Templated vs. Generative: Explaining Robot Failures

Gregory LeMasurier[1], Christian Tagliamonte[1], Jacob Breen[1], Daniel Maccaline[1], and Holly A. Yanco[1]

***Abstract*— The need for robots to explain their failures grows as the variety and number of robots deployed in public, homes, and work environments increases. This paper extends our prior work utilizing explanation templates by comparing those *Templated* explanations to *Generative* explanations created by a Large Language Model. Our study surprisingly reveals that Templated explanations result in similar or higher perceived intelligence and trust while also being more understandable. Through our findings, we aim to provide considerations for effective robot explanation systems, ultimately enabling people to be able to understand and provide assistance to robots that have encountered unforeseen circumstances.**

I. INTRODUCTION

As more robots are integrated into our daily lives, it becomes infeasible for experts to monitor every robot to provide assistance when they encounter unforeseen circumstances. Therefore, it becomes increasingly necessary that robots are able to detect and communicate their failures to people who may be nearby.

People working around robots prefer that explanations are provided when the robot needs to modify its plan or is incapable of completing its task [1]. Explanations have been shown to foster increased trust [2], [3], [4], [5], [6], transparency [3], understandability [3], [7], [8], [9] and team performance [3]. Explanations from proactive systems [10], [11], [12] lead to more understandability and better subjective ratings compared to reactive systems [12].

Explainable robot systems should adapt to the recipients' roles and experience [13] as well as their task performance [14]. These explanations should provide sufficient (but not overwhelming) detail so that non-experts can understand and act upon [15] the robot's explanation.

Many different modalities can be used by explainable systems, including visual (e.g., graphics, images, and plots) [16], [17], [18], motion [4], [17], and natural language (e.g., rules and numeric responses) [16], [17], [18]. In this work we focus on language-based explanations.

Several studies [7], [8], [9] have explored different explanation structures for language-based explanations. Including reasons for failures in explanations helps improve understandability and helpfulness [7]. Providing an action history and potential reasons for failures has been shown to improve understandability [8], [9], desirability [9], and enables non-experts to detect and solve a robot's errors [8].

One method for generating robot explanations is through training machine learning models, which are designed for the specific task of explaining failures [5], [8]. A slightly more generalizable approach is to use explanation templates [7], [12], [19], though these templates can be grammatically incorrect in some cases [12], potentially hindering understandability and trust in the robot. Large Language Models (LLMs) can also be used to generate explanations for robot failures [20].

LLMs can complete a variety of language understanding and generation tasks and perform well on zero shot reasoning [21]. Recently, we have seen many applications for LLMs in robotics [22] and human-robot interaction (HRI) [23]. Few-shot learning techniques can be applied during prompting to enable the LLM to complete a variety of tasks, without the need to fine-tune for the desired task [24]. Therefore, LLMs have been applied to a variety of applications such as social robotics [25], motion control [26], task planning [27], [28], interpreting robot log files [29], and expressive behaviors [30]. LLMs can also be prompted to generate explanations [20], [31], [32], [33] and to answer follow up questions [20]. While LLMs can be used as powerful text generation tools, consideration and understanding of the appropriateness of using LLMs in HRI is necessary [34].

Templated and generative approaches are also common in Natural Language Generation (NLG) [35]. It is important to consider when generative approaches add value compared to simpler approaches such as using templated text [36]. While generative systems offer fluency, potentially at a trade-off for accuracy, it has been argued that accuracy is more important than fluency in NLG [37]. Based on our previous observations [12], we hypothesized that fluency might play a larger role in HRI because of the robot embodiment.

Our work contributes to the existing literature by conducting a user study to evaluate both Templated and Generative explanation methods which are communicated through a robot embodiment. Our analysis provides insights into people's perception of the two types of explanation systems, the understandability of the explanations, and their perception of the timing of each system's explanations.

II. SYSTEM DESIGN

As the foundation of a system designed to provide explanations of robot failures, we have utilized Behavior Trees

*This research has been funded by the Office of Naval Research (N00014-18-1-2503 and N00014-23-1-2744). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Office of Naval Research.

[1]Richard A. Miner School of Computer & Information Sciences, University of Massachusetts Lowell, Lowell, MA, USA

Corresponding author:
gregory_lemasurier@student.uml.edu

(BTs) [38], which are a task sequence and execution method that can be used to represent a robot's internal states and actions. This architecture is very effective for automatically generating robot explanations [39]. BTs have been framed into semantic sets: {goal, subgoals, steps, actions} which enables hierarchical explanations and also the generation of answers to follow-up questions when asked for clarifying details by users, an explanation method preferred by users [15].

BTs can be complimented with Assumption Checkers (ACs) [40], [41], which track information about the robot's internal states or environmental conditions throughout task execution. These ACs can be used to estimate system performance [40], [41] and take corrective actions [42], ultimately resulting in a proactive failure detection and explanation system [12]. Proactive explanation systems are capable of predictively detecting and explaining failures before they occur, ultimately resulting in better human perception and more understanding of the robot's failure [12].

BTs can leverage Robot and Object profiles to abstract out robot and object specific information, promoting a robot and task agnostic design [20]. Abstracting out robot and object specific information also enables the system to easily reference this information when generating explanations.

In this work, we compare explanations generated from templates to those generated by GPT-4 using the prompt template that we have previously evaluated through a proof of concept analysis [20].

III. HYPOTHESES

We proposed a set of hypotheses when we described our user study plans to evaluate our new explanation framework [20]. We have since revised our study design to remove the human crafted explanation condition since the templated condition already includes segments of human crafted phrases. The base condition has also been removed as we used the exact scenario from our prior study [12] where we had shown that proactive systems result in higher rated human perception and higher explanation quality compared to the base. We have adjusted our hypotheses to reflect these changes. Additionally, we have added sub-hypothesis 2b to have a full set of comparable results to our prior study [12].

***Hypothesis 1 (Human Perception)*:** The Generative System (GEN) will have higher ratings for perceived intelligence and trustworthiness compared to the Templated System (TEM). When reviewing the free response questions from our prior work [12], we found that several participants commented on the fact that our systems used templates and were not grammatically correct. One participant even stated: "I lose a little faith in a supposed smart robot when its explanations aren't spoken in correct English." Therefore, we hypothesize that the GEN explanations, which are grammatically correct, will be perceived better than the TEM explanations, as measured by perceived intelligence and trustworthiness.

***Hypothesis 1a (Perceived Intelligence)*:** GEN explanations, which are grammatically correct, will be perceived as more intelligent compared to the TEM explanations.

***Hypothesis 1b (Perceived Trustworthiness)*:** The grammatically correct explanations from GEN will result in the system being perceived as more trustworthy than TEM.

***Hypothesis 2 (Explanation Quality)*:** GEN and TEM will perform on par with each other in regards to explanation quality as measured through understandability and the timing of the explanations.

***Hypothesis 2a (Understandability)*:** GEN and TEM explanations will perform on par with each other. Prior studies have shown that generated explanations can perform on par with human crafted explanations [8]. Both GEN and TEM produce explanations which utilize an action history with context, but they notably use different language to communicate this information. Therefore, because both explanations contain the same type of information, we hypothesize that there will not be differences in understandability across system conditions.

***Hypothesis 2b (Explanation Timing)*:** There will not be differences in terms of the explanation timing across the GEN and TEM systems. Both use the exact same videos, with replaced audio explanations, and therefore do not have any actual differences in explanation timing.

IV. METHODS

A. Participants and Power Analysis

We developed a mixed online user study to evaluate our explanation systems. An *a priori* power analysis using G*Power 3.1.9.7 [43] Goodness-of-fit test was conducted to determine our sample size. The parameters used in this analysis include Degree of Freedom = 1, a large effect size w of 0.5, α error probability = 0.05, Power ($1 - \beta$ error probability) = 0.95. This power analysis determined that we needed to recruit 52 participants for each of the two conditions (104 participants).

A total of N = 135 participants were recruited through Prolific. Participants were selected from a standard sample across all available countries and pre-screened to be fluent in English and to have a 100% approval rate across more than 100 but fewer than 10,000 prior submissions. We recruited a total of 31 extra participants to account for people who may have failed attention check questions. A total of 23 participants failed attention check questions, and were excluded from our study. The last participants recruited for each condition were removed until all conditions had an equal number of participants and to match the 104 participants as determined through our power analysis. This resulted in the removal of eight participants.

Out of the 104 included participants, 66 identified as male, 37 as female, and 1 as non-binary. The participants were aged from 18–74 ($M = 32.47, SD = 11.87$). Participants were then asked to rate their agreement on a seven point Likert-type item, ranging from "Strongly Disagree" (1) to "Strongly Agree" (7), "I am experienced with robots" where 31 rated higher than neutral agreement ($M = 3.42, SD = 1.62$). This study was approved by the University of Massachusetts Lowell's Institutional Review Board.

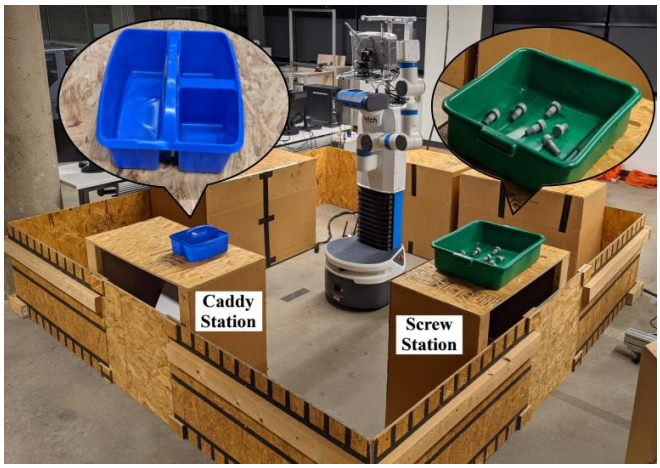

Fig. 1. The FetchIt! challenge arena [44]. The Fetch robot navigates to the screw station to pick a screw from the green bin, navigates to the caddy station, and places the screw in the blue caddy.

B. Scenario and Task

In this study, we follow the same task and scenario as used in our prior work [12]; by reusing them, our results can be compared across studies. The study uses a scenario of a manufacturing company who produces kits for gearboxes using a combination of people and newly introduced robots. The Fetch mobile manipulator robot [45] with a 7 degree-of-freedom arm and an RGBD camera, as seen in Figure 1 starts in the center of the human-robot shared assembly line. The experimenter played the role of a worker whose job was to sort gearbox parts and fill up part of a caddy to create a gearbox kit. The robot would complete the kit by driving around, picking up, and placing the final screw into a caddy.

The participants' job was to observe the newly introduced robot and evaluate its performance on the assembly line. The videos started with the experimenter and robot working on their respective tasks in the shared workspace. Then the experimenter acted as a *challenger* who manipulated the environment causing potential real-world failures (e.g., placing the caddy in a location that the robot can not reach). The robot then proactively detected the failure and provided an explanation when an assumption was violated.

C. Conditions

The study was a mixed 2 (System Type: *Templated (TEM)*, *Generative (GEN)*) $\times$ 2 (Failure Type: *Screw Bin Empty*, *Caddy Out of Reach*) design. As we only needed two failure conditions for this study, to match the two System Type conditions, we selected the two Failure Types that were most confused by participants in our prior study [12], ensuring that the robot's explanation rather than the participants' observations influenced their understanding. This resulted in four videos[1] of the robot interacting with a person which reflected the four possible combinations of System Type $\times$ Failure Type. Across the two videos watched by an individual participant, they experienced both explanation and Failure Types, one of each per video. Each participant had a randomly assigned configuration, and the ordering of the scenarios was counterbalanced to reduce ordering effects.

The Fetch robot was running a proactive explanation system utilizing the BT and AC framework described in [12]. The only difference between System Types is the method used to generate explanations, as described below.

Templated (TEM): TEM gives explanations using the same explanation template as the proactive condition in our prior study [12]. This explanation template is filled out using information from the behavior tree and its assumption checkers. Templated explanations enable consistent and robust explanations for different failures, without having to hand-craft explanations for every possible failure that could occur. Our TEM used the following explanation template: "[*Assumption checker description*] so I will not be able to [*failed sub goal node name*]". As the node names were directly inserted into our templates, they did not always grammatically fit, as seen in Table I.

Generative (GEN): GEN gives explanations using a LLM (GPT-4). This system utilizes information from the robot and object profiles [20], as well as the behavior tree and its assumption checkers to fill out a prompt template, as previously described in [20]. For this condition, we selected the first explanation generated by our system for each Failure Type. The prompt guided the LLM to generate an explanation that indicated what and why the robot failed, to ensure that our generated explanations aligned with the suggestions from prior work on explainable robots [7], [8], [9].

D. Procedure

The procedure followed in this study is almost identical to the procedure followed in our prior study [12]. Once participants signed up to participate in our study through Prolific, they were redirected to our Qualtrics survey. Each participant would read and fill out an informed consent form, then answer a set of demographics questions. Next, participants listened to an audio clip and were required to select the corresponding phrase from a list before continuing, to ensure that the participants' audio was enabled and working. (The videos also had captions to help people interpret the robot's speech.) Then participants read through the description of the scenario, watched a slide show video describing the robot's capabilities, and viewed a labeled image of the arena to understand what was going on in the videos. After this introduction to the study, participants would watch two videos according to their randomly assigned condition.

After each video, participants were asked to respond to a questionnaire which evaluated their experience with the robot. Participants first responded to a simple attention check question, such as what color band did the robot have on its hand, to make sure that the participant was paying attention. Next participants responded to a set of questions consisting of the metrics described in Section IV-E. After watching both videos, participants were asked to indicate which system they preferred. We anticipated that participants would take around

[1] youtube.com/watch?v=BBxQ5_ozEcg&list=PLIwwT33Qq2HStj1RxY414uX7pOfg_ooD3

TABLE I

EXPLANATIONS USED FOR EACH CONDITION.

System Type	Screw Bin Empty	Caddy Out of Reach
Templated (TEM)	I do not see any screws on the table so I will not be able to pick screw.	My arm can not reach the caddy so I will not be able to place object into caddy.
Generative (GEN)	I couldn't find the screw on the table, so I'm unable to pick up the screw right now.	I couldn't generate a path to the caddy, so I'm unable to position the screw for placement in the caddy.

30 minutes to complete the study; the median completion time was 23 minutes. Participants were compensated with USD $7.50 at an hourly rate of $15 through Prolific for their time and effort.

E. Measures

1) ***Perceived Intelligence****:* To measure the perceived intelligence of each system, we used the scale proposed by Warner and Sugarman. Participants responded on 7-point response scaling from "Strongly Disagree" (1) to "Strongly Agree" (7) after each of the two videos that they watched. Internal consistency was measured for each administration of the scale, Cronbach's $\alpha = 0.81$ for the first administration and 0.89 for the second. Responses for each element were averaged to create a perceived intelligence score.

2) ***Trustworthiness****:* To evaluate the perceived trustworthiness of each system, we asked participants to complete a Muir Trust Scale [46] after watching each of the two videos. This scale measures performance-based trust through four constructs: predictability, reliability, competence, and overall trust. Participants responded on a 7-point response scaling from "Strongly Disagree" (1) to "Strongly Agree" (7). Internal consistency was measured for each administration of the scale, for administration one, Cronbach's $\alpha = 0.86$ and 0.87 for the second administration. The responses to each element of the scale were averaged to create a perceived trustworthiness score. By measuring trust with Muir's scale, we have directly comparable results to our prior study [12].

While Muir's trust scale is a well established metric, newer trust scales also exist, such as MDMT v2 [47], which evaluates moral and capacity trust. Participants filled out the full MDMT v2 trust scale after watching each of the two videos. Participants responded on a 6-point response scaling from "Not at all" (0) to "Very Much" (5), and they could select "Does not fit" if they believed a concept was not applicable. Internal consistency was measured for each administration of the scale, for moral trust Cronbach's $\alpha = 0.95$ and 0.95; for capacity trust, Cronbach's $\alpha = 0.91$ and 0.94. The responses to each element of the scale were averaged to create a perceived trustworthiness score, after removing all "Does not fit" responses which were treated as missing values [47].

3) ***Understandability****:* Understandability of explanations was evaluated through a multiple choice question about what participants believed was the cause of failure: "Please select the option that best matches what failures or errors you observed in the video."

Participants were asked two 7-point responses scaling from "Strongly Disagree" (1) to "Strongly Agree" (7): "The robot's explanation changed my initial understanding of the robot's failure" and "The robot's explanation helped me understand its failure." This aimed to evaluate whether the explanation enhanced understanding, as participants could have used their observations when identifying failure's cause.

Finally, for more insight on the participants' responses, we asked participants to rate their *confidence*, from "Very Unsure" (1) to "Very Confident" (7), with their response to the multiple choice item regarding the cause of failure. This question was designed to help us identify if participants guessed the cause of failure or if they believed that they understood. When designing explanation systems, it is also important that they are clear so that people can confidently identify how to resolve issues rather than needing to guess.

4) ***Explanation Timing****:* Participants were asked to respond to two 7-point responses scaling from "Strongly Disagree" (1) to "Strongly Agree" (7) regarding explanation timing: "When something went wrong, the robot explained so at an appropriate time" and "The robot should have explained that something went wrong sooner". These inverses were used to ensure a reliable measure.

5) ***Preference****:* As a supplemental analysis, we investigated which explanation participants preferred. To do so, we asked the following multiple choice question: "In the two videos that you watched, each had the robot giving a different explanation. Which did you prefer?" The answers that participants could select from consisted of the two different explanations in the same order that they saw them based on their randomly assigned conditions. We also provided a third option "Other" which allowed participants to respond in a free response format.

6) ***Follow Up Questions****:* Participants responded to a free response question asking them to list any questions they would have liked to ask the robot based on its explanation. We analyzed these free responses to identify the number of questions that participants would like to ask. The answers will be used to inform a future extension of this work.

V. RESULTS

A. Human Perception

To evaluate H1, we conducted two independent sample t-tests, one for each Failure Type, to evaluate the between subject effects of System Type on the corresponding dependant variables. For each of these tests, we applied a Bonferroni correction to adjust for family-wise error, where our criterion for statistical significance was adjusted by setting $\alpha = 0.025$.

1) Perceived Intelligence: First, we evaluated the perceived intelligence of the System Types. We did not observe significant differences between System Types for the *Screw*

Bin Empty Failure Type ($t(102) = -0.05, p = 0.962$). We did however, observe significant differences between TEM ($M = 4.90, SD = 1.10$) and GEN ($M = 4.24, SD = 1.18$) for the *Caddy Out of Reach* Failure Type ($t(102) = -2.92, p = 0.004$). Therefore, with TEM and GEN performing similarly in the *Screw Bin Empty* scenario and with TEM being perceived as more intelligent than GEN in the *Caddy Out of Reach* scenario, we did not find support for H1a.

2) Trustworthiness: We then evaluated Muir's trust scale. We did not observe significant differences between System Types for the *Screw Bin Empty* Failure Type ($t(102) = -1.77, p = 0.079$). After Bonferroni correction, we did not observe significant differences between TEM ($M = 4.82, SD = 1.07$) and GEN ($M = 4.38, SD = 1.17$) for *Caddy Out of Reach* ($t(102) = -2.02, p = 0.046$).

Next, we investigated the MDMT capacity and moral trust scales. We did not observe significant differences between System Types for the *Screw Bin Empty* Failure Type ($t(102) = -1.26, p = 0.21$). We observed significant differences between TEM ($M = 3.51, SD = 0.99$) and GEN ($M = 3.01, SD = 1.01$) for the *Caddy Out of Reach* Failure Type ($t(102) = -2.54, p = 0.013$).

We did not observe a significant difference between System Types for moral trust, across both Failure Types, *Screw Bin Empty*: ($t(95) = 0.66, p = 0.514$) and *Caddy Out of Reach*: ($t(93) = -1.13, p = 0.26$). Therefore, H1b is not supported as GEN was rated as less trustworthy or equally trustworthy across all three evaluated trust sub-scales.

B. Explanation Quality

1) Understandability: To evaluate participants' understanding of the robot's failures, we analyzed their responses to the multiple choice question on the **cause of failure**. The confusion matrix in Figure 2 shows the full distribution of responses to this question. We performed chi-squared goodness of fit tests and found that the distribution of our responses was not equivalent to random choice ($p < 0.0001$) for each condition. For each Failure Type, we compared the distributions of responses across System Types using Fisher's exact test for count data with a false discovery rate correction. Then we further investigated frequency of participants' responses to understand these differences.

For the *Screw Bin Empty* scenario, a Fisher's exact test for count data found a significant relationship ($p = 0.012$). As highlighted in blue in the first column of Figure 2, the true cause of the failure was "There were no screws in the bin on the table". In TEM, 69.23% of responses were correct, and the most confused answer was "The robot could not detect the screws that were in the screw bin" at 19.23%. In GEN, the most selected answer was "The robot could not detect the screws that were in the screw bin" at 46.15%, then 46.15% with the correct response.

As seen in the confusion matrix, the most popular incorrect choice was "The robot could not detect the screws that were in the screw bin". This shows that while participants were incorrect, they showed that they partially understood the robot's failure.

For the *Caddy Out of Reach* scenario, a Fisher's exact test for count data found a significant relationship ($p < 0.0001$). In the second column in Figure 2, the true cause of the failure is highlighted in orange: "The robot could not place a screw in the caddy because the caddy was too far away". In TEM, 78.85% of responses were correct, and the most confused answer was "The robot could not detect the caddy that was on the table" at 9.62%. In GEN, the most selected answer was "The robot could not detect the caddy that was on the table" at 38.46%, then 36.54% with the correct response.

Again, the most popular incorrect selection was "The robot could not detect the caddy that was on the table", showing that participants partially understood the robot's failure.

2) Influenced Understanding: As in our prior work [12], we asked participants two questions to isolate the impact that the explanation had on their understanding. First participants respond to the question: "The robot's explanation changed my initial understanding of the robot's failure." This question was used to ensure that participants are understanding the failure due to the robot's explanation and not because of visual observations that they made while watching the videos. We did not observe a significant difference between systems as shown by Mann-Whitney U-tests across both the *Screw Bin Empty* Failure Type ($U = 1377, p = 0.871$) and the *Caddy Out of Reach* Failure Type ($U = 1446, p = 0.537$). Figure 3a. shows the distribution of responses to this question.

To determine if the explanations had a positive or negative influence on a person's understanding of the robot's failure, we analyzed participants' responses to "The robot's explanation helped me understand its failure." We conducted two Mann-Whitney U-tests, one for each Failure Type, and included a Bonferroni correction to adjust for family-wise error, where our criterion for statistical significance was adjusted by setting $\alpha = 0.025$. We did not observe significant differences between System Types for the *Screw Bin Empty* Failure Type ($U = 1389, p = 0.94$), however we did notice significant differences in the *Caddy Out of Reach* Failure Type ($U = 921, p = 0.004$) where TEM ($M = 5.96, SD = 1.31, Mdn = 6$) was rated as having helped participants understand the failure more than GEM ($M = 5.19, SD = 1.55, Mdn = 6$). A distribution of responses to this question can be found in Figure 3b.

3) Confidence: To evaluate participants' confidence in their selection to the understandability question, we conducted two Mann-Whitney U-tests, one for each Failure Type, and included a Bonferroni correction to adjust for family-wise error, where our criterion for statistical significance was adjusted by setting $\alpha = 0.025$. We did not observe significant differences between System Types for the *Screw Bin Empty* Failure Type ($U = 1350, p = 0.992$), and after Bonferroni correction we did notice significant differences in the *Caddy Out of Reach* Failure Type ($U = 1050, p = 0.035$) where TEM ($M = 6.37, SD = 0.84, Mdn = 7$) was rated as having helped participants understand the failure more than GEM ($M = 5.94, SD = 1.11, Mdn = 6$). This shows that participants were equally confident in their selections regard-

Cause of Failure		Screw Bin Empty			Caddy Out Of Reach	
		Templated	Generative		Templated	Generative
An obstacle was blocking the robot from driving to the table					1.92%	9.62%
The robot hit an obstacle while driving around						
There were no screws in the bin on the table		**69.23%**	**46.15%**			
There was no caddy on the table						5.77%
The robot could not detect the screws that were in the screw bin		19.23%	**46.15%**		1.92%	
The robot could not detect the caddy that was on the table		3.85%	1.92%		9.62%	**38.46%**
The screw bin was moved while the robot was looking for screws						1.92%
The caddy was moved while the robot was looking for it			1.92%			5.77%
The robot could not reach the screws because they were too far away					5.77%	
The robot could not place a screw in the caddy because the caddy was too far away					**78.85%**	36.54%
The robot reached for a screw, but missed						
The robot dropped the screw						
The robot's arm malfunctioned					1.92%	
The robot's camera malfunctioned		3.85%	3.85%			
Not sure						
I did not observe any failures or errors						
Other		3.85%				1.92%

Fig. 2. Confusion matrix for each Failure Type across System Types. The row corresponding to the ground truth choice is outlined in bold and highlighted with the corresponding color for each Failure Type, for example, the correct cause of failure for Screw Bin Empty is highlighted in blue and the correct cause of failure for Caddy Out of Reach is highlighted in orange. The percentage of responses is indicated in each cell, or is blank for no responses. The bold percentage number in each explanation column indicates the most selected response. The cells are filled from grey (low number of responses) to purple (high number of responses).

less if explanations are templated or generative. Figure 3c. shows the distribution of responses.

4) Explanation Timing: To evaluate the timeliness of our explanations, we asked participants to respond to two questions: (Q1): "When something went wrong, the robot explained so at an appropriate time" and (Q2): "The robot should have explained that something went wrong sooner". We asked two versions of this question to ensure response consistency as we were conducting an online user study. No significant differences between System Types were observed through through Mann-Whitney U-tests across both the *Screw Bin Empty* Failure Type Q1: ($U = 1181, p = 0.244$) Q2: ($U = 1241, p = 0.467$) and the *Caddy Out of Reach* Failure Type Q1: ($U = 1196.5, p = 0.293$) Q2: ($U = 1500.5, p = 0.328$). A distribution of responses to this question can be found in Figure 3d and 3e.

C. Preference

Participants were asked to select the explanation that they most preferred from a list containing both of the explanations that they had observed in their two videos. There were significant differences in the distribution of responses from random chance ($\chi^2(1) = 9.707, p = 0.002$). A total of 65 participants preferred templated explanations, 34 preferred the generative explanations, and 5 selected "Other".

D. Follow Up Questions

Participants were asked to list any follow up questions that they would have liked to ask the robot. We conducted two independent sample t-tests, one for each Failure Type, to evaluate the effects of System Type on the number of follow up questions asked. We did not observe a significant difference between System Types, across both Failure Types, *Screw Bin Empty*: ($t(102) = 1.97, p = 0.052$) and *Caddy Out of Reach*: ($t(102) = 0.0, p = 1.00$). Follow up questions involved asking the robot clarification questions about its perception or behavior, asking the robot to take specific corrective actions, or asking the robot for clarification on its capabilities so that they can prevent the error in the future.

VI. DISCUSSION AND FUTURE WORK

Through our online study, we evaluated Templated and Generative systems for providing explanations of robot failures. Contrary to our hypotheses, we observed surprising results that demonstrate the limitations of using LLMs in this context.

We found that Generative explanations were less understandable compared to Templated explanations, thus rejecting H2a. We did not anticipate a difference between System Types as they both generated explanations which contained the same level of information and were both factually correct, as determined by the experimenters. Notably, while the Generative System's explanations used different language with the same conceptual backing, there were drastic differences in understanding. Participants also reported that the Templated explanations helped them understand the failure similarly or more than the Generative explanations. Despite these differences in understanding, we did not observe a significant difference in participants' confidence to their responses for the understanding question.

Our study has shown that it is important to be careful when considering the usage of LLMs to generate explanations for robot failures in HRI. There are several important trade-offs and considerations that should be investigated in future work to enable researchers to make appropriate and informed decisions on whether or not to use LLMs, described below.

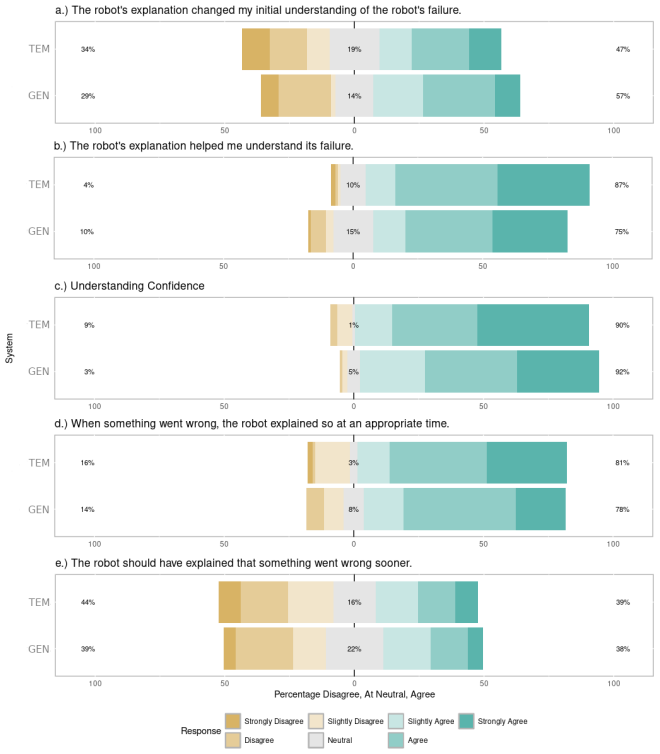

Fig. 3. The distribution of responses to the influence on understanding (a,b) understanding confidence (c), and explanation timing (d,e). Percentages indicate the percent of participants who responded below neutral (left), neutral (center), and above neutral (right).

A. Consistency

In our work, the Generative System leveraged a prompt template [20] which provided the LLM with context of the scenario and the same information that the Templated System had. For our Generative System, we ran our prompts once per Failure Type and selected the first response generated. This is a potential limitation of our work, as responses from LLMs vary, meaning that explanations generated in the real world could take multiple forms at different times. On the other hand, templated systems generate very consistent and repeatable explanations.

Though our results suggest that generative explanations are perceived as similarly or less intelligent than templated systems, variability in the generated explanations may increase perceived intelligence since the robot could explain the same thing in different ways. This trade-off between consistent and varying explanations should be explored in future work.

B. Generalizability

Templated Systems utilize hand-crafted templates, and therefore its explanations are only as generalizable as the templates themselves. For example, the wording of a template might be suitable for explaining failures, but not for answering questions. LLMs on the other hand are more capable of generalizing across tasks. In domains where it is infeasible to have experts craft templates for every necessary scenario, it may be necessary to leverage generative systems, which have a greater capacity for generalization.

C. The Human Factor

While the trade-offs between accuracy and fluency in generated text is important [37], our results suggest that there may be more to this trade-off in a HRI context. Experimenters validated our generated explanations to ensure they were accurate and contained the same level of information. By ensuring the same accuracy, we proposed hypotheses regarding the grammatical correctness, or fluency, of the explanations based on our previous observations where participants had negative perceptions of a templated system due to its incorrect grammar [12]. To our surprise the accurate and grammatically correct explanations from the Generative System received ratings on par or lower than the accurate but grammatically incorrect explanations from the Templated System. We believe this is due to the different phrasing between the Templated and Generative systems.

It is not surprising that LLMs generated explanations with different phrasing than the experimenter crafted templates that were leveraged by the Templated System, however, we observed that sometimes the language used by the LLM seemed more technical. This can be seen in the Caddy Out of Reach Failure Type (Table I), where the system mentions generating a path to the caddy and positioning a screw for placement. This terminology is likely not how a person would describe the robot's failure to another person – especially to a robot novice. Robot explanations should take into account the recipients' role(s) and experience [13], therefore, it is especially important to ensure that generative systems do so as well. Future work should explore different prompt structures to prompt LLMs with the recipients' role(s) and experience with robot systems in general as well as the particular system.

ACKNOWLEDGMENTS

Section IV, particularly Subsections IV-B and IV-D, contain similar content to the corresponding sections in our prior work [12] as we used the same underlying architecture and experimental design in order to have comparable results.

Thank you to Elizabeth Phillips for her suggestions and clarification on the statistical analyses.

REFERENCES

[1] L. Wachowiak, A. Fenn, H. Kamran, A. Coles, O. Celiktutan, and G. Canal, "When do people want an explanation from a robot?" in *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction (Boulder, CO, USA)(HRI'24)*. ACM, 2024.

[2] M. T. Dzindolet, S. A. Peterson, R. A. Pomranky, L. G. Pierce, and H. P. Beck, "The role of trust in automation reliance," *International Journal of Human-Computer Studies*, vol. 58, no. 6, pp. 697–718, 2003.

[3] N. Wang, D. V. Pynadath, and S. G. Hill, "Trust calibration within a human-robot team: Comparing automatically generated explanations," in *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2016, pp. 109–116.

[4] M. Kwon, S. H. Huang, and A. D. Dragan, "Expressing robot incapability," in *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, 2018, pp. 87–95.

[5] N. Wang, D. V. Pynadath, E. Rovira, M. J. Barnes, and S. G. Hill, "Is it my looks? Or something I said? The impact of explanations, embodiment, and expectations on trust and performance in human-robot teams," in *International Conference on Persuasive Technology*. Springer, 2018, pp. 56–69.

[6] M. Edmonds, F. Gao, H. Liu, X. Xie, S. Qi, B. Rothrock, Y. Zhu, Y. N. Wu, H. Lu, and S.-C. Zhu, "A tale of two explanations: Enhancing human trust by explaining robot behavior," *Science Robotics*, vol. 4, no. 37, 2019.

[7] R. Thielstrom, A. Roque, M. Chita-Tegmark, and M. Scheutz, "Generating explanations of action failures in a cognitive robotic architecture," in *2nd Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence*, 2020, pp. 67–72.

[8] D. Das, S. Banerjee, and S. Chernova, "Explainable AI for robot failures: Generating explanations that improve user assistance in fault recovery," in *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, 2021, pp. 351–360.

[9] S. Stange and S. Kopp, "Effects of a social robot's self-explanations on how humans understand and evaluate its behavior," in *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, 2020, pp. 619–627.

[10] A. Alvanpour, S. K. Das, C. K. Robinson, O. Nasraoui, and D. Popa, "Robot failure mode prediction with explainable machine learning," in *2020 IEEE 16th International Conference on Automation Science and Engineering (CASE)*. IEEE, 2020, pp. 61–66.

[11] M. Diehl and K. Ramirez-Amaro, "A causal-based approach to explain, predict and prevent failures in robotic tasks," *Robotics and Autonomous Systems*, vol. 162, p. 104376, 2023.

[12] G. LeMasurier, A. Gautam, Z. Han, J. W. Crandall, and H. A. Yanco, "Reactive or proactive? How robots should explain failures," in *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, 2024, pp. 413–422.

[13] M. Ribera and À. Lapedriza García, "Can we do better explanations? A proposal of user-centered explainable AI." CEUR Workshop Proceedings, 2019.

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

[15] Z. Han, E. Phillips, and H. A. Yanco, "The need for verbal robot explanations and how people would like a robot to explain itself," *ACM Transactions on Human-Robot Interaction (THRI)*, vol. 10, no. 4, pp. 1–42, 2021.

[16] G. Vilone and L. Longo, "Classification of explainable artificial intelligence methods through their output formats," *Machine Learning and Knowledge Extraction*, vol. 3, no. 3, pp. 615–661, 2021.

[17] Z. Han and H. Yanco, "Communicating missing causal information to explain a robot's past behavior," *ACM Transactions on Human-Robot Interaction*, vol. 12, no. 1, pp. 1–45, 2023.

[18] E. Cambria, L. Malandri, F. Mercorio, M. Mezzanzanica, and N. Nobani, "A survey on XAI and natural language explanations," *Information Processing & Management*, vol. 60, no. 1, p. 103111, 2023.

[19] M. Lomas, R. Chevalier, E. V. Cross, R. C. Garrett, J. Hoare, and M. Kopack, "Explaining robot actions," in *Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction (HRI '12)*, 2012, pp. 187–188.

[20] C. Tagliamonte, D. Maccaline, G. LeMasurier, and H. A. Yanco, "A generalizable architecture for explaining robot failures using behavior trees and large language models," in *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, 2024, pp. 1038–1042.

[21] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, "Large language models are zero-shot reasoners," *Advances in neural information processing systems*, vol. 35, pp. 22 199–22 213, 2022.

[22] F. Zeng, W. Gan, Y. Wang, N. Liu, and P. S. Yu, "Large language models for robotics: A survey," *arXiv preprint arXiv:2311.07226*, 2023.

[23] C. Zhang, J. Chen, J. Li, Y. Peng, and Z. Mao, "Large language models for human-robot interaction: A review," *Biomimetic Intelligence and Robotics*, p. 100131, 2023.

[24] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," *Advances in Neural Information Processing Systems*, vol. 33, pp. 1877–1901, 2020.

[25] J. Sevilla-Salcedo, E. Fernádez-Rodicio, L. Martín-Galván, Á. Castro-González, J. C. Castillo, and M. A. Salichs, "Using large language models to shape social robots' speech," 2023.

[26] Y.-J. Wang, B. Zhang, J. Chen, and K. Sreenath, "Prompt a robot to walk with large language models," *arXiv preprint arXiv:2309.09969*, 2023.

[27] I. Singh, V. Blukis, A. Mousavian, A. Goyal, D. Xu, J. Tremblay, D. Fox, J. Thomason, and A. Garg, "ProgPrompt: program generation for situated robot task planning using large language models," *Autonomous Robots*, pp. 1–14, 2023.

[28] Y. Cao and C. Lee, "Robot behavior-tree-based task generation with large language models," *arXiv preprint arXiv:2302.12927*, 2023.

[29] M. Á. González-Santamarta, L. Fernández-Becerra, D. Sobrín-Hidalgo, Á. M. Guerrero-Higueras, I. González, and F. J. R. Lera, "Using large language models for interpreting autonomous robots behaviors," in *International Conference on Hybrid Artificial Intelligence Systems*. Springer, 2023, pp. 533–544.

[30] K. Mahadevan, J. Chien, N. Brown, Z. Xu, C. Parada, F. Xia, A. Zeng, L. Takayama, and D. Sadigh, "Generative expressive robot behaviors using large language models," *arXiv preprint arXiv:2401.14673*, 2024.

[31] 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.

[32] N. Kroeger, D. Ley, S. Krishna, C. Agarwal, and H. Lakkaraju, "Are large language models post hoc explainers?" *arXiv preprint arXiv:2310.05797*, 2023.

[33] P. Mavrepis, G. Makridis, G. Fatouros, V. Koukos, M. M. Separdani, and D. Kyriazis, "XAI for all: Can large language models simplify explainable ai?" *arXiv preprint arXiv:2401.13110*, 2024.

[34] T. Williams, C. Matuszek, R. Mead, and N. Depalma, "Scarecrows in oz: The use of large language models in hri," pp. 1–11, 2024.

[35] A. Gatt and E. Krahmer, "Survey of the state of the art in natural language generation: Core tasks, applications and evaluation," *Journal of Artificial Intelligence Research*, vol. 61, pp. 65–170, 2018.

[36] "Nlg vs. templates." Citeseer.

[37] E. Reiter. (2019) Generated texts must be accurate! [Online]. Available: https://ehudreiter.com/2019/09/26/generated-texts-must-be-accurate/

[38] M. Colledanchise and P. Ögren, *Behavior Trees in Robotics and AI: An Introduction*. CRC Press, 2018.

[39] Z. Han, D. Giger, J. Allspaw, M. S. Lee, H. Admoni, and H. A. Yanco, "Building the foundation of robot explanation generation using behavior trees," *ACM Transactions on Human-Robot Interaction (THRI)*, vol. 10, no. 3, pp. 1–31, 2021.

[40] A. Gautam, T. Whiting, X. Cao, M. A. Goodrich, and J. W. Crandall, "A method for designing autonomous agents that know their limits," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2022.

[41] X. Cao, A. Gautam, T. Whiting, S. Smith, M. A. Goodrich, and J. W. Crandall, "Robot proficiency self-assessment using assumption-alignment tracking," *IEEE Transactions on Robotics*, vol. 39, no. 4, pp. 3279–3298, 2023.

[42] S. Zudaire, F. Gorostiaga, C. Sánchez, G. Schneider, and S. Uchitel, "Assumption monitoring using runtime verification for UAV temporal task plan executions," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 6824–6830.

[43] F. Faul, E. Erdfelder, A. Buchner, and A.-G. Lang, "Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses," *Behavior Research Methods*, vol. 41, no. 4, pp. 1149–1160, 2009.

[44] "Fetchit! A mobile manipulation challenge," https://opensource.fetchrobotics.com/competition, 2019, accessed: 2022-02-04.

[45] M. Wise, M. Ferguson, D. King, E. Diehr, and D. Dymesich, "Fetch & Freight: Standard platforms for service robot applications," in *Workshop on Autonomous Mobile Service Robots*, 2016.

[46] B. M. Muir, "Operators' trust in and use of automatic controllers in a supervisory process control task," 1989.

[47] D. Ullman and B. F. Malle, "MDMT: Multi-dimensional measure of trust," 2019.