# Personalised Explanations in Long-term Human-Robot Interactions

Ferran Gebellí*, Anaís Garrell†, Jan-Gerrit Habekost‡, Séverin Lemaignan*, Stefan Wermter‡ and Raquel Ros§

***Abstract*— In the field of Human-Robot Interaction (HRI), a fundamental challenge is to facilitate human understanding of robots. The emerging domain of eXplainable HRI (XHRI) investigates methods to generate explanations and evaluate their impact on human-robot interactions. Previous works have highlighted the need to personalise the level of detail of these explanations to enhance usability and comprehension. Our paper presents a framework designed to update and retrieve user knowledge-memory models, allowing for adapting the explanations' level of detail while referencing previously acquired concepts. Three architectures based on our proposed framework that use Large Language Models (LLMs) are evaluated in two distinct scenarios: a hospital patrolling robot and a kitchen assistant robot. Experimental results demonstrate that a two-stage architecture, which first generates an explanation and then personalises it, is the framework architecture that effectively reduces the level of detail only when there is related user knowledge.**

## I. INTRODUCTION

In eXplainable Human-Robot Interaction (XHRI), explainability is acknowledged as a key factor for improving human understanding of robots' behaviours and decision-making [19], [34]. From the perspective of *Theory of Mind (ToM)*, XHRI has been framed as a model reconciliation problem [9], [31]. Within these theoretical frameworks, a human is assumed to possess a mental model of the robot. In this context, the objective of explainability is to minimise the discrepancy between the robot's internal model of itself and the human's mental model of the robot, which the robot should estimate as precisely as possible.

Since robots must infer different models for each individual, personalisation becomes essential [27], but very few studies have explicitly addressed the challenge of personalisation in XHRI [2]. Although providing detailed explanations can improve human understanding, excessive detail can lead to cognitive overload and reduced attention [14]. Overly complex or redundant information can confuse users [21]; therefore, explanations should be appropriate and concise [20], as simpler explanations are generally more comprehensible [13]. In scenarios involving prolonged interactions, which has been identified as a common yet understudied setting in HRI [3], the need for clear explanations becomes even more critical. In such cases, personalisation is particularly important to adjust the level of detail and prevent redundant explanations of previously understood information.

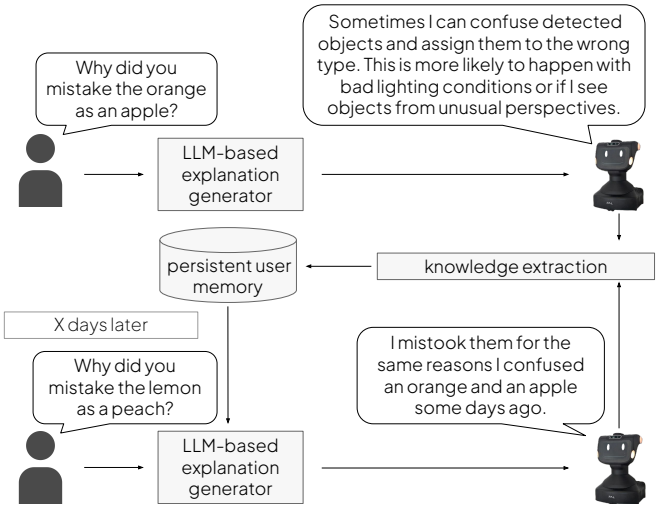


Fig. 1. Illustrative personalisation of an explanation's level of detail in a second interaction after some days elapsed since the first interaction.

In this study, we propose a framework that uses Large Language Models (LLMs) to generate personalised explanations for end users, as illustrated in Fig. 1. Our key contribution is the integration and retrieval of an external, user-specific memory that is continuously updated to track the user's estimated knowledge. We validate the effectiveness of the framework concerning the explanation's level of detail while comparing three architectures for querying and retrieving user-specific known concepts[1]. Future research will focus on evaluating the impact of these personalised explanations on user usability. We determine two research questions:

- **RQ1**: Can LLMs provide explanations about robot behaviours and decision-making that provide an adapted level of detail?
- **RQ2**: Which knowledge retrieval strategies are most effective in producing robust personalised explanations?

Our paper is organised as follows. Sec. II reviews the related literature. Sec. III details the proposed approach. Then, Sec. IV details the experiments' procedure, while Sec. V presents the results. A discussion of the results is provided in Sec. VI. Finally, Sec. VII concludes the paper.

* Ferran Gebellí and Séverin Lemaignan belong to PAL robotics (Barcelona, Spain) ferran.gebelli@pal-robotics.com, severin.lemaignan@pal-robotics.com

† Anaís Garrell belongs to the Institut de Robòtica i Informàtica Industrial (IRI-CSIC), and Universitat Politècnica de Catalunya - BarcelonaTech (UPC), (Barcelona, Spain) anais.garrell@upc.edu

‡ Jan-Gerrit Habekost and Stefan Wermter belong to the University of Hamburg, (Hamburg, Germany) jan-gerrit.habekost@uni-hamburg.de, stefan.wermter@uni-hamburg.de

§ Raquel Ros belongs to the Artificial Intelligence Research Institute (IIIA-CSIC), (Barcelona, Spain) raquel.ros@iiia.csic.es

[1] The implementation code and experiments are available at https://github.com/fgebelli/personalised_XHRI

## II. RELATED WORK

In HRI, user modelling and personalisation have been extensively studied [25], primarily to adapt decision-making processes and behaviours. This has been approached through classical methods [1] and, more recently, with LLMs [37]. However, personalisation in HRI has been largely overlooked in adapting dialogues to explain the reasoning behind these decisions and behaviours.

### A. Users' knowledge tracing

A related area to personalised explanations is intelligent tutoring systems, which track users' knowledge. Various approaches have been proposed, including Learning Factors Analysis [5], Bayesian Knowledge Tracing [6], and Deep Knowledge Tracing [23], along with numerous extensions.

Nevertheless, these methods typically assume the availability of abundant and structured user feedback to update the learning process. In HRI scenarios, it is often impractical to solicit frequent feedback from users as it interrupts interactions, posing a significant challenge to direct application.

### B. Personalised XAI

In non-embodied eXplainable Artificial Intelligence (XAI), the personalisation of explanations has been explored from multiple perspectives. In [30], explanations are tailored to different types of users, which are identified using clustering techniques. Other studies focus on individual-level personalisation rather than general user categories [33]. Personalisation can also be based on user preferences [4] or a combination of preferences and performance [28].

However, these studies focus on non-embodied systems that do not encounter some of the complexities inherent to robots. In robotic systems, explanations are intrinsically more complex due to the robot's ability to physically interact with the environment, which extends the behaviours and decision-making to be explained. Additionally, the anthropomorphic attributes ascribed to robots influence human beliefs and expectations in ways that differ from interactions with purely digital systems, pushing the need for explainability.

### C. Personalised LLMs

LLMs are a promising tool for personalising conversations. In addition to tracking different chat histories for individual users, more advanced techniques have been proposed. Some approaches leverage user context through prompting mechanisms [37], though these methods are not well-suited for handling extensive historical interactions. To address this limitation, alternative strategies employ Retrieval-Augmented Generation (RAG) to extract relevant excerpts from previous conversations [26], [24]. Another widely adopted approach involves fine-tuning LLMs based on past user interactions [36], [32] or constructing user-specific embeddings [16]. Furthermore, other works have studied explicit external memories that are used to store corrections that users provide about the system's knowledge [18] or user-specified preferences [12].

Personalised LLMs have also been investigated in the context of tabletop robots. In [11], an LLM personalises interactions by incorporating summaries of prior open-domain conversations with older adults. Similarly, [22] explores adapted conversations based on the social network accounts followed by each user.

Although these studies provide personalisation, they do not incorporate complex associations to the embodiment capabilities nor an explainability perspective to explicitly model the knowledge that users have about robots.

### D. LLM-based explanations in robots

LLM-based explanations have been proposed in robotics, yet they lack the personalisation component. Template-based approaches have been presented [15], where the robot's behaviour trees are integrated into prompts to generate explanations. Another approach utilises RAG to leverage the robot's logs [29], addressing the limitations of overly long context windows. Additionally, other studies propose the generation of episodic memories based on the automatic detection of key events, capturing both the robot's internal states and relevant environmental information [17], [35].

However, these studies do not incorporate the personalisation perspective, an unaddressed issue not only in LLM-based explanation methods but more broadly in XAI [2].

Therefore, we have introduced in our previous work [8] an initial framework for personalised LLM-based explanations, drawing inspiration from works such as [29], [17], [35], while adding the personalisation component. In the present work, we further refine this framework and conduct extensive validation of the adapted explanations' level of detail.

## III. A FRAMEWORK FOR PERSONALISED XHRI

In this section, we introduce a framework for personalised explanations in HRI that tracks and updates individual memory models for interacting users[2]. In our framework, LLMs provide personalised explanations based on relevant *user context* and *robot context* that are included in the prompts.

Computing the *user context* is central within the framework, as it is directly relevant to personalisation. We present 4 key modules in our framework: (1) a *persistent user memory* keeps track of a discrete list of concepts that constitute the user's knowledge about the robot, (2) a *knowledge extraction* module populates the *persistent user memory* from previous explanations, (3) a *retrieval of relevant user knowledge* exploits embedding similarity between the user question and previously known concepts in the *persistent user memory*, and (4) a *post-processing* step transforms each of the retrieved concepts to compute the *user context*.

Since the primary objective of our work is personalisation rather than complex explanation generation, our framework assumes that the relevant robot internal information to generate the explanation is available within the *robot context*, following available approaches in the literature, e.g., [29], [17], [35]. In this work, we employ a straightforward *relevant*

[2]We assume that an application-specific and privacy-aware system provides the user re-identification that selects the adequate memory model.

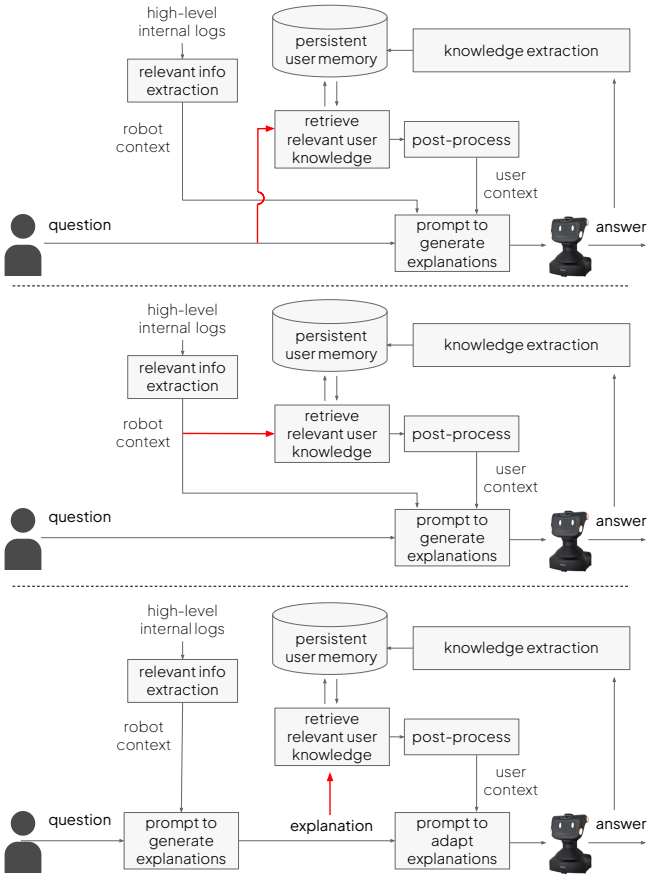


Fig. 2. Overview of the proposed architectures that employ either the user *question query* (top), *robot context query* (mid) or full *explanation query* (bottom) to retrieve the user memory. Red arrows highlight the main differences.

*info extraction* of recent high-level logs that provide the *robot context*, which is a *string* containing only the root causes and circumstances needed for the explanation generation.

In the next subsections, we first present 3 different architectures that implement our framework, and then we provide a detailed description of the core framework modules.

### A. Architecture variants

We propose 3 different architectures in Fig. 2. Since our focus is on the representation and retrieval of the *persistent user memory*, we aim to compare different ways to retrieve that memory to obtain the user-related knowledge, which will be later used in the personalisation of the explanation. The *question query* architecture (top in Fig. 2) uses the user question to retrieve related concepts. For the *robot context query* architecture (middle in Fig. 2), the *robot context* is employed to retrieve the memory. Finally, the *explanation query* architecture (bottom in Fig. 2) represents a two-stage process, in which the first stage generates a full explanation, which is used as the *persistent user memory*'s query to later adapt the initial explanation in a second LLM prompt.

We hypothesise that the query employed for the retrieval of the *persistent user memory*, which we detail in the next subsection and later compare in Sec. V, will be significantly influenced by the content of the query. We expect that when the query content is semantically closer to the *persistent user memory* entries, the retrieval process will be more accurate. In this direction, we expect the *explanation query* architecture to perform the best, since the query content is expected to have a higher similarity to previous memories, when compared to the *question query* or *robot context query*. A further advantage of the *explanation query* architecture is its modular design, enabling integration with any other explanation-generating systems, including those not LLM-based, to add the personalisation component. However, a disadvantage of the two-stage architecture is a potential increase in inference time. We discuss this trade-off later in Sec. VI.

### B. Persistent User Memory and Knowledge Extraction

Each user's memory model consists of a dynamic list of concepts that the user has previously acquired regarding the robot's functionality and behaviour. These concepts are *strings* that capture key insights, such as "*The robot cannot reach all objects on the table*" or "*The robot can be obstructed by obstacles*".

Each *string* concept is stored with two metadata fields: a timestamp indicating the most recent update of the concept and an embedding vector generated using the *text-embedding-3-small* model from *OpenAI*. When a new concept is added to the memory, its embedding vector is compared against existing entries using cosine similarity. If the similarity exceeds a predefined threshold, the new concept replaces the existing one since they essentially refer to the same user knowledge with a different rephrasing.

The *persistent user memory* is populated using concepts extracted from previous explanations by a dedicated LLM (example prompt in Fig. 3).

> **Prompt knowledge extraction:** You are an system that helps to identify the knowledge that a user has learned about a robot from an explanation. I will share with you the explanation about the robot that the user has received. You need to provide a list of the knowledge that the user has learned within that explanation. [...]

Fig. 3. Example prompt to extract the concepts from new conversations.

### C. Retrieval of Relevant User Knowledge

Following the RAG paradigm, the system retrieves concepts from the *persistent user memory* based on cosine similarity. Unlike traditional RAG implementations, which retrieve the top-$K$ similar entries, we introduce a similarity threshold ($\tau_{retr}$), since no user context should be retrieved in case there are no related known concepts. Filtering out unrelated prior knowledge contributes to both scalability and effectiveness, ensuring that the system incorporates in *user context* only truly relevant concepts. The $\tau_{retr}$ threshold can be adjusted in a trade-off that we discuss later in Sec. VI.

### D. Post-processing of Retrieved Knowledge

Our approach explicitly models the natural decline of knowledge due to lack of use over long-term interactions. Once the relevant concepts have been retrieved, we update

**Prompt one-stage:**

You are a system that helps a robot to explain its actions to the users of the robot. {use_case_robot_description}

Provide an answer to the user's questions as if you were the robot. Do not use bullet points or lists.

These are related and context relevant descriptions of what the robot can do and what the robot has been doing the past 5 minutes: {robot_context}

This is the related knowledge that the user may have about the robot: {user_processed_knowledge}. Make references to what the user already knows, mentioning when you explained it if more than several weeks have passed. Do this if and only if there is a very clear relation between the user's knowledge and the explanation. {detail_level_instructions}

Fig. 4. Prompt structure for the one-stage explanation generation process.

the probability that the user retains each concept by applying Ebbinghaus' exponential forgetting curve [7], $P = e^{-\lambda t}$, where $t$ is the time elapsed from the timestamp in the concept metadata and $\lambda$ is a decay parameter that can be adjusted based on the user type, such as young or elderly adults. This probability is post-processed to finally obtain the *user context*, which we split into two components that are later introduced in the exemplified prompts in Fig. 4 and 5:

- *user_processed_knowledge*: First, the probability values are converted into qualitative descriptors inspired by [10], e.g., translating probabilities above 0.9 to *very likely*. The elapsed time is also represented in a textual descriptor such as *a few days ago* or *about a month ago*. Each retrieved concept is then summarised in the following template: *"The user {probability descriptor} knows that {concept} from a conversation {elapsed time descriptor}."*
- *detail_level_instructions*: The probabilities are weighted by the retrieval similarity scores, and the obtained values are thresholded into one of the following instructions that specify detail level target: (1) *Provide a complete explanation with all the details*, (2) *The explanation should be moderately short*, (3) *The explanation should be concise, consisting of one or two sentences* or (4) *Provide a very short summary in a single sentence*.

### E. Explanation Generation

In the *question* and *robot context query* architectures, a single LLM simultaneously processes the *user context* and *robot context* (Fig. 4). Conversely, in the *explanation query* architecture, a first LLM generates an initial explanation based on the *robot context*, while a second LLM refines this explanation to align with the user's knowledge (Fig. 5). Importantly, the second step is only executed if there are retrieved concepts; otherwise, the full explanation is forwarded.

## IV. EXPERIMENTS

Our approach is evaluated in two distinct use cases: a hospital patrolling robot and a kitchen assistant robot. Inspired by real-world robot tests, we have identified a set of 7 topics that relate to knowledge related to robot events, failures and capabilities that users seek to gain. For each topic, we prepare a pool of 3 user questions. The evaluation

**Prompt A two-stage:**

You are a system that helps a robot to explain its actions to the users of the robot. {use_case_robot_description}

Provide an answer to the user's questions as if you were the robot. Do not use bullet points or lists.

These are related and context relevant descriptions of what the robot can do and what the robot has been doing the past 5 minutes: {robot_context}

**Prompt B two-stage:**

You are a system that helps a robot to reduce the level of detail of its explanations to a user, given the knowledge that the user has about the robot. Adapt the original explanation without changing its core idea. Provide only the adapted explanation, as if you were talking to the user.

This is the original explanation that needs to be adapted: {explanation}

This is the related knowledge that the user may have about the robot: {user_processed_knowledge}. Make references to what the user already knows, mentioning when you explained it if more than several weeks have passed. Do this if and only if there is a very clear relation between the user's knowledge and the explanation. {detail_level_instructions}

Fig. 5. Prompt structure for the two-stage explanation generation process.

presented here is conducted based on synthetic data, enabling controlled interactions and a systematic analysis.

### A. Kitchen assistant robot use case

The kitchen assistant robot helps by placing objects in boxes, handing them to the user or pointing to them. Several kitchen-related objects are used, like cans bowls or fruits. We employ the Neuro-Inspired COLlaborator (NICOL) robot from the University of Hamburg (Fig. 6). It is a robotic platform with a humanoid head and two 8 DoF arms that can verbally interact with users. Fig. 7 details the seven topics identified for this use case, along with an example question for each one.

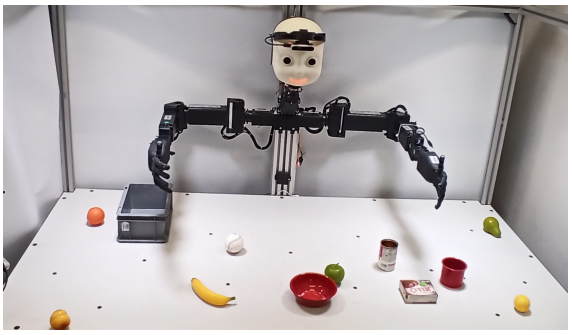

Fig. 6. Use case of a kitchen assistant robot.

### B. Hospital patrolling robot use case

In this use case, a robot is in charge of autonomously patrolling the rehabilitation unit of an intermediate healthcare facility. The robot is tasked to alert the nursing staff in case of hazardous situations for the patients, e.g., a fallen patient on the floor, as well as for internal failures. The robot-triggered alerts are received on all staff's phones, with an optional interface to obtain explanations. We use PAL's Tiago LITE mobile robot (Fig. 8). Fig. 9 lists the topics and sample questions for this use case.

| Topic | Sample question |
|---|---|
| Speech recognition failure | Why did you point to the object when I said "pick"? |
| Perception wrong detection | Why did you choose the apple instead of the pear? |
| Unreachable object | Why didn't you pick the jello box? |
| Grasping failure | Why didn't you put the bowl in the box? |
| Unconfigured object | Why didn't you pick the peeler? |
| Unseen object | Why didn't you give me the lemon? |
| Unavailable action | Why did you ignore my request to cut the apple? |

Fig. 7. Topics and sample questions for the kitchen assistant use case.

Fig. 8. Use case of a robot patrolling a healthcare facility.

### C. Procedure

We define 4 test cases, which are presented in Fig. 10. The motivation behind them is two-fold. On the one hand, we pursue to verify that the reduction of the level of detail is higher for recently learned concepts than for concepts that were learned longer ago, since they are less fresh in the user's memory. On the other hand, we aim to validate that the system does not produce summarised explanations in cases where the user does not hold any related knowledge, even if the user already knows concepts from unrelated topics.

Each test case repetition consists of two interactions between a user and a robot, as in Fig. 1. In the first interaction, we simulate 2 user questions given 2 hypothetical events related to 2 topics from Fig. 7 or 9. After a given *elapsed time* = {*1 week, 1 month*} we simulate a second interaction which takes into account the first one. In this case, the questioned topic is either a known one (from the first interaction) or a completely new one, i.e. a topic never asked before (*topic type* = {*known, unknown*}). The topics *X, Y, Z* in Fig. 10 are always different among them. We evaluate only the explanation generated in the second interaction.

Each test case is repeated 70 times per robot use case, randomly selecting the topics and questions from the pool.

We repeat the same procedure for each of the architectural variants in Fig. 2, which have been extended with four levels of the retrieval threshold to analyse different degrees of retrieval strictness: zero ($\tau_{retr} = 0$), low ($\tau_{retr} = 0.35$), mid ($\tau_{retr} = 0.45$) and high ($\tau_{retr} = 0.55$). For the $\tau_{retr} = 0$ the entire memory is retrieved, thus, the query content becomes irrelevant, and only the modules' distribution and connections matter. Therefore, the *question query* and *robot query* have been merged into a single ablation for $\tau_{retr} = 0$.

All LLMs utilised within the proposed architecture employ the *OpenAI 4o-mini* model with a temperature of 0.3.

| Topic | Sample question |
|---|---|
| Fallen person alarm | What does the robot do after raising this alarm? |
| Standing person alarm | Why didn't the robot raise the alarm in room 102? |
| Closed door alarm | How does the robot detect closed doors? |
| Emergency button pressed | How can I reset the robot? |
| Motor overheated | How much should I wait for the motor to cool down? |
| Robot lost | Why is the robot not moving? |
| Obstructed path | Why is the robot blocked? |

Fig. 9. Topics and sample questions for the hospital robot use case.

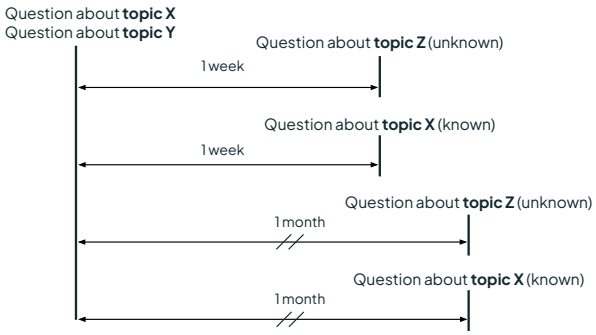


Fig. 10. Timeline for the 4 test cases in the experiment.

## V. RESULTS

The consolidated results are detailed in Table I, where each row is the average across 140 runs (*2 use cases x 70 repetitions*). We have computed separate metrics for the retrieval process and the final adapted explanation. Regarding the retrieval process, we compute the precision, recall and false positive rate based on the ground-truth topics associated with each retrieved concept. Concerning the final response, we compute the explanation length reduction compared to a baseline explanation generated with a blank user memory for the same event, topic and question. As a reference, the average baseline explanation length is 509 characters. Fig. 11 plots the length reduction values.

Instead of comparing the inference speed, which depends on the LLM model size and available computing resources, we report the prompt and response token size (Fig. 12), which proportionally influences the inference speed. In our case, we have used the *OpenAI API*, achieving a mean inference time for the whole pipeline of 3.88 seconds per explanation. In Fig. 12, we report separately experiments for the two-stage architecture, that is, the *explanation query* architecture, and the one-stage architectures, which are the *question query* and *robot context query*. We also separate experiments where no memory concepts have been retrieved versus when at least one concept has been retrieved. The rationale behind this separation is that for the two-stage architecture, the second stage is not executed if no concepts are retrieved, forwarding the initial full explanation. We segment results for the token lengths among stages for the *explanation query* architecture.

| Test case | | Ablation | | Results | | | |
|---|---|---|---|---|---|---|---|
| **Type of Topic** | **Elapsed Time** | **Architecture** | **Thresh ($\tau_{retr}$)** | **Retrieval Precision (%)** | **Retrieval Recall (%)** | **Retrieval False Positive Rate (%)** | **Response Length Reduction (%)** |
| known | 1 week | question / robot context query | zero | 51.67 | 100.00 | 48.33 | 65.97 |
| | | question query | low | 54.61 | 50.36 | 22.62 | 38.33 |
| | | question query | mid | 73.44 | 40.00 | 10.18 | 34.75 |
| | | question query | high | 88.88 | 15.00 | 1.07 | 20.04 |
| | | robot context query | low | 77.83 | 68.21 | 16.79 | 57.79 |
| | | robot context query | mid | 83.14 | 45.71 | 7.86 | 45.72 |
| | | robot context query | high | 92.28 | 25.00 | 1.96 | 29.53 |
| | | explanation query | zero | 52.29 | 100.00 | 47.14 | 43.14 |
| | | explanation query | low | 85.22 | 82.5 | 11.25 | 43.89 |
| | | **explanation query** | **mid** | **94.27** | **49.64** | **2.32** | **38.98** |
| | | explanation query | high | 96.42 | 18.57 | 0.36 | 16.64 |
| unknown | 1 week | question / robot context query | zero | - | - | 100.00 | 64.71 |
| | | question query | low | - | - | 44.11 | 34.68 |
| | | question query | mid | - | - | 24.29 | 32.38 |
| | | question query | high | - | - | 3.27 | 7.03 |
| | | robot context query | low | - | - | 32.68 | 26.45 |
| | | robot context query | mid | - | - | 17.68 | 26.18 |
| | | robot context query | high | - | - | 3.04 | 6.21 |
| | | explanation query | zero | - | - | 100.00 | 38.35 |
| | | explanation query | low | - | - | 19.88 | 23.86 |
| | | **explanation query** | **mid** | - | - | **1.79** | **2.51** |
| | | explanation query | high | - | - | 0.36 | 1.63 |
| known | 1 month | question / robot context query | zero | 51.67 | 100.00 | 48.33 | 60.17 |
| | | question query | low | 56.17 | 53.21 | 23.51 | 34.17 |
| | | question query | mid | 70.15 | 36.79 | 10.54 | 28.55 |
| | | question query | high | 80.00 | 15.00 | 1.96 | 16.30 |
| | | robot context query | low | 76.49 | 67.86 | 17.14 | 50.48 |
| | | robot context query | mid | 83.33 | 45.36 | 7.68 | 39.57 |
| | | robot context query | high | 92.94 | 25.36 | 1.96 | 25.22 |
| | | explanation query | zero | 52.86 | 100.00 | 47.14 | 39.04 |
| | | explanation query | low | 84.37 | 80.71 | 11.37 | 36.52 |
| | | **explanation query** | **mid** | **96.15** | **48.57** | **1.25** | **25.84** |
| | | explanation query | high | 98.53 | 16.07 | 0.18 | 10.31 |
| unknown | 1 month | question / robot context query | zero | - | - | 100.00 | 53.64 |
| | | question query | low | - | - | 45.71 | 34.51 |
| | | question query | mid | - | - | 22.50 | 25.19 |
| | | question query | high | - | - | 2.68 | 5.27 |
| | | robot context query | low | - | - | 33.21 | 25.41 |
| | | robot context query | mid | - | - | 18.39 | 20.73 |
| | | robot context query | high | - | - | 2.68 | 3.87 |
| | | explanation query | zero | - | - | 100.00 | 37.03 |
| | | explanation query | low | - | - | 20.78 | 23.81 |
| | | **explanation query** | **mid** | - | - | **2.68** | **4.25** |
| | | explanation query | high | - | - | 0.54 | 2.00 |

TABLE I

RESULTS FOR THE DIFFERENT TEST CASES AND ABLATIONS.

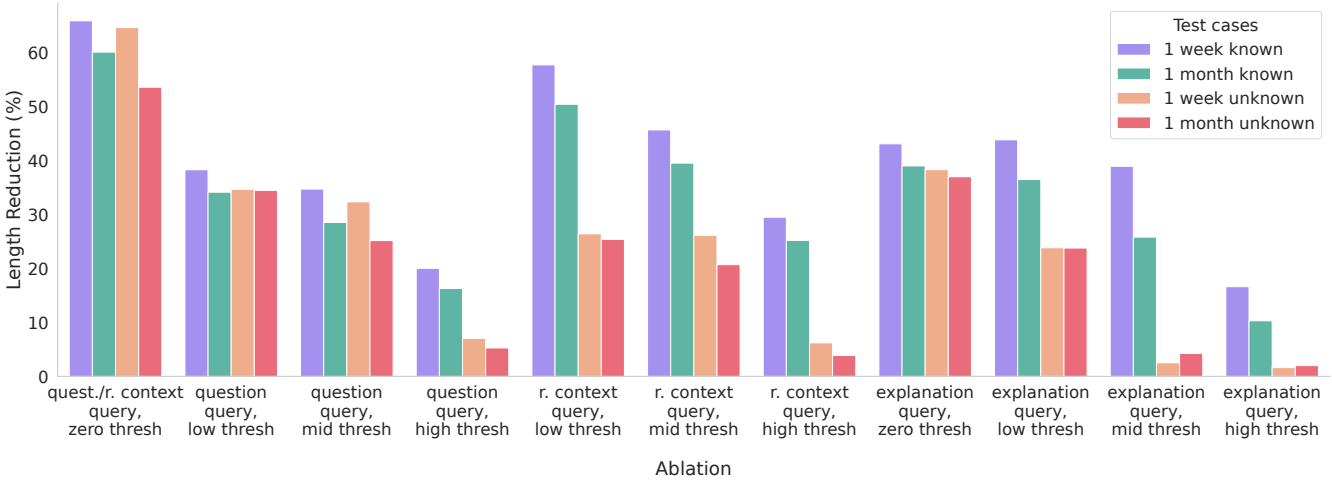


Fig. 11. Length reduction of the personalised explanations compared to the baseline explanation length.

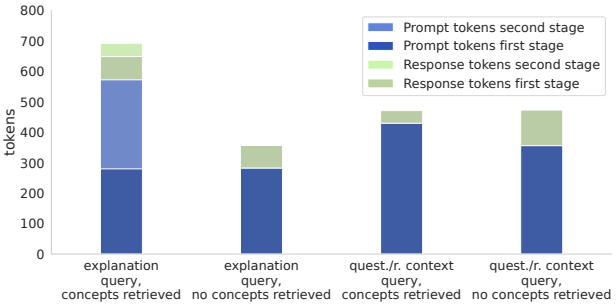


Fig. 12. Processed tokens in the prompt and response.

## VI. DISCUSSION

Before the results' analysis, we manually inspected 10% of the generated explanations across ablations, test cases and topics. This confirmed that the generated explanations were truthful and adequately addressed the root cause. This consistency is expected due to curated *robot context* provided to the LLM. Our primary focus in this section is therefore analysing the adaptation of the explanation's level of detail.

### A. Retrieved concepts

When comparing the possible architectures in Table I, the best results—higher precision and recall and lower false positive rate—are obtained for the *explanation query* setting, followed by the *robot context query* and then *question query*. These results are expected since the semantic similarity between the previously known concepts and the query will be higher for the explanation that is to be delivered, followed by the *robot context* that is used to generate that explanation and, finally, by the user question. An illustrative example of a potential low similarity in the *question query* is the following: the same question, e.g., *"Why didn't you pick the apple?"*, can lead to different explanations because either the apple was not reachable or not seen.

Moreover, results in Table I report a general trend of higher precision, lower recall and lower false positive rates as the retrieval is more strict, moving towards higher $\tau_{retr}$ thresholds. This precision-recall trade-off needs to be assessed upon each application and tuned via $\tau_{retr}$ to balance the cases where relevant concepts are missed versus when incorrect concepts are introduced. For the *unknown* topics, it is important to validate that the false positive rate is significantly low, which the ablations with $\tau_{retr} = high$ clearly achieve, as well as the *explanation query* with $\tau_{retr} = mid$.

### B. Length reduction

When comparing the discrepancy within the *topic type* for the same *elapsed time*, the length reduction difference between the *known* and *unknown* settings should be as high as possible, indicating that the system effectively reduces detail only when the user has related knowledge. Moreover, the *unknown* topics should have a very low length reduction since there is no related knowledge, and thus, no adaptation of the level of detail should take place. As expected, the ablations with a $\tau_{retr} = zero$ do not fulfil these requirements because all concepts in the memory, including completely irrelevant ones, are retrieved and considered relevant. The *question query* ablations moderately fulfil the requirements, but only for $\tau_{retr} = high$, which limits the length reduction for the *known* topics. The *robot context query* ablations present a wider difference between *known* and *unknown topic types* due to closer similarity between the query and memory concepts that lead to a more accurate retrieval. The *explanation query* ablation delivers the broadest difference between the *known* and *unknown* topics and a remarkably low length reduction for *unknown* topics, with $\tau_{retr} = mid$ showing the best tradeoff.

These outcomes align with the retrieval results since a less accurate retrieval due to a lower semantic similarity directly influences the target level of detail, as presented in Sec. III.

When comparing the *1 week* and *1 month elapsed time* settings for the *known topic type*, the length reduction is higher for the *1 week* case, as the concepts have a higher probability to be remembered and the level of detail is further reduced when compared to the *1 month* case, where the probability that users retain the knowledge is lower.

### C. Token lengths

Although the *explanation query* ablation produces better results, an expected disadvantage is that the two-stage LLM calls can lead to a higher inference time. In Fig. 12, we observe how the token length is effectively higher for *explanation query* in the cases where concepts are retrieved, but shorter when there are no retrieved memories. In the latter, the second stage is skipped, forwarding the full initial explanation without processing the *user context*, as explained in Sec. III. Therefore, depending on the frequency of users asking about known vs. unknown topics, the two-stage approach can run faster or slower than the one-stage one.

### D. Future Work

In this contribution, we have technically evaluated the proposed framework through 3 architecture implementations. Future work should study the effect of such personalised explanations on users due to an adequate level of detail. We are planning a longitudinal study where only part of the participants will receive personalised explanations, to measure if such adaptation improves the usage and usability without decreasing the understanding of the system.

Moreover, future research should investigate how to populate the memory with other multimodal sources, such as graphical user interfaces or observational systems. This way, facts that the user has experienced can update the estimated user knowledge in the memory representation on top of verbal explanations. For this purpose, a new metadata field could be added to encapsulate the initial probability of being known, populated by attention models or adapted depending on the multimodal source. This approach could require evaluating the differences between the estimated user memory and the concepts that users have actually learned.

## VII. CONCLUSIONS

In this work, we introduced a framework for generating personalised explanations. By integrating a user memory

that provides relevant user context to an LLM, the system can dynamically adjust the explanations' level of detail. We compared three architectures and evaluated the levels of correct information retrieval and length reduction. Results indicate that the proposed framework effectively leverages the user's previous knowledge, especially in the two-stage architecture. Future research should evaluate in a user study the impact of the personalised explanations on users.

## ACKNOWLEDGMENT

This work has been partially supported by the Horizon Europe Marie Skłodowska-Curie grant agreement N. 101072488 (TRAIL), the Horizon Europe grant agreement N. 10107025 (CoreSense) and the Horizon 2020 grant agreement N. 857188 (SAFE-LY-PHARAON).

## REFERENCES

[1] 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)*, pages 1–8, 2019.

[2] S. Anjomshoae, A. Najjar, D. Calvaresi, and K. Främling. Explainable agents and robots: Results from a systematic literature review. In *International Conference on Autonomous Agents and Multiagent Systems*, pages 1078–1088, 2019.

[3] T. Belpaeme. Advice to new human-robot interaction researchers. *Human-Robot Interaction: Evaluation Methods and their Standardization*, pages 355–369, 2020.

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

[5] H. Cen, K. Koedinger, and B. Junker. Learning factors analysis–a general method for cognitive model evaluation and improvement. In *International Conference on Intelligent Tutoring Systems*, pages 164–175. Springer, 2006.

[6] A. T. Corbett and J. R. Anderson. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-adapted Interaction*, 4:253–278, 1994.

[7] H. Ebbinghaus. Über das gedächtnis: untersuchungen zur experimentellen psychologie. *Duncker & Humblot*, 1885.

[8] F. Gebellí, L. Hriscu, R. Ros, S. Lemaignan, A. Sanfeliu, and A. Garrell. Personalised explainable robots using llms. In *Companion of the ACM/IEEE International Conference on Human-Robot Interaction*, 2025.

[9] T. Hellström and S. Bensch. Understandable robots-what, why, and how. *Paladyn, Journal of Behavioral Robotics*, 9(1):110–123, 2018.

[10] D. Hillson. Describing probability: The limitations of natural language. Project Management Institute, 2005.

[11] B. Irfan, S.-M. Kuoppamäki, and G. Skantze. Between reality and delusion: challenges of applying large language models to companion robots for open-domain dialogues with older adults. 2023.

[12] 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*, pages 796–806, 2024.

[13] R. Korpan, D. Tiourine, S. Chen, and S. Epstein. Evaluation of a robot navigatior's explanations. In *Companion of the ACM/IEEE International Conference on Human-Robot Interaction*, 2025.

[14] T. Kulesza, S. Stumpf, M. Burnett, and I. Kwan. Tell me more? the effects of mental model soundness on personalizing an intelligent agent. In *Proceedings of the Sigchi Conference on Human Factors in Computing Systems*, pages 1–10, 2012.

[15] G. LeMasurier, C. Tagliamonte, J. Breen, D. Maccaline, and H. A. Yanco. Templated vs. generative: Explaining robot failures. In *33rd IEEE International Conference on Robot and Human Interactive Communication*, pages 1346–1353. IEEE, 2024.

[16] 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.

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

[18] 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.

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

[20] Y. Mualla, I. Tchappi, T. Kampik, A. Najjar, D. Calvaresi, A. Abbas-Turki, S. Galland, and C. Nicolle. The quest of parsimonious xai: A human-agent architecture for explanation formulation. *Artificial Intelligence*, 302:103573, 2022.

[21] Y. Mualla, I. H. Tchappi, A. Najjar, T. Kampik, S. Galland, and C. Nicolle. Human-agent explainability: An experimental case study on the filtering of explanations. In *ICAART*, pages 378–385, 2020.

[22] T. Onorati, Á. Castro-González, J. C. del Valle, P. Díaz, and J. C. Castillo. Creating personalized verbal human-robot interactions using llm with the robot mini. In *International Conference on Ubiquitous Computing and Ambient Intelligence*, pages 148–159. Springer, 2023.

[23] C. Piech, J. Bassen, J. Huang, S. Ganguli, M. Sahami, L. J. Guibas, and J. Sohl-Dickstein. Deep knowledge tracing. *Advances in Neural Information Processing Systems*, 28, 2015.

[24] 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.

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

[26] A. Salemi, S. Kallumadi, and H. Zamani. Optimization methods for personalizing large language models through retrieval augmentation. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 752–762, 2024.

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

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

[29] 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.

[30] U. Soni, S. Sreedharan, and S. Kambhampati. Not all users are the same: Providing personalized explanations for sequential decision making problems. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 6240–6247. IEEE, 2021.

[31] S. Sreedharan, T. Chakraborti, and S. Kambhampati. Foundations of explanations as model reconciliation. *Artificial Intelligence*, 301:103558, 2021.

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

[33] 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 International Conference on Autonomous Agents and Multiagent Systems*, pages 2316–2318, 2023.

[34] 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*, pages 119–138. Springer, 2021.

[35] 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 *Annual Conference on Robot Learning*, 2024.

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

[37] 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*, 47(8):1087–1102, 2023.