

Selective Exploration and Information Gathering in Search and Rescue Using Hierarchical Learning Guided by Natural Language Input

Dimitrios Panagopoulos¹, Adolfo Perrusquía¹ and Weisi Guo¹

Abstract—In recent years, robots and autonomous systems have become increasingly integral to our daily lives, offering solutions to complex problems across various domains. Their application in search and rescue (SAR) operations, however, presents unique challenges. Comprehensively exploring the disaster-stricken area is often infeasible due to the vastness of the terrain, transformed environment, and the time constraints involved. Traditional robotic systems typically operate on predefined search patterns and lack the ability to incorporate and exploit ground truths provided by human stakeholders, which can be the key to speeding up the learning process and enhancing triage. Addressing this gap, we introduce a system that integrates social interaction via large language models (LLMs) with a hierarchical reinforcement learning (HRL) framework. The proposed system is designed to translate verbal inputs from human stakeholders into actionable RL insights and adjust its search strategy. By leveraging human-provided information through LLMs and structuring task execution through HRL, our approach not only bridges the gap between autonomous capabilities and human intelligence but also significantly improves the agent's learning efficiency and decision-making process in environments characterised by long horizons and sparse rewards.

I. INTRODUCTION

Autonomous intelligent robots are expected to be deployed in a broader range of real-world applications [1]. For instance, in the aftermath of natural or man-made disasters, search and rescue (SAR) robots are utilised to assist in tasks such as searching operations, incidents reporting, locating missing people, and providing aid to those affected within the impacted area [2]. In this high-stakes domain, despite considerable advancements, SAR robots continue to face significant limitations in terms of decision-making, task execution, and adaptability. These limitations stem from the robot's dependency on preset behaviors, significant environment change/degradation, and accurate data input by human operators (e.g., upload updated viable roads/bridges [3], [4] and triage priority areas for SAR [5]). However, this is in contention with the future vision of having fully-autonomous agents with the capacity to emulate human decision-making processes and adapting in real-time [6], especially within human-robot interaction (HRI) regimes.

A. State-of-the-Art and Gaps

Here, we consider how to make use of the user contributed information provided into the learning scheme. The literature suggests various forms of human guidance to aid learning

algorithms, ranging from demonstrations and advice to preference and online evaluative feedback [7]. The proposed work, nevertheless, is specifically designed for scenarios that humans have no access to the agent's internal software implementation. Instead, the feedback in these instances is communicated through natural language and converted into grounded insight. In fact, we know information theoretical bounds to information aided human-robot SAR exist [8], but how this can be achieved in reality is not clear.

In this research area, there is widespread acknowledgment that the utility of informative feedback is beneficial to the learning agent [9]. Information about the environment's structure is densely informative, helping the agent reduce exploration and quickly find the optimal strategy [10]. Early work examined the use of side information from mobile signals or sensor networks [11], however this lacks semantic detail and availability.

Current SAR systems deployed in these scenarios fail to actively seek, collect, and exploit contextual information from human stakeholders, which can be important to speeding up the learning progress and enhancing the efficiency of the search. This oversight results in a significant under-utilisation of potentially vital information and insights that could otherwise direct efforts more accurately and promptly [12]. In contrast, humans naturally tend to seek assistance when faced with challenging tasks, particularly when there is a lack of sufficient information and knowledge regarding the operational environment. This inclination to request help originates from our inherent desire for collaboration, problem-solving, and leveraging collective knowledge and expertise [13].

In view of the above, the challenge lies in extending SAR robots' capabilities beyond mere execution of tasks to becoming active participants in the problem solving process. This is especially crucial in large-scale and dynamic disaster environments, where much of the necessary information, initially unknown or inaccessible to rescue personnel, is scattered across various sources [14]. Recognising the dynamic nature of such environments, where prior knowledge may be limited or outdated, we propose the integration of human linguistic inputs into the agent's learning process as a critical enhancement over reliance solely on environmental cues.

B. Opportunities in LLMs and HRL

Large Language Models (LLMs) [15], with their advanced natural language understanding capabilities, emerge in bridging the communication gap between SAR robots and humans by providing a variety of constructive roles in

¹School of Aerospace, Transport and Manufacturing, Cranfield University, Bedford, UK, {d.panagopoulos, adolfo.perrusquia-guzman, weisi.guo}@cranfield.ac.uk

solving planning tasks [16]. Furthermore, the complexity and scale of disaster environments necessitate an approach that goes beyond simple task execution. Hierarchical Reinforcement Learning (HRL) [17] offers a structured method to address this challenge by breaking down complex tasks into more manageable subtasks. HRL's emphasis on learning at multiple levels of abstraction and operating on a subset of the state space makes it particularly suited for environments with delayed rewards, such as those encountered in SAR operations.

Assuming that the agent is capable of interpreting the information signal into an actionable insight, the retained feedback can promote or discourage behaviour before it is presented [18]. However, encoding human input into a shape that an agent understands can be a complicated process. This is where the integration of LLMs with robotic systems marks a significant advancement in the field, bridging the communication gap between humans and machines [19]. LLMs, especially those tailored with domain-specific knowledge and trained on relevant datasets, show promise and highlight the potential in enhancing decision-making processes in emergency situations [20]. Recent advancements in integrating LLMs into RL paradigms have shown promise in addressing some of these limitations, as highlighted in [21]. The LLM can act as an information processor, extracting meaningful insights for the agent from natural language, thus enhancing the agent's natural language understanding. However, the potential of combining RL with the nuanced comprehension abilities of LLMs has not been fully exploited. A survey has called for the potential uses of NLP techniques in RL, but the capabilities of LLMs were limited at that time [22]. This synergy [23] could revolutionise how SAR robots process and act upon human-provided information in real-time.

C. Novelty

In this paper, the approach considers the agent acting as a first responder in disaster scenarios. Leveraging LLMs to interpret and convert human linguistic inputs into actionable commands enables interaction between SAR robots and non-technical individuals¹ on-site. By doing so, the proposed system facilitates real-time communication and allows the robot to actively collect, process, and utilise the insights provided by humans, which are crucial in the early stages of disaster response. In addition to that, an HRL framework is adopted to structure the agent's task process [24], effectively addressing the challenges caused by long horizons and sparse rewards. Acting as a first responder in these language-rich environments, the agent uses a hierarchy of decision making to not only solve immediate problems efficiently, as dictated by RL principles, but also to strategically and safely integrate and act upon collected information, improving its learning and operational efficiency.

The main contributions of this paper are summarised as follows: **(1)** The development of a general, task-oriented hierarchical planning framework for the operation of SAR robots,

which includes the use of specialised agents for different subtasks. **(2)** A novel architecture for human-in-the-loop integration and policy shaping, merging LLMs with HRL to empower SAR robots to interpret and act upon human linguistic inputs in real-time. **(3)** An extension of the utility of HRL within SAR operations through the proposed method, further enriched by infusing domain-specific knowledge into the LLM using the Retrieval-Augmented Generation (RAG) pipeline [25], enabling the agent to identify, prioritise, and efficiently seek out specific information.

This paper delves into a novel and critical challenge in the realm of emerging crisis response. Crucial information distributed across a crisis scene often remains unexploited, leading to inefficiencies in response strategies and outcomes. By investigating this issue, our work highlights the importance of leveraging this dispersed information within the learning frameworks of SAR operations, ultimately leading to context-sensitive learning. The findings show that descriptive information about the environment (e.g., new hazards or updated victim locations), which may not be immediately available through standard data sources, significantly improves the agent's success on completing the task.

The rest of the paper is organised as follows. Section II presents an overview of the proposed conceptual architecture. Section III provides the case study setting followed by experimental evaluation and analysis. In Section IV, the results of our experiment are presented and discussed and we point directions for future work. Finally, Section V concludes our proposed work.

II. PROBLEM STATEMENT AND MODELLING

A. Conceptual Architecture

Our aim is to design a novel formulation that incorporates mechanisms to accelerate the learning process of an agent within a HRL framework, integrated with a LLM for processing verbal inputs. The proposed system is illustrated in Fig. 1 and comprises several key components:

- **Context Extractor:** This module processes verbal inputs communicated to the robot, using a pre-trained LLM to parse and interpret these inputs, and generate a structured contextual representation. The latter, encapsulating key information from the verbal inputs, is then fed to the Strategic Decision Engine (SDE).
- **Information Space:** This set of predefined information types serves as a guide or map for the agent, aligning its actions with the mission's strategic goals. It serves as a reference for the SDE, ensuring that actions remain on track with these objectives.
- **Strategic Decision Engine (SDE):** Operating as a manager within the hierarchical framework, the SDE makes strategic decisions based on the environment's state, the context from the Context Extractor, and directives from the Information Space. The decisions guide the agent towards actions aligned with context and mission priorities.
- **Attention Space:** Situated within the SDE, this dynamic space influences the agent's decision-making by em-

¹those not involved with the agent's operational software implementation

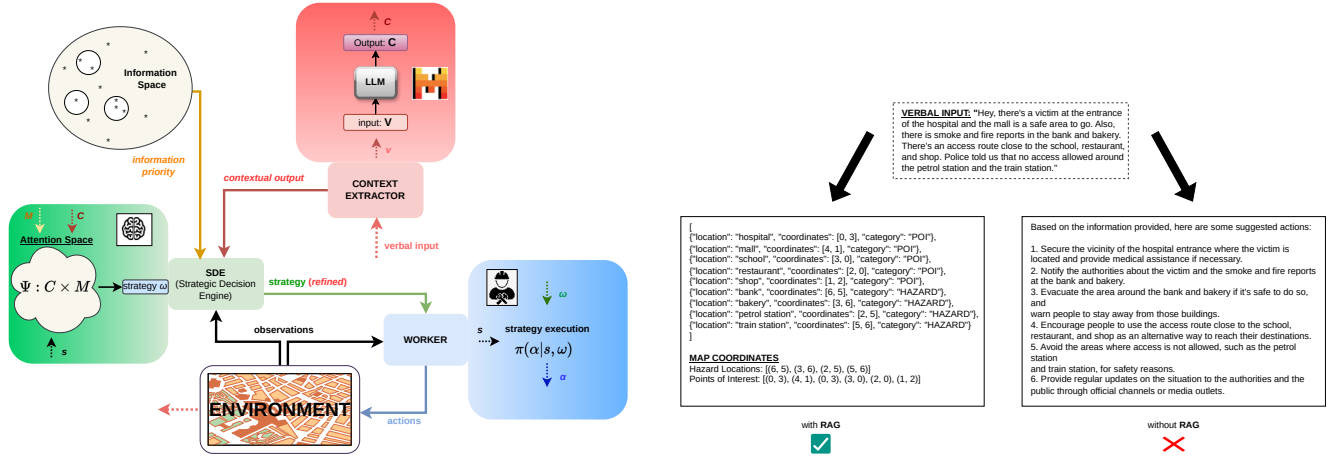


Fig. 1. **Left:** The figure illustrates the proposed pipeline within a hierarchical decision-making framework. The *Environment* provides observations s to both the *SDE* and the *Worker* modules. When these observations s contain verbal input v , the latter is directed to the *Context Extractor*, which then generates contextual outputs c . These outputs c , along with observations s from the *Environment* and information priorities M specified by the *Information Space*, are channelled into the *SDE*. Within the *SDE*, strategies ω are refined, taking into account the *Attention Space*. The *Worker* module, informed by these refined strategies ω , executes primitive actions α within the *Environment*. As a result of these interactions, the system continually adjusts and updates its policies, creating a dynamic feedback loop that evolves over time. **Right:** Comparison of outputs from LLMs with and without RAG integration in simulated SAR operations, demonstrating enhanced task-specific detail and notation in outputs.

phasizing certain aspects of the context and directive information, guiding policy adjustments. It steers the agent towards more context-informed decisions.

- **Worker:** This execution module is activated once the *SDE* selects a strategy, carrying out the corresponding sequence through primitive actions that interact with the environment.

B. Modelling

In HRL, a common approach involves using a hierarchy of decision-makers, such as a manager and workers, where the manager selects which subtask to execute, and each subtask is governed by its own policy.

We use the formalism of the Markov decision process (MDP), given by a 7-tuple $(S, A, \Omega, \beta_\omega, P, R, \gamma)$, where S is the state space, A is the action space, Ω is the set of strategies each with termination condition β_ω , P is the transition function, R is the reward function, γ is the discount factor.

Several elements in the proposed hierarchical framework extend beyond the basic MDP tuple but are crucial for the decision-making process. Specifically, V is a set of verbal inputs $\{v_1, \dots, v_m\}$, where each v_i represents a piece of information related to the task; C denotes contextual details $\{c_1, \dots, c_n\}$, derived from V ; $M = \{(i_1, p_1), \dots, (i_k, p_k)\}$ represents the information types and their associated priorities, where i_j is the type and p_j is its priority; $\pi_\Omega: S \rightarrow \Omega$ is a meta-policy function that maps states to strategies; $\pi_\omega: S \rightarrow A$ is a function that maps states to actions under each strategy ω ; $Q(s, \omega)$ evaluates the expected reward of choosing strategy ω in state s ; $Q(s, \omega, \alpha)$ evaluates the expected reward of choosing action α under strategy ω in state s ; $L: V \rightarrow C$ is a transformation function that processes verbal inputs to generate contextual information; Ψ is the attention space that refines policy functions based on the encoded context and information priorities.

Within this framework, the agent operates on two levels of hierarchy: **(1) High-Level: SDE** The manager decomposes the overall task into smaller, manageable subtasks, sets priorities and the hierarchy for executing these tasks. It selects appropriate strategies for each subtask and monitors and adjusts these strategies based on performance and changes in the environment. **(2) Low-Level: Worker** The worker is responsible for executing the subtasks assigned by the manager. This involves directly interacting with the environment through specific actions determined by sub-policies π_ω . Both layers can either be learned over time through experience or follow deterministic rules.

The objective is to find the optimal policies π_Ω and π_ω that maximise the expected reward over time, while also considering the influence of the attention space Ψ in dynamically refining and guiding the policy selection process based on context and information priorities:

$$J(\pi_\Omega, \pi_\omega) = \mathbb{E} \left[\sum_{t=0}^T \gamma^t R(s_t, a_t) \mid \pi_\Omega^\Psi, \pi_\omega^\Psi \right] \quad (1)$$

where γ is the discount factor, and T the time horizon. The action a_t at time t is determined by the sub-policy π_ω if the policy ω is active, and s_t is the state of the environment at time t .

III. EXPERIMENTS

A. Simulated Environment Setup

To evaluate the efficiency of our proposed system integrating LLM with a hierarchical framework, we design a SAR simulated environment with discrete state and action space. Specifically, in the designed 2D experimental setup (see Fig. 2), an agent is tasked with operating in a disaster-stricken area, aiming to rescue victims while avoiding obstacles. The experiment simulates a complex scenario, where the agent

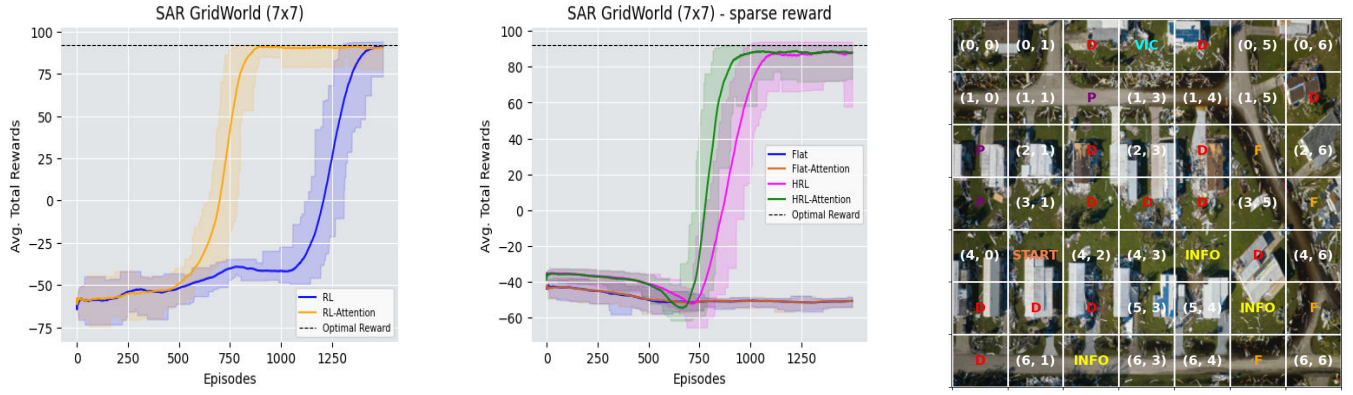


Fig. 2. Comparative Analysis of Learning Agents in SAR Scenarios. **Left:** Performance of flat RL agents, with and without attention guidance, receiving intrinsic rewards. **Middle:** Comparison of hierarchical (HRL) and flat RL agents, with and without attention guidance under sparse reward - reward is given upon successful task completion - conditions. **Right:** SAR environment configuration showing information locations marked as ‘INFO’, obstacles ‘D’, victim locations ‘VIC’, hazards ‘F’, and points of interest ‘P’.

must not only locate and save victims but also optimise the collection and processing of crucial information within the environment. As the agent navigates through designated information locations, they must learn to gather data prioritised by a predefined information space. This data includes updates on victim locations, safe routes, and potential hazards, mimicking real-world SAR operations where such information is crucial for effective decision-making. Upon receiving verbal information, the hazards and points of interest are revealed to the agent and are marked as ‘F’ and ‘P’, respectively. The agent transforms this input into contextual representation, allowing it to dynamically adapt and enhance its decision-making process through an attention space. This space refines the policies within the hierarchical framework, steering the agent’s behavior towards more context-informed decisions by exploiting the feedback. While we restrict our analysis to a simplified toy example, allowing for complete control over all variables and extensive experimentation across various configurations, this simulation is intended to replicate the complex nature of actual disaster situations. The development of the environment largely adheres to the structure and conventions of a custom environment in the OpenAI Gym framework [26]. The code is available at <https://github.com/dimipan/HRL-LLM>.

TABLE I
PERFORMANCE METRICS OF DIFFERENT MODELS

Model	Collisions	Steps	Avg. Reward
Flat (no sparse)	3	24	-19.8
Flat-Att (no sparse)	0	26	26.5
HRL (sparse)	3	24	14.12
HRL-Att (sparse)	0	26	22.8

B. Implementation Details

Here, we elaborate on the functionality of key elements and demonstrate the practical application of our proposed system through a case study. All components of the system

run on a single machine equipped with an RTX A2000 GPU. Our system employs the Mistral 7B [27] pre-trained LLM, integrated within a local RAG pipeline. Embedding models and LLMs are hosted using Ollama, with a local Chroma instance as the vector store, and Langchain orchestrates all processes. The model was selected for its suitability in real-time applications, offering quick responses and outperforming even models with larger parameter counts. Additionally, we infuse a JSON object containing structured data with sets of critical keywords and mapped locations, complete with precise coordinates. The information space categorises essential SAR information [28] into three types: victim details (X), navigation routes (Y), and environmental hazards (Z), with the requirement that this sequence must be respected during the collection process. The SDE dynamically selects strategies based on conditions reflected in the agent’s state representation. The latter includes the agent’s position on the grid, the status of each type of information—indicated in a binary format to show whether it has been collected or not—and an additional indicator that specifies whether a victim has been saved or not. Specifically, strategy π_{EXP} is selected when the agent needs to navigate to information points or the final location. Strategy π_{COL} is chosen when the agent reaches information points to gather data. Strategy π_{OPE} is selected when the agent must perform critical triage tasks. The worker interacts with the environment by executing the strategies decided by the SDE through primitive actions. In particular, these actions are abstracted into different types depending on the current strategy. Within strategy ‘EXPLORE’ movement actions {‘up’, ‘down’, ‘left’, ‘right’} are enabled. When strategy ‘COLLECT’ is active a set of actions {‘A’, ‘B’, ‘C’, ‘X’, ‘Y’, ‘Z’} reflects the agent’s ability to get information. Finally, within strategy ‘OPERATE’ the agent can choose from {‘save’, ‘use’, ‘remove’, ‘carry’}. Both strategies and primitive actions can be refined by the attention space via policy shaping, allowing the agent to prioritise certain aspects according to the context extracted from human feedback. The execution of each worker associated with a strategy is

trained and learned through RL. The training process entails running the system across 1500 episodes for a total of 50 runs, utilising the Q-learning algorithm. Throughout these episodes, the performance of the system is evaluated by averaging the rewards obtained in each run. A decaying ϵ -greedy action selection method is employed with an initial ϵ value 1.0 and linearly decaying to a minimum of 0.01 at a decay rate of 2. The discount factor γ and learning rate α are set at 0.998 and 0.1, respectively. When the agent exploits the collected information and acts based on the attention space, the ϵ decay is steeper.

C. Hypotheses

In our experimental evaluation, we study the following hypotheses: **(1)** The use of an LLM infused with domain-specific knowledge through the RAG pipeline produces context-informed outputs that are more relevant and accurate in SAR scenarios than those generated without RAG [29]. **(2)** The integration of the attention space into a flat RL agent accelerates the learning process, resulting in better and faster convergence. **(3)** The hierarchical setup is particularly effective in sparse reward environments, where rewards are only granted upon task completion, outperforming flat RL setups. **(4)** The integration of the attention space into the hierarchical further improves the agent’s performance. **(5)** Utilising the attention space in the decision making leads to a reduction in encounters with dynamic obstacles, highlighting the value of real-time, contextually aware feedback in operational settings.

IV. RESULTS & DISCUSSION

We compare the performance of agents in both flat and hierarchical structures, with and without attention guidance. The experiments were tailored to test specific hypotheses mentioned above related to the efficacy of incorporating domain-specific knowledge and attention mechanisms into learning agents.

A. Hypothesis 1: Domain-Knowledge Infused LLMs

The first hypothesis suggests that in simulated SAR operations LLMs would lack the capability to produce context-informed outputs properly aligned with the demands of the task without the RAG integration. Fig. 1 supports this hypothesis and shows the practical enhancements in output accuracy and detail achieved through RAG integration.

B. Hypothesis 2: RL with Attention Space

For the second hypothesis, we expect that integrating the LLM into a flat RL agent with an attention space would accelerate the learning process. The results from Fig. 2 indicate that agents guided by the attention space not only perform better in terms of reward obtained across episodes but also show improved learning speed. This improvement is particularly evident in scenarios where agents receives ongoing intrinsic rewards throughout the episode.

C. Hypothesis 3 & 4: HRL (with Attention Space) in Sparse Reward Environment

The third and fourth hypotheses examine the effectiveness of hierarchical task formulation in sparse reward environments, both with and without the integration of LLM and attention space. Hierarchical agents consistently outperform flat agents, which fail to learn any effective policy in environments where rewards are only dispensed upon the completion of tasks. This highlights the effectiveness of hierarchical structures in managing sparse reward environments. The addition of LLM and attention space further boosts the performance of the hierarchical agents as results indicate.

D. Hypothesis 5: Safe Navigation

Lastly, the fifth hypothesis addresses the system’s potential to reduce encounters with dynamic obstacles, as detailed in Table I. The number of collisions and steps taken refer to the learned policy, while the average reward is over the entire learning curve. The results show that the use of LLM and attention space significantly reduces the frequency of these encounters. It is worth pointing out that for the flat agents, we only consider the intrinsic reward setup where they successfully converge, thus disregarding the sparse reward configuration. Both flat and hierarchical agents without attention mechanisms complete the task with fewer steps (24 steps on average), but frequently collide with dynamic obstacles that become apparent after verbal input is communicated (3 collisions on average). In contrast, flat and hierarchical agents utilising attention mechanisms take slightly more steps to complete the task (26 steps on average) but successfully avoid collisions (0 collisions on average), demonstrating their ability to effectively use real-time, context-aware feedback to adapt their policies.

The results from these experiments highlight the potential of integrating advanced LLMs and attention mechanisms into RL systems, particularly in challenging and dynamic environments like SAR operations. Context-aware decision-making capabilities facilitated by these refinements not only improve the performance in terms of task-specific metrics but also enhance the adaptability and safety of agents operating in real-world conditions. The demonstrated effectiveness of hierarchical structures in managing complex and sparse reward scenarios suggests a promising avenue for developing more robust systems for HRI.

While our experiments provide a validation of our hypotheses, there are several limitations to address. In particular, when validating these systems in continuous domains employing deep RL techniques, policy shaping is intricate and less intuitive. Therefore, it may be advantageous to prefer action space bounding or shaping over policy shaping, as discussed here. This approach allows for a more straightforward implementation, since actions can be intuitively bounded based on well-defined physical constraints or safety requirements.

Another challenge emerges with the use of language. Real-world deployment in such scenarios often entails interacting with non-standardised, potentially unreliable linguistic inputs

from humans. This can mislead the system and make the extraction of actionable information complicated. A solution to that could be the infusion of additional documents, expert knowledge, and detailed contextual data into the LLM.

Lastly, the computational cost of employing more advanced LLMs with larger numbers of parameters must also be acknowledged. While these models may offer superior reasoning performance and better natural language understanding, their computational and resource demands are significantly higher.

V. CONCLUSIONS

This paper investigates the integration of LLMs and hierarchical learning within the context of SAR operations, demonstrating how the sophisticated interplay between advanced computational tools and human input can revolutionise SAR autonomous systems. Our approach, specifically, addresses the critical need for SAR robots to adapt rapidly and make context-informed decisions in dynamic disaster scenarios. The system's ability to prioritise predefined types of information and adjust its strategy based on immediate feedback marks a significant advancement over traditional methods, achieving a balance between rapid response capabilities and strategic information gathering. Our work highlights the untapped potential of leveraging advanced language and reasoning capabilities of LLMs to convert verbal inputs into actionable insights. By optimising learning and decision-making processes and introducing an attention space that refines strategy based on context, our paper opens new avenues for more context-sensitive and efficient robotic responses in high-stakes environments, potentially setting new standards for future implementations.

REFERENCES

- [1] M. Soori, B. Arezoo, and R. Dastres, "Artificial intelligence, machine learning and deep learning in advanced robotics, a review," *Cognitive Robotics*, vol. 3, pp. 54–70, 2023. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2667241323000113>
- [2] M. El Debeiki, S. Al-Rubaye, A. Perrasquía, C. Conrad, and J. A. Flores-Campos, "An advanced path planning and uav relay system: Enhancing connectivity in rural environments," *Future Internet*, vol. 16, no. 3, p. 89, 2024.
- [3] P. Arya, W. Guo, S. K. Ahmed, and L. Irwanda, "Deep learning for bridge load capacity estimation in post-disaster and -conflict zones," *Royal Society Open Science*, vol. 6, 2019.
- [4] L. Lu and W. Guo, "Automatic quantification of settlement damage using deep learning of satellite images," in *2021 IEEE International Smart Cities Conference (ISC2)*, 2021, pp. 1–6.
- [5] Y. Liu and G. Nejat, "Robotic urban search and rescue: A survey from the control perspective," *Journal of Intelligent & Robotic Systems*, vol. 72, pp. 147–165, 2013.
- [6] Z. X. et al., "The rise and potential of large language model based agents: A survey," 2023.
- [7] R. Zhang, F. Torabi, L. Guan, D. H. Ballard, and P. Stone, "Leveraging human guidance for deep reinforcement learning tasks," *arXiv preprint arXiv:1909.09906*, 2019.
- [8] R. Krzysiak and S. Butail, "Information-based control of robots in search-and-rescue missions with human prior knowledge," *IEEE Transactions on Human-Machine Systems*, vol. 52, no. 1, pp. 52–63, 2022.
- [9] A. Bignold, F. Cruz, R. Dazeley, P. Vamplew, and C. Foale, "Human engagement providing evaluative and informative advice for interactive reinforcement learning," *Neural Computing and Applications*, pp. 1–16, 2022.
- [10] Z. Guo, C. Yao, Y. Feng, and Y. Xu, "Survey of reinforcement learning based on human prior knowledge," *Journal of Uncertain Systems*, vol. 15, no. 01, p. 2230001, 2022.
- [11] A. Albanese, V. Sciancalepore, and X. Costa-Pérez, "Sardo: An automated search-and-rescue drone-based solution for victims localization," *IEEE Transactions on Mobile Computing*, vol. 21, no. 9, pp. 3312–3325, 2022.
- [12] L. Lawrie, K. Gillies, E. Duncan, L. Davies, D. Beard, and M. K. Campbell, "Barriers and enablers to the effective implementation of robotic assisted surgery," *PLOS ONE*, vol. 17, no. 8, pp. 1–21, 08 2022. [Online]. Available: <https://doi.org/10.1371/journal.pone.0273696>
- [13] S. W. Kozlowski and D. R. Ilgen, "Enhancing the effectiveness of work groups and teams," *Psychological Science in the Public Interest*, vol. 7, no. 3, pp. 77–124, 2006, PMID: 26158912. [Online]. Available: <https://doi.org/10.1111/j.1529-1006.2006.00030.x>
- [14] S. Waring, L. Alison, N. Shortland, and M. Humann, "The role of information sharing on decision delay during multiteam disaster response," *Cognition, Technology & Work*, vol. 22, pp. 263–279, 2020.
- [15] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang, W. Ye, Y. Zhang, Y. Chang, P. S. Yu, Q. Yang, and X. Xie, "A survey on evaluation of large language models," *ACM Trans. Intell. Syst. Technol.*, vol. 15, no. 3, mar 2024. [Online]. Available: <https://doi.org/10.1145/3641289>
- [16] S. Kambhampati, K. Valmeekam, L. Guan, K. Stechly, M. Verma, S. Bhabri, L. Saldy, and A. Murthy, "LLMs can't plan, but can help planning in llm-modulo frameworks," 2024.
- [17] S. Pateria, B. Subagdja, A.-h. Tan, and C. Quek, "Hierarchical reinforcement learning: A comprehensive survey," *ACM Comput. Surv.*, vol. 54, no. 5, jun 2021. [Online]. Available: <https://doi.org/10.1145/3453160>
- [18] S. McCallum, M. Taylor-Davies, S. Albrecht, and A. Suglia, "Is feedback all you need? leveraging natural language feedback in goal-conditioned rl," in *NeurIPS 2023 Workshop on Goal-Conditioned Reinforcement Learning*, 2023.
- [19] J. Wang, Z. Wu, Y. Li, H. Jiang, P. Shu, E. Shi, H. Hu, C. Ma, Y. Liu, X. Wang et al., "Large language models for robotics: Opportunities, challenges, and perspectives," *arXiv preprint arXiv:2401.04334*, 2024.
- [20] A. Chandra and A. Chakraborty, "Exploring the role of large language models in radiation emergency response," *Journal of Radiological Protection*, 2024.
- [21] Y. Cao, H. Zhao, Y. Cheng, T. Shu, G. Liu, G. Liang, J. Zhao, and Y. Li, "Survey on large language model-enhanced reinforcement learning: Concept, taxonomy, and methods," 2024. [Online]. Available: <https://arxiv.org/abs/2404.00282>
- [22] J. Luketina, N. Nardelli, G. Farquhar, J. Foerster, J. Andreas, E. Grefenstette, S. Whiteson, and T. Rocktäschel, "A survey of reinforcement learning informed by natural language," *arXiv preprint arXiv:1906.03926*, 2019.
- [23] M. Pternea, P. Singh, A. Chakraborty, Y. Oruganti, M. Milletari, S. Bapat, and K. Jiang, "The rl/llm taxonomy tree: Reviewing synergies between reinforcement learning and large language models," 2024. [Online]. Available: <https://arxiv.org/abs/2402.01874>
- [24] M. Eppe, C. Gumbsch, M. Kerzel, P. D. Nguyen, M. V. Butz, and S. Wermter, "Intelligent problem-solving as integrated hierarchical reinforcement learning," *Nature Machine Intelligence*, vol. 4, no. 1, pp. 11–20, 2022.
- [25] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel et al., "Retrieval-augmented generation for knowledge-intensive nlp tasks," *Advances in Neural Information Processing Systems*, vol. 33, pp. 9459–9474, 2020.
- [26] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba, "Openai gym," *arXiv preprint arXiv:1606.01540*, 2016.
- [27] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. d. I. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier et al., "Mistral 7b," *arXiv preprint arXiv:2310.06825*, 2023.
- [28] E. Katsadouros, D. G. Kogias, C. Z. Patrikakis, G. Giunta, A. Dimou, and P. Daras, "Introducing the architecture of faster: A digital ecosystem for first responder teams," *Information*, vol. 13, no. 3, p. 115, 2022.
- [29] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, and H. Wang, "Retrieval-augmented generation for large language models: A survey," *arXiv preprint arXiv:2312.10997*, 2023.



DSPACE

<https://dspace.org/>

Selective exploration and information gathering in search and rescue using hierarchical learning guided by natural language input

Panagopoulos, Dimitris; Perrusquía, Adolfo; Guo, Weisi

2024-10-06

Institute of Electrical and Electronics Engineers (IEEE)

<https://doi.org/10.1109/smc54092.2024.10831125>

Panagopoulos D, Perrusquia A, Guo W. (2024) Selective exploration and information gathering in search and rescue using hierarchical learning guided by natural language input. In: 2024 IEEE International Conference on Systems, Man, and Cybernetics (SMC). In: 2024 IEEE International Conference on Systems, Man, and Cybernetics (SMC), 6-10 October 2024, Kuching, Malaysia