arXiv:2402.04206*, 2024.

[31] D. Sobrín-Hidalgo, M. Á. González-Santamarta, Á. M. Guerrero-Higueras, F. J. Rodríguez-Lera, and V. Matellán-Olivera, "Enhancing robot explanation capabilities through vision-language models: a preliminary study by interpreting visual inputs for improved human-robot interaction," *arXiv preprint arXiv:2404.09705*, 2024.

[32] Z. Liu, A. Bahety, and S. Song, "Reflect: Summarizing robot experiences for failure explanation and correction," *arXiv preprint arXiv:2306.15724*, 2023.

[33] Z. Wang, B. Liang, V. Dhat, Z. Brumbaugh, N. Walker, R. Krishna, and M. Cakmak, "I can tell what i am doing: Toward real-world natural language grounding of robot experiences," in *8th Annual Conference on Robot Learning*.

[34] K. Vafa, J. Y. Chen, J. Kleinberg, S. Mullainathan, and A. Rambachan, "Evaluating the world model implicit in a generative model," *arXiv preprint arXiv:2406.03689*, 2024.