Personalising Explanations for Robot Failures in Robot Operating System using Parameter-Efficient Fine-Tuning

by

Ella Scheltinga

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Tuesday December 17th, 2024 at 14:00

Student number:	4833856	
Project duration:	February 2024 – December 2024	
Thesis committee:	Dr. C. Pek. ,	TU Delft, supervisor
	J. M. Prendergast,	TU Delft
	C. R. M. M. Oertel,	TU Delft

An electronic version of this thesis is available at http://repository.tudelft.nl/.

TU Delft

Personalising Explanations for Robot Failures in Robot Operating System using Parameter-Efficient Fine-Tuning

Ella Scheltinga

***Abstract*— Autonomous robots are increasingly improving at performing navigation tasks, however they will likely fail at some point or not perform as intended due to uncertainties or unforeseen situations in the real world. In such scenarios, explaining the robot's behaviour to humans is crucial to build trust and resolve potential issues. Recently, large language models (LLMs) have shown great potential in analysing robot log data, e.g., obtained in Robot Operating System (ROS), and providing users with useful explanations. Yet, these models often fail to consistently generate high quality answers. This study develops an approach using parameter-efficient fine-tuning (PEFT) to improve explanations generated by LLMs and tailoring them towards a target audience (expert, non-expert) and preferred lengths (short, medium, long). We collected ROS log data from the TIAGo robot in simulation, combined them with user questions, and corresponding answers generated using GPT-4o to create a dataset for fine-tuning Mistral 7B with PEFT. Furthermore, we use a panel of LLMs (GPT-4o, Mistral-Large, Llama3-8B) as judges to evaluate these explanations based on quality criteria and user study (N=17) to validate these results on a group of roboticists. Our findings show that personalisation significantly improves both the suitability of explanations, with personalised answers consistently outperforming non-personalised ones. Furthermore, tailored explanations achieved higher clarity and user understanding. Additionally, a single feedback loop iteration using textual feedback from LLMs further enhanced explanation relevance and contextual quality, demonstrating the value of iterative improvement in explainability systems, despite minor trade-offs in other criteria.**

I. INTRODUCTION

Robots are becoming increasingly integrated into our daily lives, making encounters with their failures more frequent. Therefore, the importance of explainability in human-robot interaction (HRI) becomes evident. Explainability fosters understanding between the user and a robot, contributing to building a trustworthy relationship between humans and robots. One effective approach is generating explanations that improve the user's understanding about the behaviour of the robot and why failures may have occurred.

Consider a simple navigation task, as shown in Figure 1, where a TIAGo robot is instructed to move along three waypoints while avoiding obstacles. TIAGo reaches waypoints 1 and 2, but fails to reach waypoint 3 and no path is shown towards this waypoint. Explainability systems are designed to address such scenarios by providing clear answers to user questions, such as: ”Why has the robot failed to reach waypoint 3?”, and ”Why is there no feasible path?”. In this situation, a path towards waypoint 3 is not feasible because it is located within a known obstacle.

Author is with Delft University of Technology, Faculty of Mechanical Engineering, Department of Cognitive Robotics, Delft, the Netherlands. E.M.Scheltinga@student.tudelft.nl

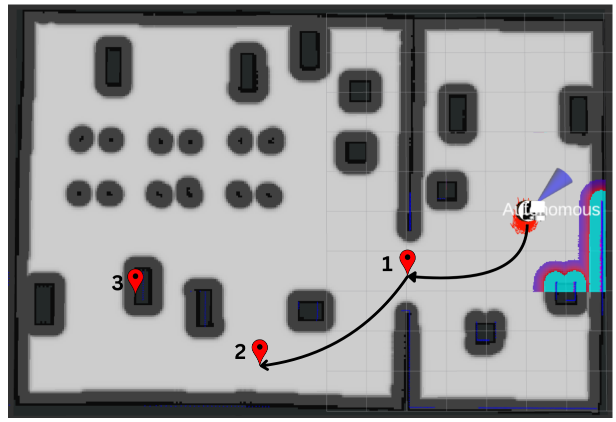


Fig. 1: Navigation task with three waypoints that our simulated TIAGo moves towards. However, it fails to reach waypoint three, since there is an obstacle.

Although the precise definition of explainability varies in the literature [1], [2], it is widely associated with interpretability and transparency [3], [4]. In the context of HRI, explainability aims to improve the user's understanding of the robot through clear and truthful explanations that align with the robot's logic [5]. Explanations are also viewed as answers to diverse user questions about failures, and as an interaction between a human and a robot [6].

The typical users of explainability systems for autonomous robots can be categorised into experts and non-experts, each with distinct explainability needs and objectives [7]. Non-expert users generally expect explanations for robot failures to use accessible language that clearly justify the cause of the failure, in order to build trust [8]. In contrast, expert users such as researchers working on a scientific discovery will likely be willing to dedicate more time to an explanation and expect them to include more technical details [5]. However, these preferences may also depend on the failure. Experts may not always require long or detailed explanations, particularly for failures with straightforward solutions. Similarly, non-experts may sometimes prefer longer explanations, especially when they seek a deeper understanding or a thorough justification for a robot's failure. Therefore, explanations should not only align with the level of expertise of the user [8], but also be adjustable according to the user's desired answer length.

Most robots operate using the open-source Robot Operating System (ROS), which enables modular development of robot software. ROS uses a publish-subscribe architecture, where nodes exchange data via topics. Data flow between nodes, especially logs generated through the */rosout* topic, provides structured, natural language messages that give insights into the robot's internal states, decisions, and failures. These textual log messages can serve as a foundation for explanations, as they accurately reflect the robot's inner workings. This textual data is already utilised by users for debugging, though currently without the support of an explainability system. Furthermore, the ROS framework allows developers to apply the same code to different robots, meaning an explainability system using ROS developed for one robot can often be adapted to work on others with minimal modifications. Therefore, using ROS log messages for explainability systems enables them to provide explanations grounded in readily available data that reflect the robot's inner workings, while ensuring compatibility for other robots.

This study utilises ROS log messages produced by the robot during navigation tasks to ground explanations, ensuring they reflect the robot's logic. We will enhance the logs produced by ROS by adding custom log messages to increase the verbosity of recorded log data.

We propose leveraging LLMs to utilise their advanced natural language processing capabilities to effectively interpret textual data, such as ROS log messages. Additionally, they are designed to generate explanations in natural language for user questions with possibilities for interaction, as seen in chatbots. These capabilities enable LLMs to provide explanations to questions about robot behaviour based on their ROS log messages [9], [10]. However, LLMs alone have limitations, such as hallucinations or providing answers in an undesired format [11]. To mitigate these limitations, we intend to adjust the LLM's weights by fine-tuning [12] these models on a dataset containing examples of question-answer pairs based on ROS log messages. Through fine-tuning on task-specific data, the LLM provides answers that are more accurate and better aligned with the task [12].

The fine-tuned LLM is designed to interpret new ROS log messages and, based on this data, generate answers to user questions. This study builds on preliminary results from the RSS 2024 workshop on Robot Execution Failures and Failure Management Strategies, with a focus on personalising explanations to align better with user needs. This process, referred to as personalisation, adapts the content and format of responses to match the user's expertise level and preferred answer length. For instance, responses for expert users focus on technical depth and precision, while answers for non-experts prioritise clarity and accessibility. Additionally, response length is tailored to suit user preference for concise or detailed answers, by providing short, medium or long answers. This categorisation is supported by prior research, which classifies users of explainability systems into experts and non-experts, each with distinct needs and preferences [7], [8].

Overall, this study aims to enhance human-robot interaction during failures by leveraging PEFT techniques and personalised explanations tailored to user needs. Based on this approach, this work focuses on answering the following research questions:

1) *How does personalisation affect the suitability of explanations in terms of length and alignment with target audience expertise?*
2) *How does personalisation affect the quality of explanations?*
3) *How to further improve the quality of personalised explanations?*

II. RELATED WORK

This section compares previous studies on robot explainability that leverage ROS log messages.

Fernandez-Becerra et al. [14] employ conditional logic to read messages from specified ROS topics and extract relevant information to generate straightforward and concise natural language explanations for a navigation task. This method addresses two specific user questions: ”What is the robot's current status?” and ”Why has the robot changed the path?”. In response to the first question, the algorithm compares the previous goal with the current goal, and if they differ, it returns ”Navigation to a new goal has started”. For the second question, if the distance from the goal location increases above a threshold and an obstacle is located within a threshold, the algorithm prints ”I have changed the planned path because there was an obstacle”. Although this approach addresses two specific user questions for a navigation task, it restricts user understanding to these predefined questions and offers only limited context-specific explanations.

Building on the limitations of the rule-based method, another approach [10] leverages LLMs to interpret ROS log messages and generate explanations in response to user questions. This allows users to ask a wider range of questions and receive more detailed explanations about the robot's behaviour. This approach involves splitting Rosbag files, containing ROS log messages, to fit the LLM's prompt size requirement and uses prompt engineering to guide the LLM in generating relevant explanations. Specifically, this study uses single-shot prompting, where the LLM is provided with one example to generate a similar response. The key advantage of this approach is that, by using LLMs, users can ask any question about the robot's behaviour, instead of being limited to two predefined questions. Additionally, the responses to the questions contain greater depth and more context, which is provided by the LLMs' natural language processing capabilities. However, the study highlights that generic pre-trained models often fail to interpret domain-specific logs correctly and rely heavily on prompt design, which often leads to inaccurate and inconsistent explanations. To overcome these limitations, they propose investigating fine-tuning LLMs.

Another study [13] introduces an explanation system using an LLM with Retrieval Augmented Generation (RAG) to interpret ROS log messages, as shown in Figure 2. Collected

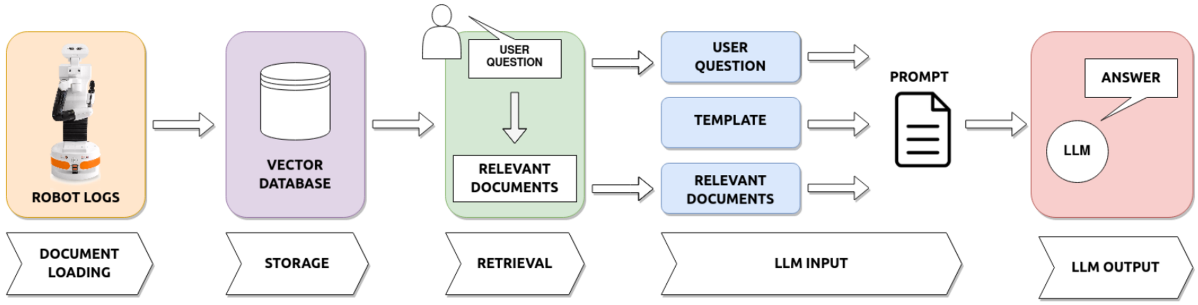


Fig. 2: Overview of the Retrieval Augmented Generation framework [13]

ROS log data is stored in a vector database. Based on the provided user question, the RAG system retrieves relevant ROS log data to provide a broader context for LLM's response. The retrieval system uses similarity-based matching, such as Maximal Marginal Relevance (MMR), to identify and rank the most relevant logs from the database. The amount of provided contextual data is limited by the prompt size of the used LLM. The process involves forming a prompt for the LLM using the user question, a predefined prompt template, and the retrieved relevant ROS log data, which then produces the answer to the user question. The advantage of this method is its ability to retrieve multiple relevant examples from the database, enabling the LLM to generate responses closely aligned with the provided context. However, the effectiveness of the system is constrained by the prompt size of the LLM, which limits the number of examples that can be included, and by the quality of the retrieval function, which may affect the relevance of the retrieved examples.

Both studies [10], [13] build on [14], demonstrating the potential of combining LLMs with prompt engineering and RAG to interpret ROS log messages to provide comprehensible explanations to user questions. These studies highlight the potential of integrating LLMs for explanation generation based on ROS log data to enhance robot explainability. While current methods offer a solution for this problem, a promising avenue that remains unexplored is fine-tuning.

Fine-tuning adapts existing LLMs to specific tasks, in this case answering user questions based on provided ROS log messages, significantly improving response quality [15]. This method addresses previous limitations by reducing reliance on prompt size constraints and retrieval system quality. Fine-tuning allows the LLM to utilise an entire dataset of ROS log messages and corresponding question-answer pairs, bypassing prompt limits and reducing reliance on prompt engineering. This retraining process enables the LLM to align outputs more effectively with the context of the data for this specific task, leading to more accurate and relevant answers. Fine-tuning ensures that the LLM can leverage the entire dataset for generating responses, thereby providing richer and more precise explanations tailored to specific queries. We will explore fine-tuning techniques to analyse ROS log data and answer user questions.

III. METHOD

We investigate leveraging Parameter-Efficient Fine-Tuning (PEFT) to generate explanations based on ROS log messages and user questions. Furthermore, we personalise these explanations towards the target audience and desired answer length and evaluate them using a panel of LLM judges and a user study. Finally, we further improve the personalised explanations using textual feedback provided by the LLM judges (See Figure 3).

A. Parameter-Efficient Fine-Tuning

Fine-tuning techniques enable the customisation of LLMs for a specific task [12], in this case question answering based on ROS log messages. Fine-tuning retrains an LLM using a task-specific dataset to optimise performance. There are various fine-tuning approaches. The most comprehensive is fine-tuning, where the entire model is retrained, adjusting all its parameters to better suit the new dataset [16]. However, this approach is highly resource-intensive and time-consuming and is therefore deemed infeasible for the scope of this study.

For ROS log interpretation tasks, a dataset of Rosbag files containing ROS log messages is required for fine-tuning. Given that such a dataset is not readily available online, it must be collected manually, which constrains its size. Given this constraint, we propose using Parameter-Efficient Fine-Tuning (PEFT), a method that selectively adjusts a small subset of the model's parameters, such as the final layers, adapter modules, or specific attention heads, while leaving the majority of the model unchanged [17]. This method leverages the existing knowledge encoded in the pre-trained model, allowing it to adapt to new tasks with minimal adjustments. By concentrating on key parameters, PEFT can achieve comparable performance to full fine-tuning while being both computationally and data efficient [18], [19]. However, PEFT can underperform on complex tasks that require extensive parameter updates across the entire model [20]. This targeted approach significantly reduces computational costs and the amount of fine-tuning data required compared to full fine-tuning, which retrains the entire model. The study by Weyssow et al. [21] demonstrates that PEFT outperforms traditional full fine-tuning, especially in limited-data scenarios, while offering benefits such as reduced memory and computational demands. Further research shows that for PEFT, increasing the size of the base model has a greater

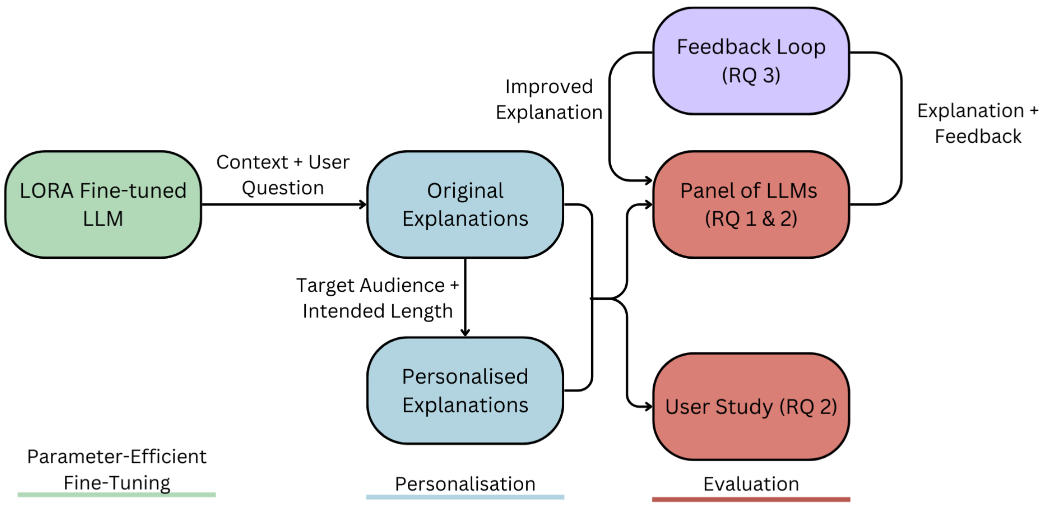


Fig. 3: Overview of our proposed method that fine-tunes Mistral 7B with LORA, which is used to generate explanations based on a context and user question. These explanations are personalised according to target audience and intended length. Both personalised and original explanations are evaluated using a panel of LLM judges and a user study. A single feedback loop is introduced to further improve explanations using textual feedback from the panel of LLMs.

impact on improving the performance than expanding the size of the fine-tuning dataset [16], [17].

Low-Rank Adaptation (LoRA) [22] is a reparametrisation PEFT technique that demonstrates potential in customising LLMs for targeted tasks like automated code generation, while efficiently managing computational load [21]. LoRA optimises rank decomposition matrices of dense layer changes during adaptation while keeping pre-trained weights frozen, achieving efficiency in storage and computation even with low rank [22]. In addition, a study [23] evaluates LoRA using datasets containing 1,000 examples to give an indication of the data size required size for effective fine-tuning. Therefore, as shown in Figure 3, we employ a large-scale LLM with LORA to maximise its performance for the ROS log interpretation task.

Algorithm 1 Personalise Explanations for Audience Type and Length

Require: Context, Question, Explanation, Audience Type (`"expert"` or `"non-expert"`), Length (`"short"`, `"medium"`, or `"long"`)
Ensure: Personalised Explanation

```
user_prompt ← "Rewrite the explanation
for an [Audience Type] audience with a
[Length] response. Context: " +
Context + "Question: " + Question +
"Explanation: " + Explanation
pers_expl ← GPT4o_CALL(user_prompt)
return pers_expl
```

B. Personalisation

After fine-tuning the LLM using LORA, we prompt the model to generate explanations for new scenarios. These outputs serve as the original explanations, as seen in Figure 3. To address the research questions, we actively personalise these explanations based on two key factors: the target audience and the desired answer length. Specifically, the original explanations are transformed into six tailored types:

- Expert: short, medium, long
- Non-expert: short, medium, long

We personalise the explanations by leveraging GPT-4o and designing targeted prompts to adapt the explanations according to the audience's expertise and preferred length. This process employs prompting principles detailed in [24] to ensure the outputs align with the intended objectives. Algorithm 1 presents the pseudo code used for this approach.

C. LLM-as-a-judge

To quantify the suitability (RQ 1) and quality (RQ 2) of the original and personalised explanations, we employ a LLM-as-a-judge framework, depicted in Figure 3. LLM-as-a-judge refers to the use of LLMs as automated evaluators to assess the quality of responses, providing a scalable alternative to traditional human evaluation [25]. The study by Zheng et al. (2023) demonstrates that LLM-as-a-judge evaluations achieve comparable accuracy to human judgments [25], underscoring the potential of the LLM-as-a-judge framework. A limitation of a single LLM-as-a-judge is that it is susceptible to intra-model bias [26]. Therefore, we employ a diverse panel of judges to correlate better with human judgements, reduce bias, and incorporate diverse perspectives [26]. A recent study shows that panels composed of smaller open source models can often outperform individual larger models in evaluation tasks, offering lower bias within the model and more cost-effective solutions without compromising their reliability [27]. Therefore, we selected a diverse panel of judges from different companies and with different model sizes, consisting of: GPT-4o, Mistral-Large, and LLaMA3-8B to ensure a balanced and comprehensive evaluation.

D. User Study

To validate the panel of judges' evaluation of the quality of original and personalised explanations, we conducted a user study, as shown in Figure 3. The study targeted roboticists (experts) with experience using ROS, simulating a scenario where they encounter a robot failure. Participants were presented with relevant ROS log messages, a related user question, and two answers: one personalised and one non-personalised version. To minimise the influence of answer length on the results, we compared the expert medium answer with the original answer, as these were most similar in length (see Appendix I). Additionally, participants were asked to indicate which answer they preferred with give brief reasoning, offering qualitative insights into the effectiveness of the personalisation.

Algorithm 2 Improving Explanation Quality Using Feedback from LLMs

Require: Context, Question, Explanation, Audience Type (`"expert"` or `"non-expert"`), Length (`"short"`, `"medium"`, or `"long"`), Feedback
Ensure: Improved Explanation

```
user_prompt ← "Improve the explanation using the feedback while keeping it tailored for an [Audience Type] audience with a [Length] response length. Context: " + Context + "\nQuestion: " + Question + "\nExplanation: " + Explanation + "\nFeedback: " + Feedback
improved_expl ← GPT4o_CALL(user_prompt)
return improved_expl
```

E. Feedback loop

To address research question 3, we propose using a single feedback loop iteration, incorporating textual feedback from LLM judges to further enhance the quality of personalised explanations (See Figure 3. During the evaluation, the LLM judges not only assess the quality of the explanations, but also provide textual feedback explaining their judgments. This feedback loop is implemented by prompting GPT-4o to refine the personalised explanations based on the textual feedback provided by the LLM judges. The pseudocode for this process is shown in Algorithm 2.

IV. EXPERIMENTAL SETUP

This section describes the processes for collecting and pre-processing ROS log data and fine-tuning a LLM using PEFT. It also details the evaluation of generated explanations by a panel of LLMs and with a user study, as well as the use of a single feedback loop to improve personalised explanations based on LLM feedback.

A. Collecting ROS log data

Fine-tuning the LLM with PEFT requires ROS log data, which was collected from the TIAGo robot simulation using the PAL Robotics software stack [1]. Using the autonomous navigation functionality, we simulated a navigation task where the TIAGo robot was instructed to move along a series of waypoints while avoiding obstacles. Specifically, we collected data where the robot was tasked with navigating waypoints in five distinct scenarios:

- **Scenario 1:** Only known obstacles, and all paths are feasible.
- **Scenario 2:** Unknown obstacle(s) in its path, yet all paths are feasible.
- **Scenario 3:** Only known obstacles, yet not all paths are feasible. (See Figure 1)
- **Scenario 4:** Unknown obstacle(s) in its path, and not all paths are feasible.
- **Scenario 5:** No unknown obstacles, but at least one of the paths goes out of bounds.

In these scenarios, known obstacles refer to those present in the robot's costmap for path planning, while unknown obstacles are not. Unknown obstacles were manually added to the environment after the cost map and global path were generated to ensure the robot encountered these obstacles during navigation. These scenarios were adapted from Sobrín-Hidalgo et al. [13], who investigated RAG for generating explanations, to the PAL robotics environment and enriched by including unknown obstacles and paths that go out of bounds. To simulate real-world variability and create a balanced dataset, waypoint positions were selected randomly to ensure their combined global path was consistent with the scenario. For each scenario, we collected 15 Rosbag files, resulting in a total of 75 recorded Rosbag files containing the ROS log messages produced by the */rosout* topic. Recording the Rosbag began simultaneously with task initialisation to capture all log messages related to the navigation task.

B. Pre-processing data

Pre-processing involves splitting the collected ROS log data into fine-tuning (80%) and testing (20%) sets, and generating question-answer pairs about the ROS log data. Each Rosbag file containing ROS log data, referred to as context, was annotated with multiple questions and corresponding answers, creating multiple data samples for each context.

Questions were adapted from Sobrín-Hidalgo et al. [13] and additional questions were proposed for this study (See Appendix II). Ten general questions, focussing on broader aspects of the navigation task, such as received waypoints and planned paths, were applied to all scenarios. Additional scenario-specific questions were applied to Scenarios 2, 3, 4, and 5 to address scenario-specific challenges, such as handling unknown obstacles or navigating infeasible paths.

Answers were generated using GPT-4o, selected for its state-of-the-art performance [28]. We leveraged the prompting principles outlined by Bsharat et al. [24], because they

[1] https://wiki.ros.org/Robots/TIAGo/Tutorials

provide a structured and effective approach for eliciting reliable and consistent responses from large language models. The full prompt is shown in Appendix III. Manual answer generation was avoided due to time constraints and GPT-4o's capability to generate reliable explanations using a structured prompt. To ensure accuracy, all answers were manually reviewed for errors and hallucinations. This process yielded 900 labelled instances, compiled into a .json file to use as input for PEFT.

C. Parameter-Efficient Fine-Tuning

Supervised PEFT was applied to Mistral-7B-v0.3 [29], chosen for its open-source nature and compatibility with Hugging Face tools, which simplifies integration for tasks like fine-tuning. This model was selected as the best fit based on the trade-off between model size and available computational resources. Additionally, Mistral-7B shows superior performance compared to larger Llama 2 models and utilises improved attention mechanisms for enhanced performance with a smaller model size [29]. The LORA configuration from the Hugging Face PEFT package was used, with fine-tuning performed using 8 NVIDIA H100 GPUs from Lambda Labs.

The fine-tuned model is publicly accessible on Hugging Face: https://huggingface.co/EllaScheltinga/SFT-Mistral-7B

Algorithm 3 Evaluate Explanation for Suitability for Target Audience and Intended Length, and Quality

Require: Context, Question, Explanation, Audience Type (`"expert"` or `"non-expert"`), Length (`"short"`, `"medium"`, or `"long"`), Rating Scale
Ensure: Rating [1-7] and Textual Feedback

```
user_prompt ← "Rate the explanation using the rating scale. Context: " + Context + "\nQuestion: " + Question + "\nExplanation: " + Explanation + "\nRating Scale: " + Rating Scale
evaluation ← JUDGES_CALL(user_prompt)
return evaluation
```

D. LLM-as-a-judge

The LLM-as-a-judge method for evaluation was employed to assess the suitability (RQ 1) and quality (RQ 2 & 3) of both personalised and non-personalised explanations.

The suitability for the target audience (expert or non-expert) and length (short, medium or long) was evaluated. The quality of explanations was evaluated based on the following five criteria:

1) **Relevance to Question:** How relevant is the explanation to the question?
2) **Contextual Accuracy:** How accurate is the explanation compared to the ROS log messages?
3) **Enhancement of Understanding:** To what extent does the explanation improve understanding?
4) **Answer Clarity:** How clear and easy to understand is the explanation?
5) **Contextual Explanation Quality:** How well does the answer provide relevant context?

The quality criteria were adapted from previous work by Sobrín-Hidalgo et al. [13] and refined to focus more on relevant contextual information, clarity, and readability.

Both suitability and quality were rated on a scale from 1 to 7. To ensure consistency, textual descriptions were added to define each score, guiding the LLMs in their evaluations, as advised by Hugging Face's guidelines [30]. Algorithm 3 provides pseudo code for the prompt used to obtain evaluations from each model.

E. User study

For the user study, we developed an online questionnaire featuring 5 scenarios that each provided relevant ROS log messages, a user question and two answers (personalised and non-personalised) for the participants to evaluate. See Appendix IV for each question-answer pair. The participants evaluated the quality of the explanations using the same five criteria as the panel of LLMs, with the addition of one new criterion:

6) **Perceived Reliability:** How reliable and trustworthy is the explanation?

Perceived reliability was excluded from LLM panel evaluation, as this subjective criterion could not be reliably assessed by the panel of LLMs.

The study involved 17 respondents (11 men and 6 women) with varying levels of ROS experience, ranging from basic to proficient experience. All participants had a formal background in robotics, including current Master's students, Master's graduates, and PhD holders. This ensured that all respondents had relevant expertise to understand the ROS log messages and evaluate the explanations effectively.

V. RESULTS

This section presents and discusses the results for each research question.

A. RQ 1: Suitability of Explanations

Two tests were conducted to assess how personalisation affects suitability for the intended target audience and answer length:

- **Test 1:** Compares personalised explanations with non-personalised explanations (e.g., expert short vs. non-personalised answers for expert suitability)
- **Test 2:** Compares explanations personalised for one characteristic with those personalised for the opposite characteristic (e.g., expert short vs. non-expert short for expert suitability)

The objective of these tests is to evaluate whether personalisation improves suitability for the intended audience and length. Test 1 focuses on the benefit of personalisation compared to original explanations, while Test 2 examines how explanations tailored for one characteristic (e.g., experts or short) perform compared to those tailored for the opposite (e.g., non-experts or long), keeping other factors constant.

TABLE I: Mean Suitability Ratings for Target Audience for Different Personalizations

Personalisation	Mean	Std. Dev	p-value
Expert Suitability			
Original	6.13	0.387	-
Expert short	6.19	0.341	0.157
Non-expert short	4.66	0.623	**1.55e-72***
Expert medium	6.51	0.212	**2.42e-24**
Non-expert medium	3.87	0.731	**1.88e-100***
Expert long	6.49	0.328	**3.65e-17**
Non-expert long	2.36	0.597	**2.59e-146***
Non-expert Suitability			
Original	4.10	0.522	-
Expert short	3.66	0.726	**8.72e-73***
Non-expert short	5.52	0.435	**1.14e-70**
Expert medium	3.20	0.595	**5.97e-115***
Non-expert medium	5.85	0.330	**2.63e-90**
Expert long	2.45	0.680	**1.05e-132***
Non-expert long	6.12	0.251	**7.52e-104**
Short Suitability			
Original	4.03	0.900	-
Expert short	5.89	0.517	**6.86e-66**
Non-expert short	5.74	0.548	**5.28e-56**
Expert long	1.99	0.209	**2.57e-179***
Non-expert long	2.10	0.250	**6.56e-179***
Medium Suitability			
Original	5.65	0.480	-
Expert medium	5.70	0.468	0.167
Non-expert medium	5.68	0.362	0.447
Long Suitability			
Original	5.33	0.972	-
Expert short	3.52	0.998	**6.78e-181***
Non-expert short	3.30	0.847	**4.09e-180***
Expert long	6.45	0.298	**3.24e-33**
Non-expert long	6.25	0.263	**9.06e-25**

Table I presents the means and standard deviations of suitability ratings for both tests across two dimensions: target audience (expert or non-expert) and answer length (short, medium, or long). Statistical differences were assessed using paired t-tests, with significant results highlighted in bold and Test 2-specific results marked with an asterisk (*).

1) Original vs Personalised:

a) Target Audience Suitability: Personalisation significantly enhances the suitability of explanations for their intended target audience:

- **Experts**: Explanations tailored for experts consistently outperformed non-personalised (original) explanations for expert suitability. Medium (6.51) and long (6.49) personalised explanations were rated significantly higher, whereas expert short (6.19) explanations performed similarly to the original (6.13). This may be due to limited opportunities for detailed tailoring in short formats.
- **Non-experts**: Personalised explanations for non-experts scored higher than the original across all lengths, with non-expert long explanations performing best. This suggests that longer explanations allow more room to adapt content to non-expert needs.
- The original explanations showed higher suitability for experts (6.13) than non-experts (4.10), likely reflecting the model's training on more expert-oriented question-answer pairs.

b) Length Suitability: Personalisation improved suitability for specific lengths:

- Short (5.89, 5.74) and long (6.45, 6.25) personalised explanations were rated significantly higher for their respective lengths compared to the original (short: 4.03, long: 5.33), aligning with differences in mean word counts shown in Appendix I.
- For medium-length suitability, personalised (5.70, 5.68) and original (5.65) explanations scored similarly, reflecting their comparable word counts.
- The original explanations exhibited some of the highest standard deviations for short (0.900) and long (0.972) suitability, indicating less consistent ratings compared to personalised explanations.

2) Personalised vs Opposites:

a) Target Audience Suitability:

- Non-expert explanations were consistently rated significantly lower for expert suitability (4.66, 3.87 and 2.36), and expert explanations scored significantly lower for non-expert suitability (3.66, 3.20 and 2.45) across all lengths.
- Suitability mismatches became more pronounced as explanation length increased, suggesting that audience alignment is easier to distinguish in longer explanations.

b) Length Suitability:

- Long explanations rated for short suitability (1.99 and 2.10) received lower scores than short explanations rated for long suitability (3.52 and 3.30). This indicates that longer explanations were more clearly identifiable as unsuitable as short.
- Short explanations rated for showed some of the highest recorded standard deviations (0.998 and 0.847), indicating less confidence from judges in their ability to distinguish the intended length of short answers.

Overall, personalisation enhances suitability for both target audience and length. Longer explanations provide the greatest benefits for audience alignment, while personalised short and long explanations outperform the original in their respective length categories. This trend is consistent with findings by Huang et al. [27], which show that LLMs tend to favour more verbose answers during evaluation.

B. Individual Judges Analysis

Due to the higher standard deviations observed in some conditions during Tests 1 and 2, we investigated the mean ratings of individual judges in more dtail to better understand the cause of this variability. We focused specifically on conditions with standard deviations above the median value, using the median as a threshold to identify the top 50% most

TABLE II: Mean Ratings for Personalisation Types Across Evaluation Criteria

Personalisation Type	Crit 1: Relevance to Question	Crit 2: Contextual Accuracy	Crit 3: Enhancement of Understanding	Crit 4: Explanation Clarity	Crit 5: Contextual Explanation Quality
Original	6.49	6.51	6.62	6.28	6.48
Expert short	6.36	6.55	6.68	**6.47**	6.49
Expert medium	6.27	**6.75**	**6.88**	**6.55**	**6.84**
Expert long	5.93	6.56	6.81	6.20	**6.75**
Expert long (improved)	**6.39**	6.52	**6.88**	**6.51**	6.75
Non-expert short	5.91	6.28	6.06	6.08	5.95
Non-expert short (improved)	**6.27**	6.34	**6.41**	**6.40**	**6.24**
Non-expert medium	6.26	6.45	6.53	**6.40**	6.43
Non-expert long	5.53	6.29	6.38	6.14	6.55
Non-expert long (improved)	**6.27**	6.34	**6.52**	**6.40**	6.45

variable conditions. Table VII in Appendix V presents the means and standard deviations for each judge individually for these conditions.

- **GPT-4o**: Ratings aligned most closely with the intended purpose of the explanations. The judge consistently rated expert explanations as highly suitable for experts, non-expert explanations as highly suitable for non-experts, and similarly rated explanations appropriately for their intended length. These ratings also exhibited lower standard deviations, suggesting greater consistency and confidence. Since this model was used for prompting and generating answers for the fine-tuned data, its alignment with expected trends is likely influenced by this.
- **Llama3-8B**: Ratings aligned least with the intended purpose with overall higher standard deviations, suggesting less confidence. This could be due to the smaller size of the model, although it added diversity to the panel of LLMs.
- **Mistral-Large**: Performed better than Llama3-8B, with lower standard deviations, but overall less than GPT-4o.

High variability in Llama3-8B's ratings suggests it has lower confidence in its ratings, however does introduce diversity to the panel, which may be valuable for capturing a broader range of perspectives. GPT-4o and Mistral-Large exhibited lower variability overall, with GPT-4o demonstrating the closest alignment with the intended purpose of explanations.

C. RQ 2: Quality of Explanations

Two tests were conducted to determine the effect of personalisation on the quality of the explanation:

- **Test 3:** Compares the quality of personalised explanations and non-personalised explanations.
- **Test 4:** Compares the quality of expert medium explanations with non-personalised explanations.

Table II presents the mean ratings for original and personalised explanations for five criteria related to explanation quality. Table VIII in Appendix VI shows the results of a paired t-test, where personalised explanations are compared with original explanations. The bold p-values signify a significant improvement in personalised explanations. The mean values reported in Table II that have a significant improvement are also in bold.

1) Panel of LLMs:

- **Criteria 1 (Relevance to Question)**: Long explanations scored lower, possibly due to dilution of relevance with additional details. Non-expert explanations were rated lower than expert ones of the same length, while original explanations scored highest, suggesting that personalisation may shift focus toward the target audience and length rather than the question itself.
- **Criteria 2 (Contextual Accuracy)**: All explanations performed well, indicating personalisation does not compromise accuracy, expert medium did show significant improvement.
- **Criteria 3 (Enhancement of Understanding)**: Expert explanations performed best, likely due to their depth and detail.
- **Criteria 4 (Explanation Clarity)**: Medium-length explanations were rated highest for clarity, with expert explanations outperforming non-expert ones. This could be caused by expert explanations using jargon, making their explanations more precise.
- **Criteria 5 (Contextual Explanation Quality)**: Medium and long explanations outperformed short and original ones, with expert explanations generally scoring higher. This indicates that longer expert explanations provide more detailed contextual information, whereas non-expert explanations often require additional words to address the question itself, leaving less opportunity to thoroughly discuss the context.

For test 4, Table III summarises the results of a Wilcoxon Signed Rank test comparing participant ratings for each evaluation criterion across original and personalised explanations. This statistical test was selected because it is suitable for paired, non-parametric data, such as Likert scale responses, where the assumption of normality may not hold. The table reports the p-values for each criterion, with statistically significant differences highlighted in bold.

To offer a visual representation of the participant ratings, Appendix VII includes box plots showing the distribution of Likert scale responses for each criterion across original and personalised explanations. These plots help illustrate the

TABLE III: Wilcoxon Signed Rank Test Results for Original vs Personalised Explanation Ratings

Question	Criterion	p-value	Better Version
1	1	0.6148	Personalised
	2	0.1987	Personalised
	3	**0.0215**	Personalised
	4	0.6559	Personalised
	5	**0.0277**	Personalised
	6	0.4537	Personalised
2	1	0.1573	Personalised
	2	0.5271	Personalised
	3	0.1317	Personalised
	4	**0.0017**	Personalised
	5	0.0511	Personalised
	6	**0.0129**	Personalised
3	1	0.5637	Original
	2	0.3951	Personalised
	3	0.2795	Personalised
	4	0.9521	Original
	5	0.1206	Personalised
	6	0.5839	Personalised
4	1	0.7055	Personalised
	2	0.4487	Original
	3	**0.0032**	Personalised
	4	**0.0013**	Personalised
	5	0.1552	Personalised
	6	0.5881	Personalised
5	1	**0.0067**	Personalised
	2	0.4142	Personalised
	3	**0.0044**	Personalised
	4	**0.0009**	Personalised
	5	**0.0240**	Personalised
	6	**0.0049**	Personalised

spread, central tendency, and variability of the ratings, further contextualising the statistical results. Table X in Appendix VIII presents participants' preferences from the survey, indicating whether they favoured the original explanations, the personalised explanations, or expressed no preference (neutral). Furthermore, Table XI in Appendix IX provides an analysis of the semantic similarity between the original and personalised explanations using Cosine similarity metrics.

2) User Study:

- **Criteria 1 (Relevance to Question)**: A significant improvement was observed for question 5. However, since personalisation was not explicitly aimed at improving relevance to the question, this result aligns with expectations and Test 3, where there was a negligible difference.
- **Criteria 2 (Contextual Accuracy)**: No significant differences were observed for any questions, suggesting that participants perceived contextual accuracy to be comparable between the original and personalised explanations. Although Test 3 did show a small improvement when comparing original and expert medium results.
- **Criteria 3 (Enhancement of Understanding) and 4 (Explanation Clarity)**: Significant improvements were observed in three questions. These results align with findings from Test 3, indicating that personalisation enhances participants' understanding and improved the clarity of explanations—two complementary criteria that the personalisation aimed to improve.
- **Criteria 5 (Contextual Explanation Quality) and 6 (Perceived Reliability)**: Significant improvements were observed in two questions. Similar trends were noted in Test 3 for criteria 5, suggesting that personalisation can improve context explanation quality and the perceived reliability.
- **Question 3**: Participant preferences were almost evenly split, with approximately half favouring the personalised explanation for its additional details and context, while the other half preferred the original explanation for its conciseness and clarity. This highlights differences in individual preferences between detail and simplicity in effective communication.
- **Question 5**: Significant improvements were observed across five out of six criteria, demonstrating the potential of personalisation. Participant feedback emphasised a preference for personalised explanations due to their improved structure, inclusion of relevant ROS logs, and conciseness.

Overall, the results show that personalisation can significantly improve explanation quality, as shown by consistent positive trends across LLM (Test 3) and human responses (Test 4). However, the benefits of personalisation are not uniform, as individual preferences, experience with ROS and content characteristics all influence its effectiveness.

D. RQ 3: Feedback Loop Improvement

We conduct one test to determine the effect of using a feedback loop to improve the personalised answers:

- Test 5: Compares personalised explanations with personalised explanations that have been improved using a feedback loop

For test 5, we identify which personalised answers score below a mean rating of 6 for the criteria from test 3. This included expert long (Criteria 1), non-expert short (Criteria 1 and 5), and non-expert long (Criteria 1). GPT-4o was prompted to refine these explanations using textual feedback from the panel of judges. The improved explanations were re-evaluated against the same five criteria to determine the effect on explanation quality.

Table II shows the mean ratings of the improved expert long, non-expert short and non-expert long for the five criteria related to explanation quality. Table VIII in Appendix VI shows the p-values from a t-test comparing the personalised and improved explanations. Values showing significant improvement are highlighted in bold in Table II.

Key results include:

- **Expert Long**: Relevance improved notably (5.91 to 6.39), with slight gains in clarity but minor decreases in other criteria.

- **Non-Expert Short**: Considerable improvements in relevance (5.91 to 6.27) and contextual explanation quality (5.95 to 6.24), with small trade-offs in other criteria.
- **Non-Expert Long**: Relevance showed clear improvement (5.91 to 6.27), with minor gains in clarity and understanding, though with a slight decline in contextual quality.

These results indicate that feedback loops are effective in enhancing targeted explanation quality criteria. However, minor decreases in other criteria suggest a potential trade-off, where improving one aspect might compromise others.

VI. CONCLUSIONS

This study demonstrates that personalisation significantly enhances the suitability and quality of explanations for robot failures based on ROS log messages. Personalised explanations consistently outperformed non-personalised ones in terms of suitability for both target audiences and desired lengths, demonstrating that personalisation effectively aligns responses with user needs. Additionally, tailored explanations achieved significant improvements in quality, validated through both LLM panel ratings and human evaluations. Furthermore, the study shows that feedback loops can further refine explanations, highlighting the potential of iterative improvement processes in explainability systems. Overall, this work underscores the potential of combining PEFT with personalisation to bridge gaps in human-robot interaction, offering scalable and adaptable solutions for explainability in autonomous robotics.

VII. FUTURE WORK

Based on the findings of this study, we suggest the following promising directions for future research:

- **Expanding Personalisation Characteristics**: Investigate beyond expertise level and explanation length to personalise explanations based on additional factors such as the user's experience with ROS, task familiarity, emotional tone preferences, or domain-specific requirements. Furthermore, research into tailoring explanations for individual users rather than predefined groups, for instance by dynamically adapting explanations to fit these needs.
- **Enhanced Feedback Mechanisms**: Explore alternative feedback systems, such as reinforcement learning or human-in-the-loop methods. These methods can iteratively improve explanations while addressing specific criteria without compromising overall quality.
- **Real-World Robot Testing**: Extend testing to physical robots operating in real-world environments. Evaluate how personalised explanations influence user trust, understanding, and interaction in dynamic and practical scenarios, ensuring that the system meets real-world application needs.
- **Multi-Modal Robot Data**: Include more types of robot data in the context used for fine-tuning, such as visual and auditory sensor data, to provide explanations grounded in a richer dataset that can capture more complexities.

References

[1] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)," vol. 6, pp. 52 138–52 160. [Online]. Available: https://ieeexplore.ieee.org/document/8466590/

[2] T. Sakai and T. Nagai, "Explainable autonomous robots: a survey and perspective," vol. 36, no. 5, pp. 219–238. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/01691864.2022.2029720

[3] A. Rosenfeld, "Better metrics for evaluating explainable artificial intelligence: Blue sky ideas track."

[4] G. Vilone and L. Longo, "Explainable artificial intelligence: a systematic review," publisher: [object Object] Version Number: 4. [Online]. Available: https://arxiv.org/abs/2006.00093

[5] A. Rosenfeld and A. Richardson, "Explainability in human–agent systems," vol. 33, no. 6, pp. 673–705. [Online]. Available: http://link.springer.com/10.1007/s10458-019-09408-y

[6] R. R. Hoffman, S. T. Mueller, G. Klein, and J. Litman, "Measures for explainable AI: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-AI performance," vol. 5, publisher: Frontiers. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fcomp.2023.1096257

[7] F. Xu, H. Uszkoreit, Y. Du, W. Fan, D. Zhao, and J. Zhu, "Explainable AI: A brief survey on history, research areas, approaches and challenges," in *Natural Language Processing and Chinese Computing*, ser. Lecture Notes in Computer Science, J. Tang, M.-Y. Kan, D. Zhao, S. Li, and H. Zan, Eds. Springer International Publishing, pp. 563–574.

[8] G. Papagni and S. Koeszegi, "Understandable and trustworthy explainable robots: A sensemaking perspective," vol. 12, no. 1, pp. 13–30. [Online]. Available: https://www.degruyter.com/document/doi/10.1515/pjbr-2021-0002/html

[9] L. Fernández-Becerra, M. A. González-Santamarta, A. M. Guerrero-Higueras, F. J. Rodríguez-Lera, and V. M. Olivera, "Enhancing trust in autonomous agents: An architecture for accountability and explainability through blockchain and large language models." [Online]. Available: http://arxiv.org/abs/2403.09567

[10] M. A. González-Santamarta, L. Fernández-Becerra, D. Sobrín-Hidalgo, A. M. Guerrero-Higueras, I. González, and F. J. R. Lera, "Using large language models for interpreting autonomous robots behaviors," publisher: [object Object] Version Number: 1. [Online]. Available: https://arxiv.org/abs/2304.14844

[11] Z. Xu, S. Jain, and M. Kankanhalli, "Hallucination is inevitable: An innate limitation of large language models." [Online]. Available: http://arxiv.org/abs/2401.11817

[12] M. R. J, K. VM, H. Warrier, and Y. Gupta, "Fine tuning LLM for enterprise: Practical guidelines and recommendations." [Online]. Available: http://arxiv.org/abs/2404.10779

[13] D. Sobrín-Hidalgo, M. A. González-Santamarta, A. M. Guerrero-Higueras, F. J. Rodríguez-Lera, and V. Matellán-Olivera, "Explaining autonomy: Enhancing human-robot interaction through explanation generation with large language models," publisher: [object Object] Version Number: 1. [Online]. Available: https://arxiv.org/abs/2402.04206

[14] L. Fernández-Becerra, M. A. González-Santamarta, D. Sobrín-Hidalgo, A. M. Guerrero-Higueras, F. J. R. Lera, and V. M. Olivera, "Accountability and explainability in robotics: A proof of concept for ROS 2- and nav2-based mobile robots," in *International Joint Conference 16th International Conference on Computational Intelligence in Security for Information Systems (CISIS 2023) 14th International Conference on EUropean Transnational Education (ICEUTE 2023)*, ser. Lecture Notes in Networks and Systems. Springer Nature Switzerland, pp. 3–13.

[15] V. B. Parthasarathy, A. Zafar, A. Khan, and A. Shahid, "The ultimate guide to fine-tuning llms from basics to breakthroughs: An exhaustive review of technologies, research, best practices, applied research challenges and opportunities," 2024. [Online]. Available: https://arxiv.org/abs/2408.13296

[16] B. Zhang, Z. Liu, C. Cherry, and O. Firat, "When scaling meets LLM finetuning: The effect of data, model and finetuning method." [Online]. Available: https://openreview.net/forum?id=5HCnKDeTws

[17] N. Ding, Y. Qin, G. Yang, F. Wei, Z. Yang, Y. Su, S. Hu, Y. Chen, C.-M. Chan, W. Chen, J. Yi, W. Zhao, X. Wang, Z. Liu, H.-T. Zheng, J. Chen, Y. Liu, J. Tang, J. Li, and M. Sun, "Parameter-efficient fine-tuning of large-scale pre-trained language models," vol. 5, no. 3, pp. 220–235, publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s42256-023-00626-4

[18] Z. Han, C. Gao, J. Liu, Jeff, Zhang, and S. Qian Zhang, "Parameter-efficient fine-tuning for large models: A comprehensive survey," publication Title: arXiv e-prints ADS Bibcode: 2024arXiv240314608H. [Online]. Available: https://ui.adsabs.harvard.edu/abs/2024arXiv240314608H

[19] Z. Fu, H. Yang, A. Man-Cho So, W. Lam, L. Bing, and N. Collier, "On the effectiveness of parameter-efficient fine-tuning," publication Title: arXiv e-prints ADS Bibcode: 2022arXiv221115583F. [Online]. Available: https://ui.adsabs.harvard.edu/abs/2022arXiv221115583F

[20] J. Xu and J. Zhang, "Random masking finds winning tickets for parameter efficient fine-tuning," 2024. [Online]. Available: https://arxiv.org/abs/2405.02596

[21] M. Weyssow, X. Zhou, K. Kim, D. Lo, and H. Sahraoui, "Exploring parameter-efficient fine-tuning techniques for code generation with large language models," publication Title: arXiv e-prints ADS Bibcode: 2023arXiv230810462W. [Online]. Available: https://ui.adsabs.harvard.edu/abs/2023arXiv230810462W

[22] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models." [Online]. Available: https://openreview.net/forum?id=nZeVKeeFYf9

[23] J. Wang, G. Yang, W. Chen, H. Yi, X. Wu, and Q. Lao, *MLAE: Masked LoRA Experts for Parameter-Efficient Fine-Tuning*.

[24] S. M. Bsharat, A. Myrzakhan, and Z. Shen, "Principled instructions are all you need for questioning LLaMA-1/2, GPT-3.5/4." [Online]. Available: http://arxiv.org/abs/2312.16171

[25] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica, "Judging llm-as-a-judge with mt-bench and chatbot arena," in *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., vol. 36. Curran Associates, Inc., 2023, pp. 46 595–46 623. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2023/file/91f18a1287b398d378ef22505bf41832-Paper-Datasets_and_Benchmarks.pdf

[26] P. Verga, S. Hofstatter, S. Althammer, Y. Su, A. Piktus, A. Arkhangorodsky, M. Xu, N. White, and P. Lewis, "Replacing judges with juries: Evaluating llm generations with a panel of diverse models," 2024. [Online]. Available: https://arxiv.org/abs/2404.18796

[27] H. Huang, Y. Qu, X. Bu, H. Zhou, J. Liu, M. Yang, B. Xu, and T. Zhao, "An empirical study of llm-as-a-judge for llm evaluation: Fine-tuned judge model is not a general substitute for gpt-4," 2024. [Online]. Available: https://arxiv.org/abs/2403.02839

[28] R. Islam and O. M. Moushi, "Gpt-4o: The cutting-edge advancement in multimodal llm," July 2024, techRxiv Preprint.

[29] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed, "Mistral 7b," 2023. [Online]. Available: https://arxiv.org/abs/2310.06825

[30] A. Roucher, "Using llm-as-a-judge for an automated and versatile evaluation," 2024, accessed: November 23, 2024. [Online]. Available: https://huggingface.co/learn/cookbook/en/llm_judge

APPENDIX

APPENDIX I
MEAN WORD COUNTS

TABLE IV: Mean Word Counts for Personalisation Types

Personalisation Type	**Mean Word Count**
Original	125
Expert short	32.1
Expert medium	118
Expert long	344
Non-expert short	40.1
Non-expert medium	124
Non-expert long	362

APPENDIX II
QUESTIONS FOR EACH SCENARIO

TABLE V: Questions for Each Scenario

Questions for all scenarios:
Q1: How many waypoints were received?*
Q2: Which waypoints were received?*
Q3: What were the coordinates of all the waypoints the robot received?
Q4: Were all the received waypoints successfully reached?*
Q5: What was the first waypoint the robot reached?
Q6: Where was the last waypoint the robot reached?
Q7: Were all paths feasible during the navigation task?
Q8: Were there any waypoints the robot failed to reach?
Q9: Were there any obstacles the robot encountered during the navigation task?
Q10: What is the task the robot had to perform?*
Additional questions for scenario 2 & 4:
Q11: Where was the encountered obstacle located?
Q12: How did the robot respond to detected obstacles?
Additional questions for scenario 3 & 4 & 5:
Q13: Why were one or more paths not feasible?
Q14: How did the robot deal with unfeasible path(s)?

The questions indicated with a * are from the work by Sobrín-Hidalgo et al. [13].

APPENDIX III
PROMPT USED FOR ANSWER GENERATION

```
{"role": "system", "content": "###Instruction### You are an explainability AI
    specialised in analyzing and interpreting ROS log messages for autonomous
    robots. Your task is to provide clear, concise and factual explanations
    based on the logs generated during robot operations. Focus on delivering
    precise answers to the user's questions."}
{"role": "user", "content": "###Context### Based on the following ROS log
    messages: [insert ROS log messages]. ###Question### [insert question]"}}
```

Appendix IV
Scenarios used in questionnaire

TABLE VI: Scenarios with Questions and Corresponding Answers

Original Answer	Expert Medium Answer
1. Which waypoints were received?	
Based on the ROS log messages, the robot received the following waypoints: 1. Waypoint 1: x=1.5, y=-3.0 2. Waypoint 2: x=4.0, y=-7.0 3. Waypoint 3: x=2.0, y=-10.0 The robot successfully navigated to waypoints 1 and 2 but failed to reach waypoint 3. The log messages indicate that the robot attempted to navigate to waypoint 3 but failed to get a valid plan, leading to the execution of recovery behaviors and ultimately an abort of the mission.	The robot received the following waypoints: 1) Waypoint 1: $x = 1.5, y = -3.0$ 2) Waypoint 2: $x = 4.0, y = -7.0$ 3) Waypoint 3: $x = 2.0, y = -10.0$ The robot successfully navigated to waypoints 1 and 2, transitioning through the standard ROS action states from `WAITING_FOR_GOAL_ACK` to `ACTIVE`, and finally to `DONE` after reaching each goal. However, at waypoint 3, the robot encountered planning failures. Despite executing recovery behaviors, including clearing costmaps and initiating a rotate recovery, it was unable to generate a valid path, resulting in mission abort. This suggests potential issues with the local or global costmap configurations or environmental obstacles that were not accounted for.
2. Were all the received waypoints successfully reached?	
Based on the ROS log messages, the following conclusions can be drawn: 1. **Waypoint 1:** - The robot attempted to navigate to waypoint 1 with coordinates (x=2.0, y=1.0). - The robot failed to get a valid plan and executed recovery behaviors, but ultimately aborted the mission due to the inability to find a valid plan. - The log indicates that the robot did not reach waypoint 1. 2. **Waypoint 2:** - The robot successfully navigated to waypoint 2 with coordinates (x=-3.0, y=-5.5). - The robot reached waypoint 2 without any issues. 3. **Waypoint 3:** - The robot successfully navigated to waypoint 3 with coordinates (x=-3.5, y=-12.0). - The robot reached waypoint 3 without any issues. Therefore, out of the three waypoints, only waypoint 2 and waypoint 3 were successfully reached. Waypoint 1 was not reached due to the failure to generate a valid plan. In summary, the robot successfully reached waypoints 2 and 3 but failed to reach waypoint 1.	Based on the ROS log analysis, the robot's waypoint navigation yielded mixed results: 1) **Waypoint 1:** The robot initiated navigation to coordinates ($x = 2.0, y = 1.0$) but encountered planning failures. Despite executing recovery behaviors, including costmap clearing and rotation, it was unable to generate a valid trajectory and aborted the mission. 2) **Waypoint 2:** Successfully navigated to coordinates ($x = -3.0, y = -5.5$) with no reported issues, indicating effective path planning and execution. 3) **Waypoint 3:** The robot reached coordinates ($x = -3.5, y = -12.0$) without complications, confirming successful navigation. In conclusion, the robot successfully reached waypoints 2 and 3 but failed at waypoint 1 due to persistent planning issues despite recovery attempts.
3. Where was the last waypoint the robot reached?	
The last waypoint the robot reached was waypoint 1, with coordinates x = -3.0 and y = 0.5. This is indicated by the log message "Reached waypoint 1". The robot successfully navigated to this waypoint but failed to reach waypoint 2 due to the goal being off the global costmap. Here is the relevant log entry: `Reached waypoint 1`. This confirms that the robot successfully navigated to waypoint 1 but failed to reach waypoint 2. In summary, the robot reached waypoint 1 but failed to reach waypoint 2.	The robot's last successfully reached waypoint was waypoint 1, located at coordinates x = -3.0 and y = 0.5. This is confirmed by the log entry "Reached waypoint 1." The subsequent attempt to navigate to waypoint 2 failed due to the goal being off the global costmap, which led to unsuccessful planning and recovery attempts. The system logs indicate that despite clearing costmaps and initiating recovery behaviors, the navigation to waypoint 2 was aborted. This highlights a common issue in autonomous navigation where the planner cannot generate a valid path due to environmental constraints or map inaccuracies.
4. Were there any obstacles the robot encountered?	

Yes, the robot encountered obstacles during the navigation task. Specifically, the log mentions a new obstacle detected at position (1.5, -3.5). This led to the robot failing to get a plan and replanning its path multiple times. The log also indicates that the robot was nearing the obstacle and had to replan its path. Despite these issues, the robot successfully reached waypoints 2 and 3. Here are the relevant log entries: 1. ”New obstacle detected at position (1.5, -3.5)” 2. ”Failed to get a plan.” 3. ”Robot is nearing the obstacle. Path is being replanned.” 4. ”TebLocalPlannerROS: trajectory is not feasible. Resetting planner...” 5. ”Failed to get a plan.” 6. ”Clearing both costmaps outside a square (3.00m) large centered on the robot.” 7. ”Clearing both costmaps outside a square (1.50m) large centered on the robot.” 8. ”Robot is nearing the obstacle. Path is being replanned.” 9. ”Failed to reach waypoint 1” These entries indicate that the robot encountered an obstacle, which caused it to fail to generate a plan and replan its path. The robot eventually cleared its costmaps and continued navigating, but it failed to reach waypoint 1. However, it successfully reached waypoints 2 and 3. In summary, the robot encountered an obstacle at position (1.5, -3.5), which led to navigation issues and replanning attempts. Despite these obstacles, the robot successfully reached waypoints 2 and 3.

Yes, the robot encountered obstacles during the navigation task. The logs indicate a new obstacle detected at position (1.5, -3.5), which triggered multiple path replanning attempts. Initially, the robot failed to generate a feasible trajectory, as indicated by the ”Failed to get a plan” and ”TebLocalPlannerROS: trajectory is not feasible” entries. The system responded by resetting the planner and clearing costmaps, first within a 3.00m and then a 1.50m radius, to manage the local environment more effectively. Despite these interventions, the robot was unable to reach waypoint 1. However, it successfully navigated to waypoints 2 and 3 after overcoming the initial obstacle-related challenges. This sequence highlights the robot’s adaptive navigation capabilities in response to dynamic environmental changes.

5. What is the task the robot had to perform?

Based on the ROS log messages, the robot’s task was to navigate to a series of waypoints. The robot received goals to move to specific coordinates (x, y) and attempted to reach these waypoints. The log messages indicate the following: 1. **Waypoint 1:** The robot was sent to coordinates (x=5.5, y=0.0). The goal was off the global costmap, and the robot failed to find a valid plan, resulting in an abort. 2. **Waypoint 2:** The robot was sent to coordinates (x=3.0, y=-7.0). The goal was reached successfully. 3. **Waypoint 3:** The robot was sent to coordinates (x=-2.5, y=0.0). The goal was reached successfully. In summary, the robot’s task was to navigate to a sequence of waypoints, and it successfully reached waypoints 2 and 3 but failed to reach waypoint 1. Here is a summary of the tasks: 1. **Waypoint 1:** Navigate to (x=5.5, y=0.0). Failed due to planning issues. 2. **Waypoint 2:** Navigate to (x=3.0, y=-7.0). Successfully reached. 3. **Waypoint 3:** Navigate to (x=-2.5, y=0.0). Successfully reached. The robot’s primary task was to navigate to these waypoints, and it successfully completed two of the three goals.

Based on the ROS log messages, the robot’s task was waypoint navigation using the move_base action server. It was tasked with reaching specific coordinates in a sequence. Here’s a breakdown of the task execution:

1) **Waypoint 1:** The robot attempted to navigate to (x=5.5, y=0.0). However, the goal was outside the global costmap boundaries, leading to planning failures despite recovery behaviors. The task was aborted as no valid path could be generated.
2) **Waypoint 2:** Successfully navigated to (x=3.0, y=-7.0). The transition from WAITING_FOR_GOAL_ACK to ACTIVE and then to DONE indicates successful goal execution and completion.
3) **Waypoint 3:** Successfully reached (x=-2.5, y=0.0). Similar to waypoint 2, the robot transitioned through the expected states, confirming successful navigation.

The robot’s primary task was to autonomously navigate through these waypoints, demonstrating effective use of the navigation stack for waypoints 2 and 3, while encountering a critical planning issue with waypoint 1 due to costmap constraints.

Appendix V
Individual judge ratings

TABLE VII: Mean Suitability Ratings for Target Audience for Different Personalizations

Personalisation	Panel Mean	Std. Dev	GPT-4o		Mistral-Large		Llama3-8B	
			Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
Expert Suitability								
Non-expert short	4.66	0.623	2.66	0.791	4.93	1.02	6.38	0.717
Non-expert medium	3.87	0.731	2.17	0.428	3.53	1.06	5.91	1.35
Non-expert long	2.36	0.597	1.92	0.268	2.18	0.441	2.99	1.50
Non-expert Suitability								
Original	4.10	0.522	3.29	0.758	3.67	0.893	5.37	0.589
Expert short	3.66	0.726	2.83	0.601	3.56	0.944	4.59	1.17
Expert medium	3.20	0.595	2.39	0.489	2.94	0.507	4.28	1.24
Expert long	2.45	0.680	1.19	0.376	2.28	0.517	3.18	1.51
Short Suitability								
Original	4.03	0.900	3.59	1.21	3.49	1.01	5.01	1.48
Expert short	5.89	0.517	6.24	0.498	5.89	0.428	5.54	1.1
Non-expert short	5.74	0.548	6.03	0.488	5.61	0.645	5.57	1.02
Long Suitability								
Original	5.33	0.972	5.00	1.35	5.82	0.687	5.17	1.35
Expert short	3.52	0.998	1.87	0.440	4.55	1.37	4.16	1.86
Non-expert short	3.30	0.847	1.98	0.408	4.61	1.19	3.33	1.46

Appendix VI
Significance Paired T-Test Test 3

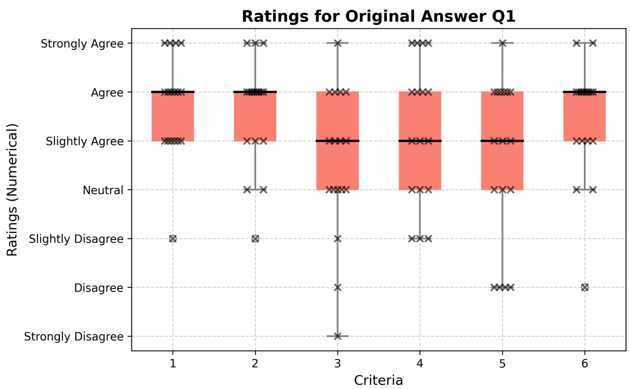
TABLE VIII: Mean Ratings and p-values for Personalisation Types Across Evaluation Criteria

Personalisation Type	Crit 1: Relevance to Question		Crit 2: Accuracy Accuracy		Crit 3: Enhancement of Understanding		Crit 4: Explanation Clarity		Crit 5: Contextual Explanation Quality	
	Mean	**p-value**	**Mean**	**p-value**	**Mean**	**p-value**	**Mean**	**p-value**	**Mean**	**p-value**
Original	6.49	-	6.51	-	6.62	-	6.28	-	6.48	-
Expert short	6.36	1.12e-02	6.55	0.361	6.68	0.256	6.45	**2.33e-04**	6.49	0.741
Expert medium	6.27	1.66e-05	6.75	**1.69e-07**	6.88	**1.51e-07**	6.55	**8.83e-08**	6.84	**3.79e-17**
Expert long	5.93	5.51e-12	6.56	0.244	6.81	**1.77e-04**	6.20	0.152	6.75	**9.48e-10**
Expert long (improved)	6.39	**6.31e-12***	6.52	0.231*	6.88	**4.64e-03***	6.51	**3.79e-13***	6.75	0.976*
Non-expert short	5.91	1.48e-17	6.28	6.06e-06	6.06	1.12e-20	6.08	9.20e-04	5.95	9.73e-23
Non-expert short (improved)	6.27	**2.53e-11***	6.34	0.158*	6.41	**1.57e-15***	6.40	**1.24e-13***	6.24	**7.22e-12***
Non-expert medium	6.26	5.90e-06	6.45	0.125	6.53	6.46e-02	6.40	**2.32e-02**	6.43	0.225
Non-expert long	5.53	3.31e-26	6.29	1.09e-06	6.38	4.38e-06	6.14	1.20e-02	6.55	6.69e-02
Non-expert long (improved)	6.27	**1.31e-17***	6.34	0.114*	6.52	**2.42e-05***	6.40	**5.03e-10***	6.45	1.25e-04*

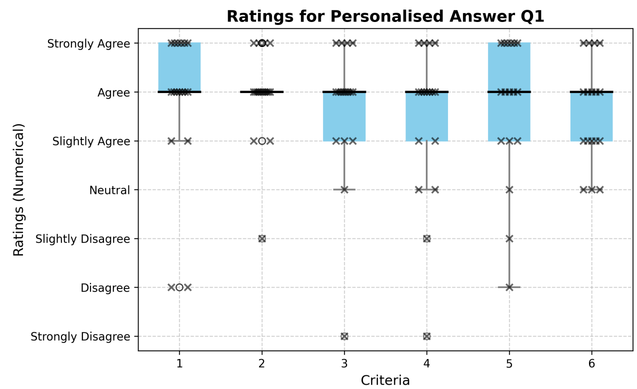
APPENDIX VII
BOX PLOTS

Criteria number	Criteria name
1	Relevance to question
2	Contextual accuracy
3	Enhancement of understaning
4	Answer clarity
5	Contextual explanation quality
6	Perceived reliability

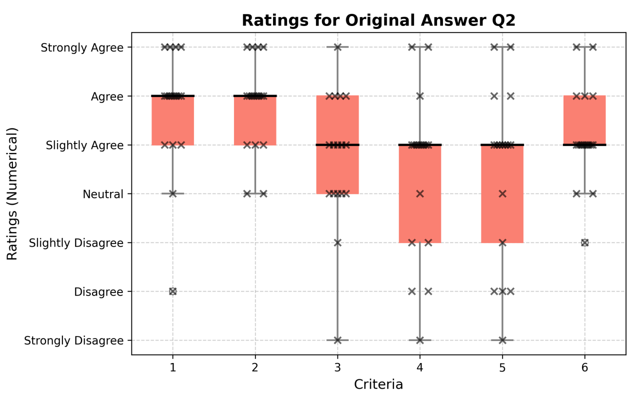
TABLE IX: Overview of criteria for box plots



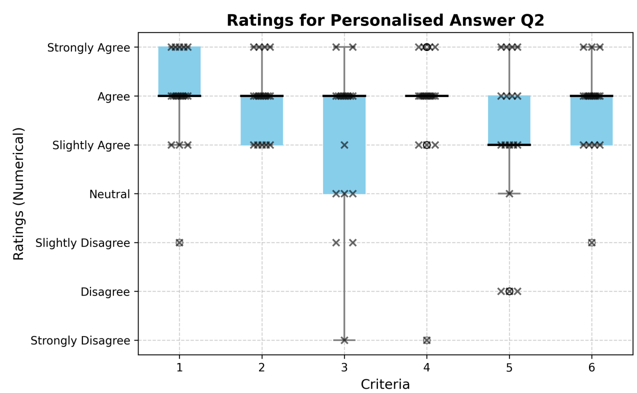
(a) Ratings per criteria for the original answer for question 1



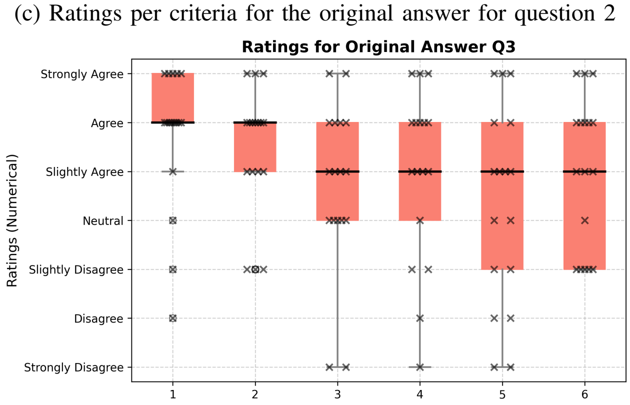
(b) Ratings per criteria for the personalised answer for question 1



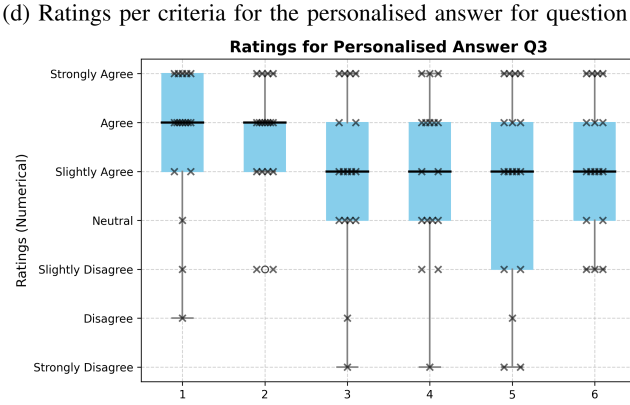
(c) Ratings per criteria for the original answer for question 2



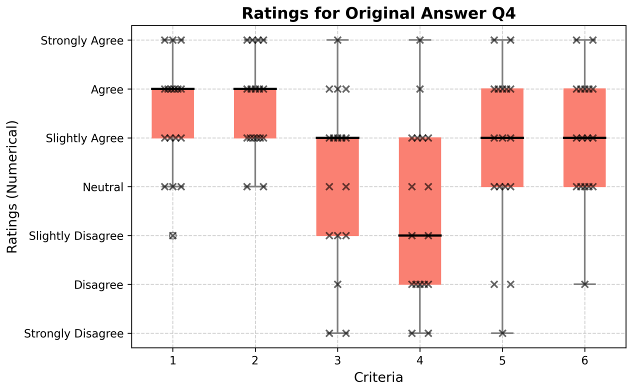
(d) Ratings per criteria for the personalised answer for question 2



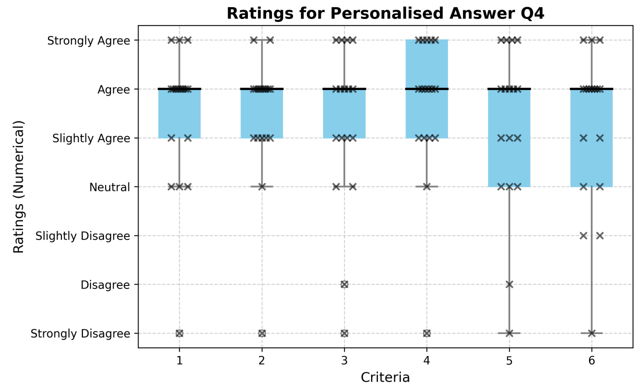
(e) Ratings per criteria for the original answer for question 3



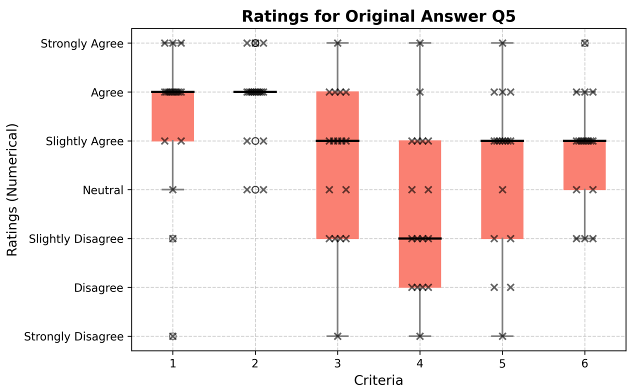
(f) Ratings per criteria for the personalised answer for question 3



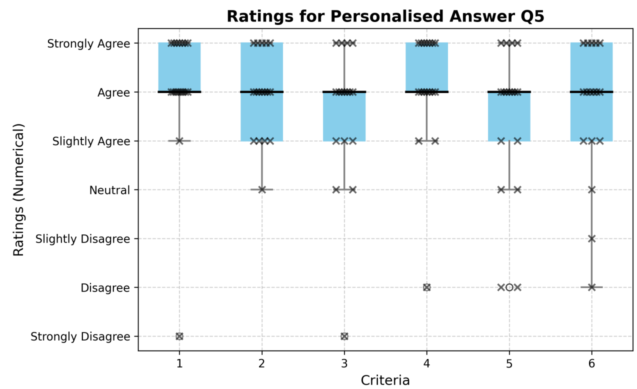
(a) Ratings per criteria for the original answer for question 4



(b) Ratings per criteria for the personalised answer for question 4



(c) Ratings per criteria for the original answer for question 5



(d) Ratings per criteria for the personalised answer for question 5

APPENDIX VIII
USER PREFERENCES FROM QUESTIONNAIRE

	Preferences			% prefer
Question	Original	Personalised	Neutral	personalised explanation
1	5	12	0	70.1%
2	2	13	2	76.5%
3	8	7	2	41.2%
4	4	12	1	70.1%
5	1	16	0	94.1%

TABLE X: Preferences of explanations of questionnaire respondents per question

APPENDIX IX
COSINE SIMILARITY RESULTS

Question	**Cosine Similarity**
1	0.6395
2	0.6878
3	0.6847
4	0.7358
5	0.7814

TABLE XI: Cosine similarity results for the provided question pairs.