LLM-MARL: Large Language Model-Augmented Multi-Agent Reinforcement Learning Framework for Interpretable and Adaptive Robotic Disaster Response

Sheema Firdous

List of Abbreviations

AI – Artificial Intelligence

RL – Reinforcement Learning

MARL – Multi-Agent Reinforcement Learning

LLM – Large Language Model

HITL – Human-in-the-Loop

PPO – Proximal Policy Optimization

MAPPO – Multi-Agent Proximal Policy Optimization

MADDPG – Multi-Agent Deep Deterministic Policy Gradient

QMIX – Monotonic Value Function Factorisation for MARL

CLIP – Contrastive Language–Image Pretraining

SOTA – State of the Art

LLM-MARL – Large Language Model-Augmented Multi-Agent Reinforcement Learning

RLHF – Reinforcement Learning with Human Feedback

SSRO – Search and Secure Rescue Operations

1. Introduction

Large-scale Language Models (LLMs) (Achiam et al., 2023; Kumar, 2024) have revolutionized the digital realm, outperforming many textual and vision-based Machine Learning (ML) and Deep Learning (DL) procedures in Artificial Intelligence (AI) realm (Tom et al., 2020). Their impact spans across digital domains, from zero-shot reasoning and task planning to vision-language understanding and instruction following, with minimal time complexity (Kernycky et al., 2024). Despite these breakthroughs, the LLMs' potential has not been fully explored in Robotics domain, particularly in the realm of Multi-Agent Reinforcement Learning (MARL), that involves multiple autonomous agents (robots) communicating together and enforcing the learning from their own actions/ moves along with environment parameters (Tang et al., 2025). While MARL enables decentralized coordination in dynamic environments, it often suffers from critical limitations including *poor interpretability, adaptiveness across diverse scenarios, and real-time human insights* that makes it a sub-optimal choice for reliable deployment in dynamic settings such as `Search and Secure Rescue Operations` (`SSRO`) (Yang et al., 2024).

This research leverages the transformative potential of LLMs to address the formidable communication challenges of robotic agents operating in `SSRO` with MARL settings. Aiming first-of-its-kind, the research study presents a novel hybrid framework that exploits the semantic reasoning power of LLMs to overcome key weaknesses in MARL-based disaster robotics. Specifically, the project aims to integrate LLMs as *behavior modulators* (via personality-conditioned prompting), *semantic communication compressors* (generating interpretable messages), and *mission-level planners,* that support human-in-the-loop decision-making. The proposed `LLM-MARL` fusion is aimed to be tested in the *Erebus disaster simulation environment*[1], for multiple disaster types in `SSRO` scenarios. The proposed system is aimed to enhance *interpretability, adaptability, and real-time interaction*, thereby advancing the current state of autonomous multi-agent coordination in high-risk environments.

2. Research Questions

This research work addresses the following research questions:

1. How can LLMs be integrated with RL methods to target the disaster scenarios with primary objective of enhancing *interpretability, adaptability*, and *communication*?
2. To what extent can LLM-driven semantic compression and role modulation enable more effective communication and coordination among heterogeneous agents in hazard scenarios?
3. Without destabilizing robot's learning, and prioritizing the time constraint, how can real-time human-in-the-loop feedback can be incorporated?
4. What ethical concerns could arise when implementing `LLM-MARL` architecture in real-world scenario?

[1] Webots-based rescue robotics platform designed for testing autonomous navigation, obstacle avoidance, and victim detection in disaster-like scenarios.

3. Aims and Objectives

The main aim of this study is to implement LLM-augmented Multi-Agent Reinforcement Learning framework that enables the agents to communicate seamlessly, while considering the time constraint and allowing human-feedback in real-time.

The primary objectives of this study are listed below:

1. Develop an independent RL-based architecture to make multiple robots learn and communicate seamlessly in dynamic hazard scenarios.
2. Implement LLM-augmented methods to enhance *interpretability, and adaptability* of robotic agents.
3. Enabling the real-time human feedback in the loop to handle the unseen scenarios and enhancing the robot's learning under these circumstances.
4. Making the robot's learning dynamic to adapt to any unseen `SSRO` procedures quickly by cross-evaluating across different test environments.

4. Key Deliverables

The key deliverable of proposed research study include a detailed report on implementation of an LLM-augmented architecture leveraging RL controller to enhance the robot's learning in dynamic `SSRO` procedures while providing real-time human feedback in the loop.

5. Literature Review

This section outlines an overview of existing literature on MARL settings while primarily focusing on major techniques used in SSRO aligned with our research settings, as detailed below.

5.1. MARL in Real-World Robotics

Reinforcement Learning (Kaelbling et al., 1996; Li, 2017) has seen great attention in past few years, dominating in decentralized coordination among autonomous agents in Multi-robot scenarios (Du & Ding, 2021; Dinneweth et al., 2022). Some primary techniques include *MADDPG* architecture (Lowe et al., 2017) based on *Deep Deterministic Policy Gradient* [2] framework, *QMIX* (Rashid et al., 2020), a *value-based MARL* approach using monotonically mixing to decentralize the training, and MAPPO (Yu et al., 2022), a MARL adaptation of *Proximal Policy Optimization* [3] employing centralized critics while decentralized agents, have been used to coordinate teams of robots in simulated environments. These systems heavily rely on centralized training and decentralized execution, ideally in best-case the ideal use-cases (Yu et al., 2022). However, they fail in dynamic and partially observable environments such as disaster scenarios, primarily due to the challenges of non-stationary environmental conditions and very limited generalization (Marinescu et al., 2017).

Some studies were conducted to address these challenges with *centralized critics* by (Foerster et al., 2018) leveraging global and local actions of robots, or *curriculum learning approach* introduced by (Wei & Ding, 2021). Still, the conducted experimentation in static or well-structured task domains poses a significant challenge for these methodologies. When implemented pipelines were evaluated with unstructured environments e.g. SSRO, the lack of *interpretability, adaptability, and human oversight* became a formidable challenge. MARL agents lack transparent behavior, making it difficult for human operators to understand or intervene during mission-critical decisions (Panait & Luke, 2005; Zhang et al., 2021). Our proposed methods will employ the LLMs to overcome the addressed challenges in dynamic environments.

5.2. Emergent Communication and Its Interpretability Problem

In traditional approaches to multi-robotics domain, many research studies focused on enhancing the communication among the robots. The primary techniques include *emergent communication mechanisms* (Foerster et al., 2016; Sukhbaatar et al., 2016) that enables agents to learn differentiable communication protocols during model training. A subsequent study involves *targeting Multi-agent Communication* framework introduced by (Das et al., 2019), where agents

[2] An actor-critic reinforcement learning algorithm designed for environments with continuous action spaces, combining Q-learning with deterministic policy gradients.

[3] A policy gradient method in reinforcement learning that improves training stability by constraining policy updates to stay within a trust region.

selectively send messages to relevant teammates. Same methodology was used by (Sukhbaatar et al., 2016) who showed that learned message passing improves collaboration, especially in cooperative tasks. However, these messages are usually high-dimensional vectors, limiting the degree of interpretation by humans.

Aligned with our research settings, (Mordatch & Abbeel, 2018) explored compositional language emergence in multi-agent games but limited the experimentation to simulation only, using artificial symbols without real-world grounding. In contrast to this study, (Jiang & Lu, 2018) proposed *attentional communication* [4], yet these efforts couldn't prioritize human-readable messages. With same research settings, (Sheng et al., 2022) argued for *structured message protocols*, but the main issues reported were: 1) limited generalization and 2) scalability to the other environmental configurations. Critically, these vector-based or symbolic messages are opaque, non-semantic, and not aligned with natural language, posing a research gap between agents and human supervisors (Dessi et al., 2021; Brandizzi, 2023).

5.3. Enabling Agent Diversity and Specialization in MARL

Given the hardware barriers of MARL robots with limited capabilities, some recent studies have focused architectural considerations with SOTA methodologies. Majority of existing robotic agents rely on homogeneous policies where all agents share the same architecture, behavior, and reward objectives (Lowe et al., 2017; Rashid et al., 2020; Yu et al., 2022). Despite of offering simplified training and scaling in symmetric tasks, the architecture fails to reflect the *diversity of agent roles* that are critical in real-world scenarios where time complexity is the main constraint. Particularly in complex, multi-phase disaster environments where agents should act differently e.g., some prioritizing exploration, others focusing on rescue or communication, handling diversity becomes highly important to increase the applicability to real-world scenarios. This problem was addressed by some recent research studies including work by (Jaques et al., Social influence as intrinsic motivation for multi-agent deep reinforcement learning, 2019; Du et al., 2019) who proposed explicit role-based behaviors but the proposed methods have dependency limitation on hardcoded roles or reward shaping to the agents, limiting the robots ability to adapt dynamically in a given mission context. In subsequent studies, the potential of LLMs was leveraged to prompt agent behavior dynamically with natural language cues (Xu et al., 2022; Sun et al., 2024), enabling *dynamic personality modulation* across agents.

Despite of these advancements, to the best of our knowledge, no existing MARL system fully integrates *LLM-based dynamic behavior* to induce *diverse, contextual agent specializations***.** Even some SOTA emerging prompt-based systems e.g. *Code-as-Policies* (Liang et al., 2023) and *SayCan* (Ahn et al., 2022) remain limited to single-agent control with fixed instruction-response dynamics. It leaves a research gap in using LLM based methods to assign, evolve, and interpret agent personalities in multi-agent systems, specifically when roles must shift in real time due to

[4] A mechanism in multi-agent systems where agents selectively focus on and exchange relevant information using attention-based models to enhance coordination and decision-making.

disaster evolution, communication loss, or environment changes, while maintaining the time complexity.

5.4. Human-in-the-Loop Systems and the Case for Interpretable Agents

Human-in-the-loop (HITL) systems are critical in high-risk domains like `SSRO` scenarios, where prompt oversight, intervention, or guidance may be needed at given timestamp (Cai et al., 2024). Considering the recent studies, most human interaction is limited to *pre-programmed rules*[5] or *manual overrides* [6]. LLMs-based methods have potentially bridged this gap in recent studies by offering language interface as introduced by (Cai et al., 2024) who leveraged transformers architecture (Vaswani et al., 2017) to provide a user interface that allow agents to share goals with humans. Visual scenes were also interpreted by (Radford et al., 2021) using *CLIP LLM*, to offer semantic-level interactions. However, all of these studies focused single-agent setup and static instruction pipelines. The research was expanded to `MARL` scenario by (Kumar, 2024) for real-time, LLM-mediated interaction between human operators and `MARL` agent teams to increase trust and effectiveness. This work serves as a baseline for our human-in-the-loop feature by additionally offering feedback integration and enhanced security features, making this study the-first-of-its-kind to address this research gap in `SSRO` simulation platform.

5.5. Semantic Compression and Personality-Based Behavior Modulation

A novel but underexplored area is using LLMs to both *compress communication semantically* and *modulate behavior* through *personality prompts*. (Kumar, 2024) provide a strong baseline for future directions by presenting a novel methodology, to use LLMs for converting the observations into brief, interpretable phrases (e.g., "Obstacle detected near Zone 2"). This refers to a form of *semantic compression* that reduces bandwidth while maintaining communication clarity. With same research aim, (Liu et al., 2024) show that distilled LLMs like *TinyLLaMA* (Zhang et al., 2024) can run on edge devices, making this vision feasible. The research argument was hypothesized by (Xu & Shen, 2022) who showed that prompting agents with behavioral roles can diversify policies and improve coverage in exploratory tasks. This *prompt-conditioning* approach can be expanded to SSRO environment where some agents explore aggressively, and other robots prioritize victim safety or structural avoidance. Such *LLM-modulated personalities* can create heterogeneity and robustness in `MARL` teams (Jaques et al., 2019). Considering the research gap to apply this conditioning in `MARL` frameworks, we hypothesize that these techniques enable LLMs to enhance both *what agents say* (communication) and *how they act* (behavior modulation) in our research settings.

[5] Static programming rules, embedded with robotics default logic.

[6] User-initiated actions that bypass automated systems to enforce direct control or correction.

6. Research Design

This section outlines the detailed methodology to propose LLM-MARL architecture. The proposed method is a 3-layered procedure, starting from *MARL Backbone* (that involves the baseline model selection for multi-agent scenario implementation, *LLM Augmentation* (this involves advancing the baseline model with LLM based methods), and *Human-In-The-Loop Feedback Module incorporation* (that leverages the SOTA techniques with LLMs to incorporate the real-time human feedback in the SSRO scenarios). This three layered architecture provides the flexibility of focusing primary research objectives individually while considering the most important factors and parameters that could lead to enhance the effective learning in minimal time constraint. The research design layers are discussed in detail below and depicted in Figure 1.

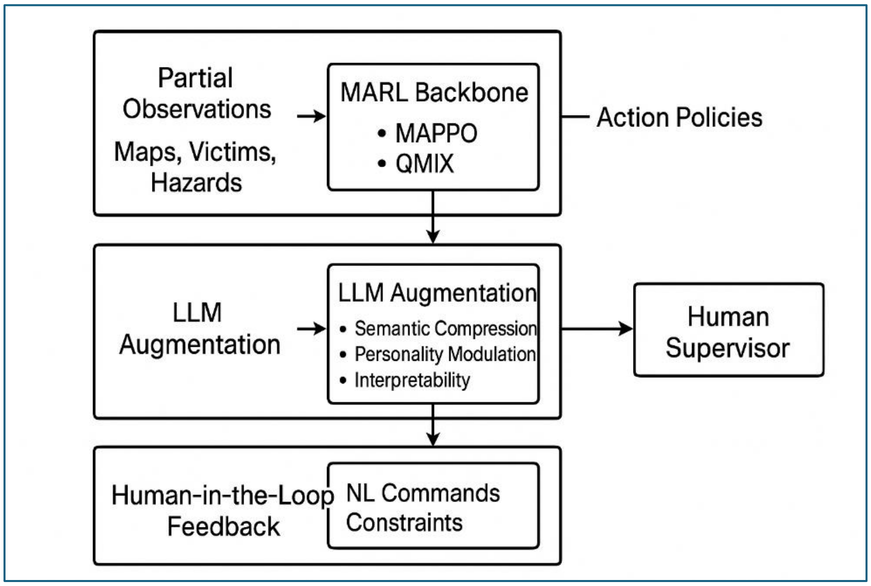

Figure 1 System Architecture Diagram

Layer 1: MARL Backbone

The core of proposed work lies in the Multi-Agent Reinforcement Learning (MARL) backbone that provides the fundamental decision-making capabilities for autonomous robotic agents operating in hazard scenarios. The architecture was selected due to its de facto paradigm enabling to learn through trail-and-error manners in complex environments (Zhang et al., 2021). To select the algorithms for MARL pipeline, rigorous study was done and pre implementation testing was done to choose **MAPPO (Multi-Agent Proximal Policy Optimization)** and **QMIX** as the baseline learning models.

- **MAPPO** extends the popular single-agent PPO framework to multi-agent domains, offering stability and scalability.
- **QMIX** introduces a value-decomposition mechanism that allows the global Q-function to be expressed as a **monotonic combination of individual agent Q-functions**.

When multiple robots operating together, each agent perceives the real-time data from its environment, but partially, including current position, current visible scene in front camera, sensors (proximity/ distance + ground sensors) values. The parameter set is aggregated with other robot's values and considered as inputs to MARL that computes action policies (e.g. forward, exploration area etc.). This training helps the robot develop the fundamentals learning including *navigation, exploration, and cooperation* with other agents.

The training leverages the *Main Supervisor* (main controller having full access to the system state of all robotic agents) that allows efficient coordination. This *Supervisor* controller will be used for initial training phase only, when the architecture is deployed, agents will rely on their local observations and learned policies that enables robust decentralized execution seamlessly.

Layer 2: LLM Augmentation

The second layer leverages the potential of LLMs to enhance the baseline system. LLMs are large-scale language models trained on trillions of data examples to produce the robust results. The LLM-based approaches have outperformed the traditional methods in thousands of machine learning tasks, making them the state-of-the-art method for ML procedures. Many LLMs were analyzed to see their transformative potential for augmentation task. *GPT-4* (Achiam et al., 2023), *LLaMA 2* (Touvron et al., 2023), and *Phi-2* (Javaheripi et al., 2023) have demonstrated remarkable capabilities in semantic reasoning, multi-modal understanding, and adaptive problem-solving, making them the best choice for this implementation layer. The purpose of the LLM augmentation is: *to compress communication into interpretable tokens, to enable personality-based role modulation, and to provide interpretability for human supervisors.*

- The robot's raw states processed in the preceding step (e.g. current position, camera RGB values etc.) are converted into natural language-like tokens using LLM. The bandwidth is hence reduced while creating interpretable logs.
- To make the dynamic role adaptation, LLM prompting was leveraged. The prompt *"focus on coordination"* will enable the robot to focus on mutual coordination more than navigation or other subtask.
- Transparency is also enhanced by translating the robot's actions into human-like explanations e.g. *Agent 2 turned left to avoid fire zone detected in front-left at 0.2mm".* This makes the system transparent and enhance the supervisor's learning to prepare emergent strategies.

We also introduce the *prompt-driven personalities,* where agent roles can be dynamically shifted through natural language conditioning. This makes the module novel in comparison with traditional MARL approaches that treats agents as interchangeable entities leading to homogeneous behavior. For example prompting an agent with *"act as a scout"* biases its policy toward exploration. With this personality modulation along with semantic compression method, this layer enhances the baseline with *communication efficiency, agent role diversity,* and *interpretability.*

Layer 3: Human-In-The-Loop (HITL) Feedback

While LLM-augmented MARL agents improve transparency, SSRO scenarios often require real-time human intervention. Supervisors may process contextual knowledge that agents cannot infer. For instance, when a victim is located near a weak building, that agent couldn't infer in real-time. So, this layer incorporates human feedback logics in natural language style. Supervisor can issue example commands listed below.

- *"All drones focus on the northern sector."*
- *"Prioritize victim extraction over hazard mapping."*
- *"Ground robot 3, avoid the collapsed passage."*

LLMs takes these commands as input, and generate structured constraints or sub-goals that the MARL agents can incorporate into their decision making seamlessly. Also, when there is an emergency scenario, requiring gall robots to leave an action for instance, HITL will enable them to translate human input to convert into sub-goals and promptly following the *Supervisor's* instructions. A safety filter is also incorporated to prevent the infeasible and unsafe actions. When there comes a safety filter, it cross-checks LLM output against a rule-based safety layer, ensuring that only valid, safe commands are executed. This dual layer of validation ensures robustness in high-risk disaster scenarios.

A critical challenge in integrating human feedback is the risk of destabilizing the learning process. If agent policies are overwritten by frequent external commands, coordination may break down. In order to mitigate this, HITL commands are introduced primarily at the model's inference time that ensures that agent's learned policies remain stable, while still being adaptable in real time to human guidance.

7. Ethical Considerations

Despite the remarkable performance of LLM-enabled frameworks in MARL procedures, its application in real-world critical scenarios i.e. SSRO introduces significant ethical and safety concerns due to the sensitivity of the application. While the proposed architecture aims to enhance adaptability, interpretability, and human oversight, its deployment must be accompanied by rigorous ethical safeguards to ensure that benefits are maximized without exposing affected communities, responders, or the environment to undue risks. The key risks and ethical considerations have been summarized to the table below that are important to get considered when deploying the LLM−MARL architecture to a real-world scenario.

Area	Key Risk	Impact	Mitigation Strategy
Human Safety & Reliability	Unsafe actions (e.g., entering unstable areas)	Harm to victims	Rule-based safety filters; fallback to baseline MARL
Transparency & Accountability	Opaque or misleading LLM explanations	Loss of trust, audit failures	Final authority with human supervisor
Data & Privacy	Sensitive victim/location data exposure	Privacy violations	Anonymization, secure storage, restricted access
Bias & Fairness	Biased or discriminatory outputs from LLM	Unequal prioritization of rescue zones	Bias audits; neutral language policies
Human-in-the-Loop Risks	Cognitive overload or automation bias	Poor supervisory decisions	Adjustable detail levels; human override training
Regulatory Compliance	Misalignment with AI governance (e.g., EU AI Act)	Legal and ethical breaches	Align with standards; IRB/ethics board approvals

8. Project Timeline and Implementation Strategy

The project implementation is proposed to be done in *Erebus Rescue Simulation* environment. The IDE is selected due to its deployment flexibility among various operating systems and lightweight feature. The program controllers for Layer 1 & 2 will be written in *Python* programming language, there is no need of external *training data,* the real-time data from simulation platform will be employed. *EPUCK* robot is found to be a best option for this task, a basic simulation world will be used for baseline. This setup is already created in selected IDE (see Figure 1) with multiple hazards (colored areas) and obstacles. Once controllers are successfully trained and tested, the trained models will be cross evaluated in real-time rescue simulation environment. The project timeline will be accomplished till system deployment, in modular manners for each pipeline step. Early finishing is aimed to give an extra leverage for system fine-tuning and cross validation procedures. The dynamic feature of trained robot will be evaluated in creating several real-time test scenarios, and creating different levels of hazards along with exception cases.

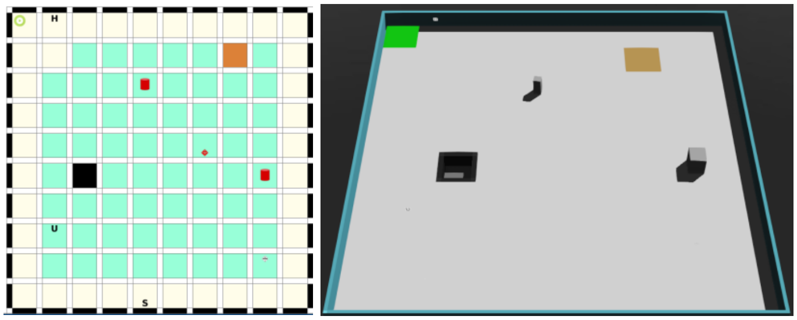

Figure 2 Erebus Rescue Simulation Baseline World

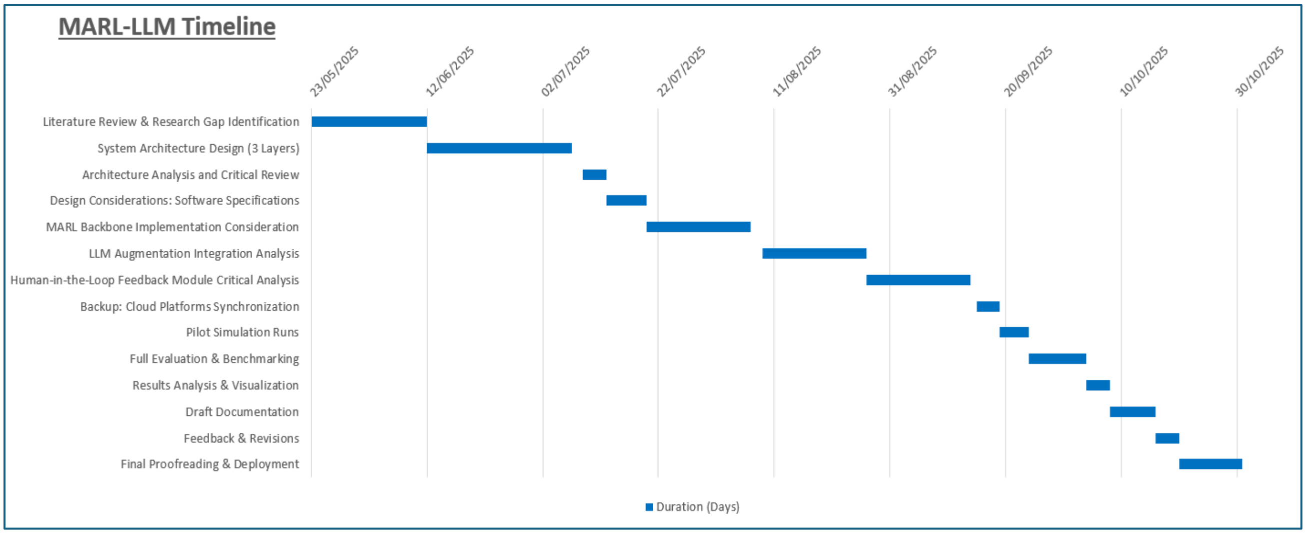

Figure 3 Project Timeline: Gantt Chart

9. Conclusive Remarks

This research study provides a critical evaluation of the key limitations that pose significant challenges in MARL-based robotic systems in SSRO environments. The evaluated aspects covers: 1) communication opacity, 2) rigid agent behavior, and 3) lack of human accessibility in multi-agent scenarios. The study propose a novel LLM−MARL framework by exploring the transformative potential of large-scale language models in enhancing multi-agent collaboration through *semantic reasoning* and *role diversity*. The research findings makes a strong baseline for implementing the first-of-its-kind LLM-enabled architecture in MARL settings.

- **Research Gap:** Existing work lacks semantic communication, real-time HITL adaptability, and behavior flexibility.
- **Aim:** We aim at developing LLM-augmented framework that supports natural language-based prompt-driven communication, and real-time feedback feature from human supervisors.
- **Methodology:** We aim to integrate lightweight LLMs with multiple MARL agents to effectively generate useful communication prompts, delivering prompts to other agents with dynamic and role-based behaviors, and to process natural language commands for mission updates in real-time in *Erebus disaster simulation platform*.

10. References

Achiam, J., Adler, S., Agarwal, S., & Ahmad, L. a. (2023). *GPT-4 Technical Report.* arXiv preprint arXiv:2303.08774.

Ahn, M., Brohan, A., Brown, N., Chebotar, Y., Cortes, O., David, B., . . . Hausman, K. a. (2022). Do As I Can, Not As I Say: Grounding Language in Robotic Affordances. *arXiv preprint arXiv:2204.01691*.

Brandizzi, N. (2023). Toward more human-like ai communication: A review of emergent communication research. *IEEE Access, 11*, 142317--142340.

Cai, Y., He, X., Guo, H., Yau, W.-Y., & Lv, C. (2024). *Transformer-based Multi-Agent Reinforcement Learning for Generalization of Heterogeneous Multi-Robot Cooperation.* IEEE.

Das, A., Gervet, T., Romoff, J., Batra, D., Parikh, D., Rabbat, M., & Pineau, J. (2019). Tarmac: Targeted multi-agent communication. *PMLR*, (pp. 1538--1546).

Dessi, R., Kharitonov, E., & Marco, B. (2021). Interpretable agent communication from scratch (with a generic visual processor emerging on the side). *Advances in Neural Information Processing Systems, 34*.

Dinneweth, J., Boubezoul, Mandiau, A., Espie, R. a., & Stephane. (2022). Multi-agent reinforcement learning for autonomous vehicles: A survey. *Autonomous Intelligent Systems*, 27.

Du, W., & Ding, S. (2021). A survey on multi-agent deep reinforcement learning: from the perspective of challenges and applications. *Artificial Intelligence Review*, 3215--3238.

Du, Y. H., Fang, M., Liu, J., Dai, T., & Tao, D. (2019). Liir: Learning individual intrinsic reward in multi-agent reinforcement learning. *Advances in neural information processing systems, 32*.

Foerster, J., Assael, I. A., De Freitas, N., & Whiteson, S. (2016). Learning to communicate with deep multi-agent reinforcement learning. *Advances in neural information processing systems, 29*.

Foerster, J., Farquhar, G., Afouras, T., & Nardelli, N. a. (2018). Counterfactual Multi-Agent Policy Gradients. *Proceedings of the AAAI conference on artificial intelligenc, 1*, 32.

Jaques, N., Lazaridou, A., Hughes, E., Gulcehre, C., Ortega, P., Strouse, D., . . . De Freitas, N. (2019). Social influence as intrinsic motivation for multi-agent deep reinforcement learning., (pp. 3040--3049).

Javaheripi, M. a., Abdin, S. a., & Marah and Aneja, J. a. (2023). Phi-2: The surprising power of small language models. *Microsoft Research Blog, 1*, 3.

Jiang, J., & Lu, Z. (2018). Learning attentional communication for multi-agent cooperation. *Advances in neural information processing systems, 31*.

Kaelbling, L. P., Littman, M. L., & Moore, A. W. (1996). Reinforcement learning: A survey. *Journal of artificial intelligence research, 4*, 237--285.

Kernycky, A., Coleman, D., Spence, C., & Das, U. (2024). International Conference on Human-Computer Interaction. *Springer*, 75--85.

Kumar, P. (2024). Large language models (LLMs): survey, technical frameworks, and future challenges. *Springer*.

Li, Y. (2017). Deep reinforcement learning: An overview. *arXiv preprint arXiv:1701.07274*.

Liang, J., Huang, W., Xia, F., Xu, P., Hausman, K., Ichter, B., . . . Zeng, A. (2023). Code as policies: Language model programs for embodied control. *IEEE*, 9493--9500.

Liu, Z., Zhao, C., Iandola, F., Lai, C., Tian, Y., Fedorov, I., . . . Krishnamoorthi, R. a. (2024). *Mobilellm: Optimizing sub-billion parameter language models for on-device use cases.*

Lowe, R., Wu, Y. I., Tamar, A., Harb, J., & Pieter Abbeel, O. a. (2017). Multi-Agent Actor-Critic for Mixed Cooperative-Competitive Environments. *Advances in neural information processing systems*, 30.

Marinescu, A., Dusparic, I., & Clarke, S. (2017). Prediction-based multi-agent reinforcement learning in inherently non-stationary environments. *ACM Transactions on Autonomous and Adaptive Systems (TAAS), 12*, 1--23.

Mordatch, I., & Abbeel, P. (2018). Emergence of grounded compositional language in multi-agent populations. *1*.

Panait, L., & Luke, S. (2005). Autonomous agents and multi-agent systems. *Springer, 11*, 387--434.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., . . . Clark, J. a. (2021). Learning transferable visual models from natural language supervision. PmLR.

Rashid, T., Samvelyan, M., De Witt, C. S., Farquhar, G., Foerster, J., & Whiteson, S. (2020). Monotonic value function factorisation for deep multi-agent reinforcement learning. *Journal of Machine Learning Research, 21*, 1--51.

Sheng, J., Wang, X., Jin, B., Yan, J., Li, W., Chang, T.-H., . . . Zha, H. (2022). Autonomous Agents and Multi-Agent Systems. *Springer, 36*, 50.

Sukhbaatar, S., Fergus, R., & others. (2016). Learning multiagent communication with backpropagation. *Advances in neural information processing systems*, 29.

Sun, C., Huang, S., & Pompili, D. (2024). Llm-based multi-agent reinforcement learning: Current and future directions. *arXiv preprint arXiv:2405.11106*.

Tang, C., Abbatematteo, B., Hu, J., Chandra, R., & Martin-Martin, R. a. (2025). Deep reinforcement learning for robotics: A survey of real-world successes}. *Proceedings of the AAAI Conference on Artificial Intelligence, 39*, 28694--28698.

Tom, B., Mann, B., Ryder, N., & Subbiah, M. a. (2020). *Language models are few-shot learners.* Advances in neural information processing systems.

Touvron, H. a., Stone, L. a., Albert, K. a., & Peter and Almahairi, A. a. (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., . . . Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*.

Wei, D., & Ding, S. (2021). A survey on multi-agent deep reinforcement learning: from the perspective of challenges and applications. *Artificial Intelligence Review, 54*, 3215-3238.

Xu, M., & Shen, Y. (2022). PMLR.

Xu, M., Shen, Y., Zhang, S., Lu, Y., Zhao, D., Tenenbaum, J., & Gan, C. (2022). *Prompting Decision Transformers for Few-Shot Policy Generalization.* PMLR.

Yang, Y., Zhou, T., Li, K., Tao, D., Li, L., Shen, L., . . . Shi, Y. (2024). Embodied multi-modal agent trained by an llm from a parallel textworld. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 26275--26285.

Yu, C. a., Vinitsky, A. a., Gao, E. a., Wang, J. a., Bayen, Y. a., & Alexandre and Wu, Y. (2022). The surprising effectiveness of ppo in cooperative multi-agent games. *Advances in neural information processing systems, 35*, 24611--24624.

Zhang, K., Yang, Z., & Bacsar, T. (2021). Handbook of reinforcement learning and control. *Springer*, 321--384.

Zhang, P., Zeng, G., Wang, T., & Lu, W. (2024). Tinyllama: An open-source small language model. *arXiv preprint arXiv:2401.02385.*