Enhancing Situation Awareness through Model-Based Explanation Generation

Konstantinos Gavriilidis[1], Ioannis Konstas[1], Helen Hastie[2], Andrea Munafo[3], Wei Pang[1]
[1]School of Mathematical and Computer Sciences, Heriot-Watt University, Edinburgh, UK
[2]School of Informatics, University of Edinburgh, Edinburgh, UK
[3]SeeByte Limited, Edinburgh, UK
[1] {kg47, I.Konstas, W.Pang}@hw.ac.uk
[2] h.hastie@ed.ac.uk [3] andrea.munafo@seebyte.com

Abstract

Robots are often deployed in remote locations for tasks such as exploration, where users cannot directly perceive the agent and its environment. For Human-In-The-Loop applications, operators must have a comprehensive understanding of the robot's current state and its environment to take necessary actions and effectively assist the agent. In this work, we compare different explanation styles to determine the most effective way to convey real-time updates to users. Additionally, we formulate these explanation styles as separate fine-tuning tasks and assess the effectiveness of large language models in delivering in-mission updates to maintain situation awareness. The code and dataset for this work are available at: https://github.com/konsgavriil/explainable_robotics_lm.

1 Introduction

Automation offers significant advantages in our society, particularly in critical sectors like manufacturing and offshore applications, as recognized in prior studies (Ballestar et al., 2021; Khalid et al., 2022). Fostering transparency and accountability within robotics is imperative to bolster trust and wider adoption (Wachter et al., 2017; Winfield et al., 2021). One pivotal cognitive process influencing trust and adoption is situational awareness, characterized by three essential stages: **perception** (understanding a robot's decision-making), **comprehension** (discerning the rationale behind these decisions), and **projection** (anticipating future automated behaviours). Recent research has shown that textual explanations presented visually within Human-In-The-Loop applications, such as autonomous driving, positively impact all facets of situational awareness (Avetisyan et al., 2022).

In this work, we include two user studies focusing on situation awareness and explanation generation. We share a dataset, licensed under Creative Commons Attribution (CC-BY), which contains categorical events related to maritime autonomous missions, user queries, and corresponding explanations. Additionally, we demonstrate the performance of multiple large language models on three downstream tasks derived from our dataset.

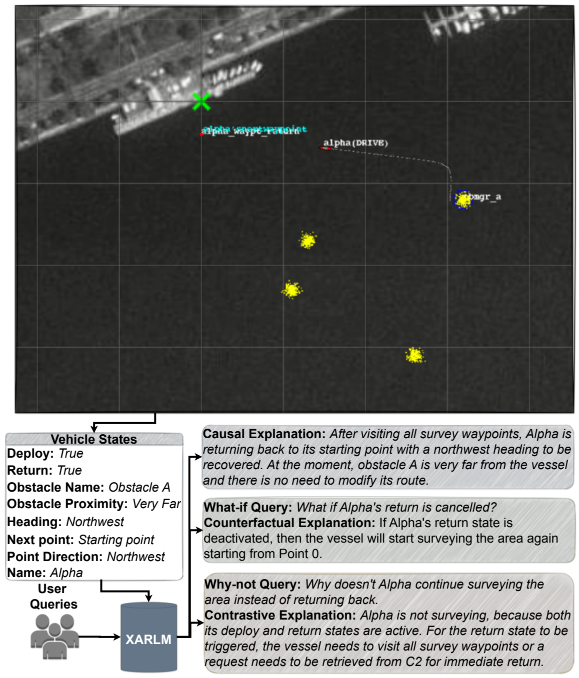

Figure 1: The eXplainable Autonomous Robot Language Model (XARLM) retrieves vehicle states and user queries to generate explanations in various styles, thereby enhancing situation awareness for vehicle operators.

Through the fine-tuning process and the user studies, we aim to answer the following research questions:

- **RQ1:** How robust are large language models in delivering explanations of autonomous mission events in causal, counterfactual, and contrastive styles?
- **RQ2:** Which of the three explanation styles

Proceedings of the 2nd Workshop on Practical LLM-assisted Data-to-Text Generation, pages 7–16

most effectively enhances situation awareness for users?

- **RQ3:** Do users prefer model-based explanations over template-based explanations?

The remainder of this paper is structured as follows: Section 2 reviews prior research that has influenced our approach. Section 3 describes the data collection and annotation processes. Section 4 outlines the fine-tuning process for the large language models and details the experiments conducted to identify the best-performing model. In Section 5, we describe the tasks included in our study to address research questions 2 and 3, as well as the participant groups that completed the questionnaire. Section 6 presents the performance of the large language models and our findings from the user studies. Finally, Section 7 examines the implications of our findings, and Section 8 discusses potential future experiments and concludes the paper.

2 Related Work

Explainable agents and robots have become a crucial research area due to the increasing demand for transparency and interpretability in autonomous systems (Langley et al., 2017; Anjomshoae et al., 2019). These systems must effectively communicate their decision-making processes to users, preferably through user-friendly modalities such as natural language (Cambria et al., 2023). Typically, natural language explanations are presented as causal explanations, which are easy to understand and clearly justify automated behaviours (Diehl and Ramirez-Amaro, 2022). Other types of explanations, such as counterfactual explanations (answering "What if" questions) and contrastive explanations (answering "Why not" questions), also facilitate the interrogation of black-box systems (Stepin et al., 2021).

Generating these explanations faithfully involves sophisticated methods for content selection, such as using Bayesian networks or surrogate models (Gyevnar et al., 2022; Gavriilidis et al., 2023). The selected content can then be communicated through controllable template-based approaches (Hastie et al., 2017). Additionally, end-to-end approaches using encoder-decoder architectures have shown promise in conveying agent rationale and improving failure and solution identification (Ehsan et al., 2019; Das et al., 2021).

The advent of causal language models with transformer-based encoder architectures (Touvron et al., 2023; Jiang et al., 2023) has significantly advanced the field of text generation. These models excel at replicating domain-specific knowledge due to extensive training on vast amounts of human-generated text (Kıcıman et al., 2023). Despite their substantial size and complexity, new techniques such as QLoRA (Dettmers et al., 2024) have made fine-tuning more computationally efficient, facilitating the adaptation of pre-trained models to specific downstream tasks.

To evaluate the semantic accuracy of models, researchers frequently compare the outputs with their corresponding inputs (Xu et al., 2021). This evaluation approach is particularly important in applications where the accuracy and reliability of natural language explanations are critical. Commonly used metrics for this purpose include BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and METEOR (Banerjee and Lavie, 2005), which measure quality based on n-gram overlap between reference labels and model-generated responses. However, these metrics have limitations, as they often fail to capture the true semantic similarity or desired verbosity of the outputs (Zhao et al., 2020). To address these shortcomings, combining n-gram-based metrics with additional metrics that perform verbatim comparisons can provide a more comprehensive evaluation of model outputs, particularly in high-stakes applications.

Given the robustness of large language models in data-to-text generation and their capability to perform multiple tasks, various domains have leveraged these models for diverse applications. For instance, they have been used for action selection in embodied tasks (Ahn et al., 2022) and for text summarization to infer sets of rules for object manipulation based on user preferences (Wu et al., 2023). In the realm of explainable robotics, language models are combined with Retrieval Augmented Generation (Lewis et al., 2020) to transform robot logs and user queries into natural language explanations, thereby enhancing human-robot interaction.

3 Data Generation

To collect a dataset for autonomous maritime vehicles, we deployed an agent that follows a pre-existing plan and attempts to complete its objectives by visiting a set of waypoints using different patterns (e.g., lawnmower, loiter). The agent prioritizes the integrity of the robotic platform and replans its behaviour in case of unexpected events. At

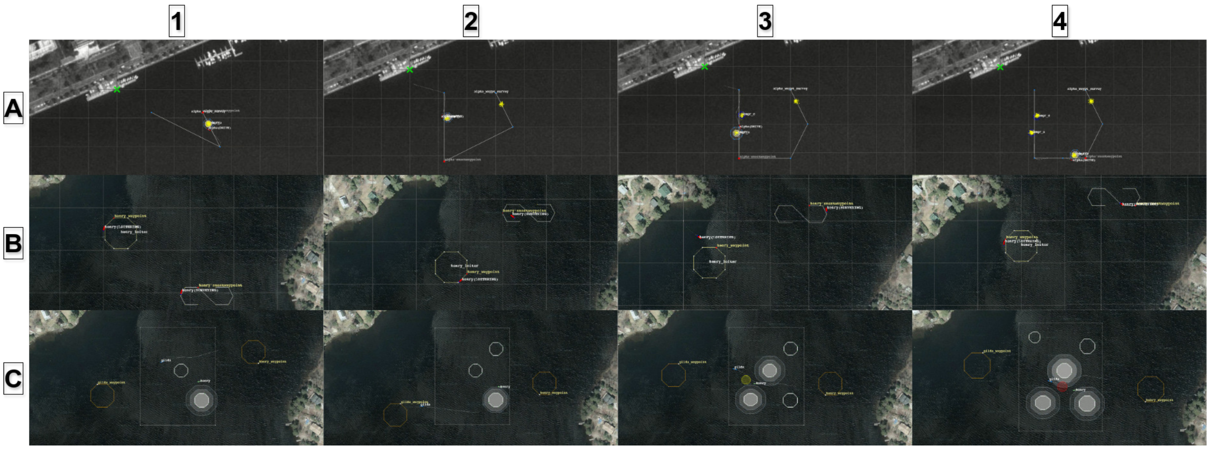

Figure 2: The MOOS-IvP scenarios used for data generation. Each of the three scenarios includes four different configurations with varying waypoints and objectives.

each simulation timestep, we recorded the robot's behaviour and the states affecting that behaviour, such as objectives and sensor-derived events (e.g., obstacle or vessel detection).

For this dataset, we utilised MOOS-IvP, an open-source behavioural agent designed for maritime robots (Benjamin et al., 2010). This simulator offers a variety of pre-built scenarios, from which we selected and refined three specific missions. We further modified the mission plans, creating four distinct configurations for each scenario. We limited the logged vehicle states to those impacting the agent's behaviour activations. Finally, we performed post-mission parsing of the log files to extract the relevant vehicle states and the corresponding activated behaviours.

In Figure 2, we illustrate three scenarios, each with four distinct task and environment configurations. Scenario **A** involves an unmanned surface vehicle (USV) conducting a survey, avoiding obstacles, and returning to its starting point for retrieval. Scenario **B** features an unmanned underwater vehicle (AUV) loitering around predefined waypoints and transitioning to a designated survey area upon receiving instructions. Scenario **C** has two vehicles loitering around a random polygon, occasionally switching sides and restarting their routine while avoiding collisions and obstacles. With each scenario, the task difficulty increases by adding more behaviours and introducing complex tasks such as collision avoidance.

3.1 Data Annotation

After completing data collection, model-based data annotation was conducted. Using a larger model,

	C	**CF**	**CT**
Dataset size	1151	3450	3450
Vocabulary size	758	993	1167
Avg Input Length	109.70	121.90	125.22
Longest Input Length	132	153	165
Shortest Input Length	86	96	97
Avg Output Length	42.38	37.13	61.41
Longest Output Length	89	122	151
Shortest Output Length	16	7	21
Inputs with spatial tokens	1151	3450	3450
Avg spatial tokens/input	18.88	22.94	22.06
Outputs with spatial tokens	1146	3128	3379
Avg spatial tokens/output	8.58	6.23	7.83

Table 1: Dataset Statistics for causal (C), counterfactual (CF) and contrastive (CT) explanations.

new annotations were generated for each data instance by providing a small number of instruction-based examples, as guided by prior research (Taori et al., 2023). Specifically, the OpenAI API's ChatCompletion functionality with the GPT-3.5-Turbo model was utilised. Task instructions and concatenated vehicle state representations were input, resulting in potential user queries along with their corresponding explanations. For counterfactual and contrastive explanations, a state or behaviour permutation was also provided, depending on the task, to validate the user query upon which the explanation was based.

Initially, 12 instructions were defined for Scenario A, 15 for Scenario B, and 21 for Scenario C, ensuring that all unique states and behaviours relevant to each scenario were addressed. This process produced an annotated dataset comprising 8,051 data instances, reflecting the state updates encountered during each mission to minimise repeated state-behaviour combinations. Detailed statistics

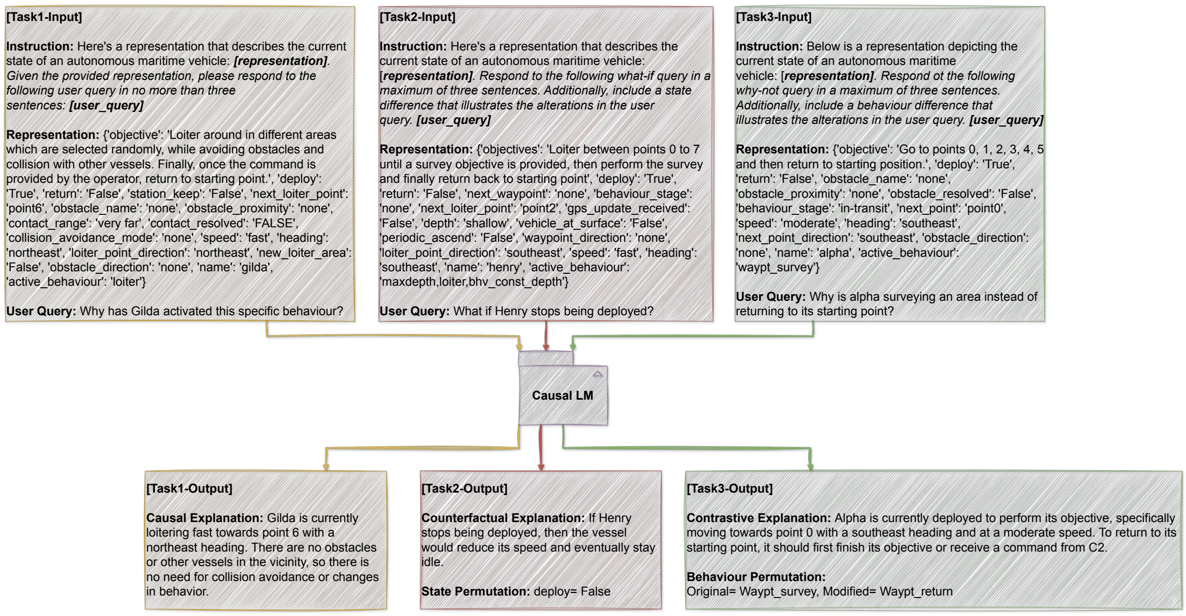

Figure 3: The defined fine-tuning tasks involve a causal language model that retrieves an instruction, a vehicle state representation, and a user query. The model then outputs an explanation, and for counterfactual and contrastive explanations, it additionally provides a permutation.

of the annotated dataset, including vocabulary size, input/output lengths, and the number of spatial tokens, are provided in Table 1.

4 Fine-Tuning

Before attempting any fine-tuning, we evaluated the performance of existing instruction fine-tuned large language models to assess their capability in generating explanations from autonomous vehicle states. Specifically, we employed three transformer-based decoder models, each with an identical number of parameters: Llama2-7B-Chat, Mistral-7B-IT, and Falcon-7B-IT, using 2-shot inference (Touvron et al., 2023; Jiang et al., 2023; Almazrouei et al., 2023). This preliminary experiment revealed that all three models demonstrated strong results in terms of semantic accuracy and precision, with Mistral and Llama2 slightly outperforming Falcon in these aspects. However, when evaluated using machine translation metrics, all three models exhibited significant shortcomings, with Mistral performing slightly better.

Upon further inspection of the model outputs, we found that these models often compensated by increasing verbosity and adding supplementary tokens. These additions were unnecessary and may increase the cognitive load for users reading the explanations. The higher semantic accuracy and precision scores can be attributed to our metric's focus on the output mentioning spatial elements, behaviours, and entities present in the input representation. In contrast, the lower scores in ROUGE-L, BLEU, and METEOR metrics were due to the generated outputs not closely resembling the dataset labels. The results of this initial experiment are illustrated in Figure 4.

Recognizing the need for further refinement for

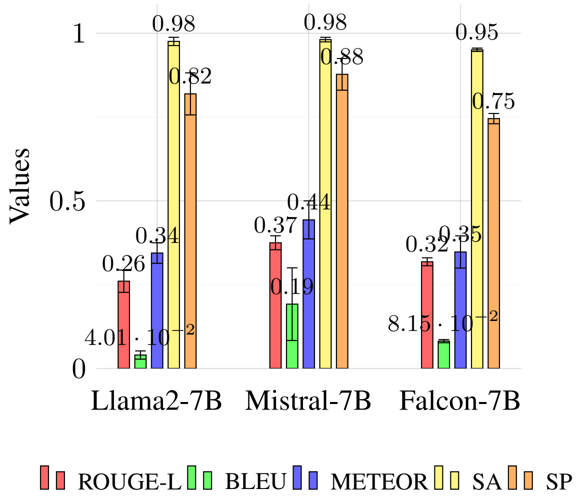

Figure 4: Performance of instruction fine-tuned large language models on two-shot inference tasks, with error bars indicating the mean and variability across various explanation types. The metrics include Semantic Accuracy (SA) and Semantic Precision (SP).

our downstream task, we defined a fine-tuning setup where each explanation type is treated as a separate task. In Figure 3, we represent our fine-tuning tasks, where a task instruction, along with a representation and a user query, are provided as input. The model output is the corresponding explanation, including permutations for counterfactual and contrastive explanations. Using our annotated dataset, we trained the three large language models on all explanation tasks utilizing the HuggingFace and PEFT (Xu et al., 2023) libraries.

4.1 Automatic Evaluation

To evaluate the performance of our models on downstream tasks, three machine translation metrics—BLEU, METEOR, and ROUGE—were utilised to measure n-gram overlap between the model outputs and reference labels. Additionally, to accurately assess the mentions of entities, landmarks, and specific details such as vessel heading, depth, speed, or behaviour, a semantic accuracy and precision metric was developed. The SA metric increases with each correct mention and decreases when elements are inaccurately identified (e.g., using 'medium' instead of 'fast' speed), ensuring the fidelity of the generated explanations.

Specifically, given a set of input tokens I and a set of output tokens O, for each token category (spatial, state, decision) that is based on a vocabulary we predefined, the sets of correct references, true positives, and false positives are defined as follows:

$$\text{Correct References} = I \cap O$$

The number of true positives (TP) and false positives (FP) are calculated as:

$$TP = |\text{Correct References}|$$

$$FP = \text{Total References} - TP$$

The semantic accuracy (Acc) and precision (Prec) are defined by:

$$\text{Acc} = \frac{TP + TN}{|O|}$$

$$\text{Prec} = \frac{TP}{TP + FP}$$

where TN (true negatives) denotes the number of tokens in O that are not references. The overall semantic accuracy and precision are computed as the average across all evaluated references within the spatial, state, and decision categories.

Section 6 presents a performance comparison of the three language models to identify the best-performing model. An ablation study was subsequently conducted on the top-performing model to explore potential improvements.

5 User Study

To estimate the effect of explanations on users and determine user preference for model-based explanations, we designed two user studies. A total of 21 participants were recruited from the robotics industry and academia, including 9 individuals very familiar with autonomous vehicles, 9 who were familiar, and 3 who were not familiar.

User Study on Situation Awareness: This study builds upon prior work (Robb et al., 2018) to investigate the effect of different explanation styles on users' situation awareness. We used recorded videos from the aforementioned maritime robot simulator, where an agent attempts to accomplish a set of objectives while considering its environment and inner state, particularly during unexpected events that require replanning. Participants encountered three different conditions, each with a different explanation type (causal, counterfactual, and contrastive), along with a tutorial video describing the task beforehand. In the first condition, explanations were presented with captions. In the second and third conditions, participants selected user queries to generate corresponding explanations that clarified alternative outcomes. After the explanations were displayed, the interface asked users about events taking place in the video at predefined timesteps. Their responses were used to estimate a performance metric representing their situation awareness per condition, thus assessing the effect of each explanation style on their mental models.

User Study on Explanation Preference: This study presented three separate scenarios, each with a map displaying the vessel and its environment, a description summarizing the events, a user query, and three potential explanations. Two default options allowed users to select all or none of the explanations to avoid restricting their choices. For each scenario, participants chose the explanation that best conveyed the current state of the robot. These explanations were derived from both domain expert templates (with low soundness and high complete-

	ROUGE-L	BLEU	METEOR	Semantic Accuracy	Semantic Precision
Causal	0.631	0.460	0.651	0.978	0.884
Counterfactual	0.665	0.538	0.670	0.969	0.857
Contrastive	0.652	0.561	0.669	0.983	0.902
All types	0.430	0.417	0.459	0.975	0.887

Table 2: Performance comparison of the top-performing large language model, Mistral, on individual tasks as well as on a combined dataset of all three tasks using a balanced dataset.

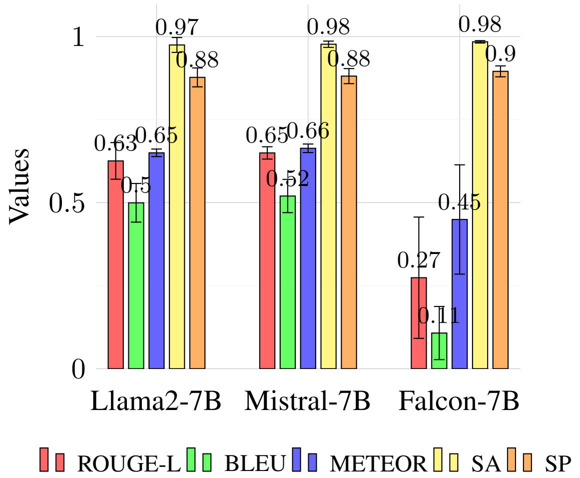

Figure 5: Performance of fine-tuned large language models, showing improved machine translation metrics compared to Figure 4.

ness) and language models, though participants were not informed of their origin. Selections were made based solely on the clarity and informativeness of the explanations provided.

6 Results

In this section, we present the results of our fine-tuned models and user study, addressing the research questions outlined in Section 1.

6.1 Automatic Evaluation

To address **RQ1**, we present the overall performance of the three large language models on the three downstream tasks, as illustrated in Figure 5. Based on the performance metrics, Mistral and Llama2 demonstrated the best results, with Mistral showing a slight edge and a significant improvement in machine translation metrics. These models also achieved high scores in Semantic Accuracy and Precision, indicating that their outputs accurately reflected the vehicle state representations provided as input.

In contrast, the Falcon model performed well on causal explanations but did not achieve comparable performance on the other two explanation types, affecting its mean performance across all tasks. Similar to its behaviour in the instruction version, the Falcon model produced verbose outputs that mixed relevant tokens with supplementary, unnecessary information. These results were evaluated for both the fine-tuned and instruction models using a test set of 100 data instances for each explanation type.

After identifying Mistral-7B as our best-performing model, we evaluated its performance on three individual datasets and a balanced dataset with equal numbers of all explanation types. As shown in Table 2, the model trained on the counterfactual dataset achieved the highest ROUGE-L and METEOR scores. The model trained on the contrastive dataset achieved the best BLEU score. The causal dataset model ranked third in machine translation metrics, with the balanced dataset model coming in last. For semantic accuracy and precision, the contrastive dataset model performed the best, while the causal and balanced dataset models had similar results. The counterfactual dataset model ranked last in semantic accuracy and precision, but not significantly behind the top models.

6.2 User Study

With the results from the two user studies, we address **RQ2** and **RQ3** as outlined in Section 1.

In the first user study, illustrated in Figure 6, we measured the total number of correct answers per condition (causal, counterfactual, and contrastive) and compared these results to the probability of randomly selecting the correct answer (33.3%) to determine the impact of explanations on situation awareness. Causal explanations led to the highest percentage of correct answers (76.19%), followed by contrastive explanations (69.84%) and counterfactual explanations (59.67%). The performance difference between random selection and explanation-assisted answers demonstrates that our explanations enhanced users' ability to correctly perceive events.

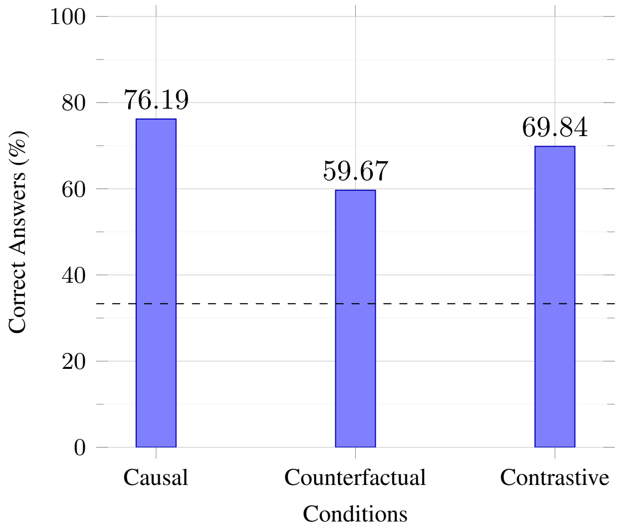

Figure 6: Percentage of correct answers for each condition in the first user study examining the impact of explanation styles on situation awareness.

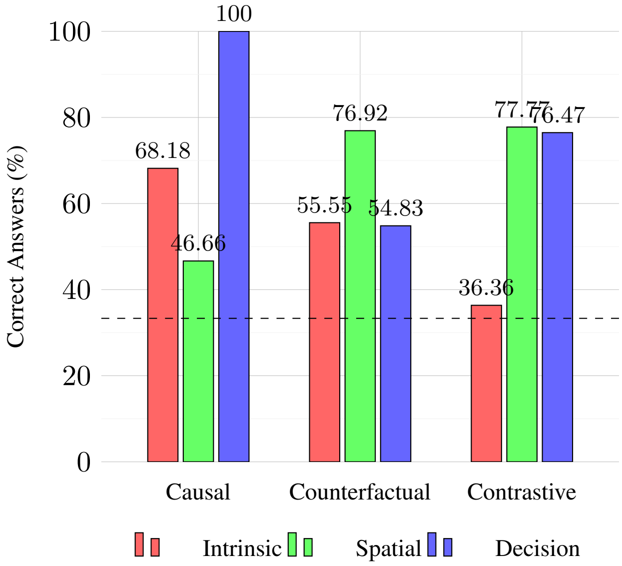

Figure 7: Percentage of correct answers for each condition on questions assessing different aspects of autonomous vehicles (intrinsic states, spatial elements, decision-making).

Counterfactual Query: What if the vehicle is at the surface and has received a GPS update?

Template-based Explanation (HiCompLowSound): The vessel has pinpointed its position and is now ready to continue its mission.

Template-based Explanation (HiCompMediumSound): If Henry is at the surface and has received a GPS update, then it should be able to move towards the loiter area and continue its mission.

Model-based Explanation: If the vehicle is at the surface and has received a GPS update, it means that the vessel has successfully surfaced to provide new GPS coordinates to the command and control. This indicates that the vessel is ready to continue its mission.

Figure 8: Three correct explanations for a counterfactual query, consisting of two template-based explanations with high completeness—one with low soundness and the other with medium soundness.

Further analysis of the first user study involved categorizing the questions into three types: **intrinsic** (inquiring about the robot's internal states, such as sensor readings), **spatial** (concerning the vessel's topology, its environment, and nearby entities or landmarks), and **decision-making** (asking about the rationale behind the robot's decisions). Figure 7 shows that causal explanations resulted in the highest accuracy for intrinsic (68.18%) and decision-making (100%) questions, but the lowest for spatial questions (46.66%), still better than random selection. Counterfactual explanations provided the second-best performance for both intrinsic and spatial questions, showing at least a 20% improvement over random selection. Contrastive explanations led to the best performance for spatial questions (77.77%) and the second-best for decision-making (76.47%), but they performed the worst for intrinsic questions, only slightly better than random selection (36.36%).

In the second study, we explored user preferences between template-based and model-based explanations. Templates created by domain experts, containing only essential information with optimal verbosity, were preferred by 70% of users. Model-based explanations were favored by 15%, while 13.33% liked both types equally, and 1.66% liked none of the explanations. These results suggest that although model-based annotations can accurately depict events, they do not fully match the preferred explanation style of users. This discrepancy indicates that the initial annotation instructions might need refinement to train models that produce explanations more closely aligned with those created by domain experts. Figure 8 presents an example of a what-if query with two template-based explanations and one model-based explanation.

7 Discussion

Our evaluation of inference performance using existing instruction fine-tuned large language models revealed that, despite their inherent capabilities and domain knowledge from pre-training, these models fall short in generating explanations with the verbosity and style expected by domain experts in autonomous vehicles. Consequently, additional fine-tuning on specific downstream tasks is necessary. Our fine-tuned models showed significant improvements in machine translation metrics, indicating a strong n-gram overlap between predictions and reference labels. Notably, our best model performed exceptionally well on counterfactual and contrastive explanations, followed by causal explanations and mixed styles when using a balanced dataset. Furthermore, the generated outputs exhibited high semantic accuracy and precision, underscoring the effectiveness of the fine-tuning process.

The results from the first user study on the effect of different explanation styles on situation awareness demonstrated that users significantly benefited from our explanations compared to random chance, as there were three potential answers per question. Specifically, users gave the most correct answers using causal explanations, followed by contrastive and counterfactual explanations. For causal explanations, users excelled in answering questions about decision-making, as the justification behind the exhibited behaviour was clear and did not require further queries.

Conversely, counterfactual and contrastive explanations allowed users to interrogate the system and learn more about the spatial elements of the mission, resulting in an almost equal percentage of correct answers. While causal explanations helped users answer spatial questions with the third-best success rate, they did not provide enough time to digest the information, potentially increasing cognitive load.

For intrinsic questions concerning the robot's inner states, such as sensor readings, causal explanations demonstrated the best performance, indicating that a straightforward approach to explaining a robot's inner states is the most effective strategy. Considering these findings, future work could tailor the explanation styles presented to users based on the type of content needing explanation.

The results from the second user study indicated a clear preference among participants for domain expert template-based explanations, though some participants preferred model-based explanations, and others expressed no strong preference, showing equal satisfaction with both types. This preference may have been influenced by the presentation format: each scenario featured two template-based explanations characterised by high completeness (the breadth of justifications behind an outcome) and low to medium soundness (the level of detail for each justification), which directly reflected the vessel's current state (Kulesza et al., 2013). In contrast, only one model-based explanation was provided per scenario. The use of model-based data annotation for labelling the dataset may have also impacted the study's outcomes.

Future work should focus on aligning language model outputs more closely with the response styles of domain experts and further refining model-based data annotation techniques, particularly for critical applications. Template-based explanations, while effective, are not scalable, require significant time to develop, and lack robustness, especially when dealing with complex or evolving scenarios. These limitations highlight the need for a data-driven approach using large language models, which offer greater adaptability, efficiency, and the potential to generate contextually relevant explanations at scale.

8 Conclusion and Future Work

This work has successfully demonstrated the impact of different explanation styles on situational awareness across various aspects of a mission, such as decision-making, spatial elements, and inner vehicle states, within the context of human-in-the-loop applications for autonomous vehicles. Additionally, we assessed user preferences between template-based and model-based explanations.

We also showcased the capabilities of our large language model in performing data-to-text tasks, transforming the states of autonomous vehicles into natural language explanations across three different styles. The fine-tuned models have shown satisfactory performance in generating coherent and contextually appropriate explanations.

For future work, several avenues for enhancement and exploration remain open. Experimenting with a broader range of explanation types could provide deeper insights into user preferences and effectiveness. Additionally, integrating additional modalities, such as map or chart-based user interfaces, would be a valuable extension. These interfaces are commonly used in conjunction with autonomous agents and could offer a more comprehensive and interactive explanatory experience for users.

Acknowledgments

This work was funded by the EPSRC CDT in Robotics and Autonomous Systems, SeeByte Ltd, and SRPe. We also extend our gratitude to Dr. Marta Romeo for her support.

References

Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, et al. 2022. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, et al. 2023. The falcon series of open language models. *arXiv preprint arXiv:2311.16867*.

Sule Anjomshoae, Amro Najjar, Davide Calvaresi, and Kary Främling. 2019. Explainable agents and robots: Results from a systematic literature review. In *18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), Montreal, Canada, May 13–17, 2019*, pages 1078–1088. International Foundation for Autonomous Agents and Multiagent Systems.

Lilit Avetisyan, Jackie Ayoub, and Feng Zhou. 2022. Investigating explanations in conditional and highly automated driving: The effects of situation awareness and modality. *Transportation research part F: traffic psychology and behaviour*, 89:456–466.

María Teresa Ballestar, Ángel Díaz-Chao, Jorge Sainz, and Joan Torrent-Sellens. 2021. Impact of robotics on manufacturing: A longitudinal machine learning perspective. *Technological Forecasting and Social Change*, 162:120348.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Michael R Benjamin, Henrik Schmidt, Paul M Newman, and John J Leonard. 2010. Nested autonomy for unmanned marine vehicles with moos-ivp. *Journal of Field Robotics*, 27(6):834–875.

Erik Cambria, Lorenzo Malandri, Fabio Mercorio, Mario Mezzanzanica, and Navid Nobani. 2023. A survey on xai and natural language explanations. *Information Processing & Management*, 60(1):103111.

Devleena Das, Siddhartha Banerjee, and Sonia Chernova. 2021. 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*, pages 351–360.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36.

Maximilian Diehl and Karinne Ramirez-Amaro. 2022. Why did i fail? a causal-based method to find explanations for robot failures. *IEEE Robotics and Automation Letters*, 7(4):8925–8932.

Upol Ehsan, Pradyumna Tambwekar, Larry Chan, Brent Harrison, and Mark O Riedl. 2019. Automated rationale generation: a technique for explainable ai and its effects on human perceptions. In *Proceedings of the 24th international conference on intelligent user interfaces*, pages 263–274.

Konstantinos Gavriilidis, Andrea Munafo, Wei Pang, and Helen Hastie. 2023. A surrogate model framework for explainable autonomous behaviour. *arXiv preprint arXiv:2305.19724*.

Balint Gyevnar, Massimiliano Tamborski, Cheng Wang, Christopher G Lucas, Shay B Cohen, and Stefano V Albrecht. 2022. A human-centric method for generating causal explanations in natural language for autonomous vehicle motion planning. *arXiv preprint arXiv:2206.08783*.

Helen Hastie, Francisco Javier Chiyah Garcia, David A Robb, Pedro Patron, and Atanas Laskov. 2017. Miriam: a multimodal chat-based interface for autonomous systems. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, pages 495–496.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Omer Khalid, Guangbo Hao, Cian Desmond, Hamish Macdonald, Fiona Devoy McAuliffe, Gerard Dooly, and Weifei Hu. 2022. Applications of robotics in floating offshore wind farm operations and maintenance: Literature review and trends. *Wind Energy*, 25(11):1880–1899.

Emre Kıcıman, Robert Ness, Amit Sharma, and Chenhao Tan. 2023. Causal reasoning and large language models: Opening a new frontier for causality. *arXiv preprint arXiv:2305.00050*.

Todd Kulesza, Simone Stumpf, Margaret Burnett, Sherry Yang, Irwin Kwan, and Weng-Keen Wong. 2013. Too much, too little, or just right? ways explanations impact end users' mental models. In *2013 IEEE Symposium on visual languages and human centric computing*, pages 3–10. IEEE.

Pat Langley, Ben Meadows, Mohan Sridharan, and Dongkyu Choi. 2017. Explainable agency for intelligent autonomous systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, pages 4762–4763.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

David A Robb, Francisco J Chiyah Garcia, Atanas Laskov, Xingkun Liu, Pedro Patron, and Helen Hastie. 2018. Keep me in the loop: Increasing operator situation awareness through a conversational multimodal interface. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, pages 384–392.

Ilia Stepin, Jose M Alonso, Alejandro Catala, and Martín Pereira-Fariña. 2021. A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence. *IEEE Access*, 9:11974–12001.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Alpaca: A strong, replicable instruction-following model. *Stanford Center for Research on Foundation Models. https://crfm. stanford. edu/2023/03/13/alpaca. html*, 3(6):7.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Sandra Wachter, Brent Mittelstadt, and Luciano Floridi. 2017. Transparent, explainable, and accountable ai for robotics. *Science robotics*, 2(6):eaan6080.

Alan FT Winfield, Serena Booth, Louise A Dennis, Takashi Egawa, Helen Hastie, Naomi Jacobs, Roderick I Muttram, Joanna I Olszewska, Fahimeh Rajabiyazdi, Andreas Theodorou, et al. 2021. Ieee p7001: A proposed standard on transparency. *Frontiers in Robotics and AI*, 8:665729.

Jimmy Wu, Rika Antonova, Adam Kan, Marion Lepert, Andy Zeng, Shuran Song, Jeannette Bohg, Szymon Rusinkiewicz, and Thomas Funkhouser. 2023. Tidybot: Personalized robot assistance with large language models. *Autonomous Robots*, 47(8):1087–1102.

Lingling Xu, Haoran Xie, Si-Zhao Joe Qin, Xiaohui Tao, and Fu Lee Wang. 2023. Parameter-efficient fine-tuning methods for pretrained language models: A critical review and assessment. *arXiv preprint arXiv:2312.12148*.

Xinnuo Xu, Ondřej Dušek, Verena Rieser, and Ioannis Konstas. 2021. Agggen: Ordering and aggregating while generating. *arXiv preprint arXiv:2106.05580*.

Wei Zhao, Goran Glavaš, Maxime Peyrard, Yang Gao, Robert West, and Steffen Eger. 2020. On the limitations of cross-lingual encoders as exposed by reference-free machine translation evaluation. *arXiv preprint arXiv:2005.01196*.