LUCIFER: Language Understanding and Context-Infused Framework for Exploration and Behavior Refinement

Dimitris Panagopoulos¹, Adolfo Perrusquía¹ and Weisi Guo¹

Abstract—In dynamic environments, the rapid obsolescence of pre-existing environmental knowledge creates a gap between an agent's internal model and the evolving reality of its operational context. This disparity between prior and updated environmental valuations fundamentally limits the effectiveness of autonomous decision-making. To bridge this gap, the contextual bias of human domain stakeholders, who naturally accumulate insights through direct, real-time observation, becomes indispensable. However, translating their nuanced, and context-rich input into actionable intelligence for autonomous systems remains an open challenge. To address this, we propose LUCIFER (Language Understanding and Context-Infused Framework for Exploration and Behavior Refinement), a domain-agnostic framework that integrates a hierarchical decision-making architecture with reinforcement learning (RL) and large language models (LLMs) into a unified system. This architecture mirrors how humans decompose complex tasks, enabling a high-level planner to coordinate specialised sub-agents, each focused on distinct objectives and temporally interdependent actions. Unlike traditional applications where LLMs are limited to single role, LUCIFER integrates them in two synergistic roles: as context extractors, structuring verbal stakeholder input into domain-aware representations that influence decision-making through an attention space mechanism aligning LLM-derived insights with the agent's learning process, and as zero-shot exploration facilitators guiding the agent's action selection process during exploration. We benchmark various LLMs in both roles and demonstrate that LUCIFER improves exploration efficiency and decision quality, outperforming flat, goalconditioned policies. Our findings show the potential of contextdriven decision-making, where autonomous systems leverage human contextual knowledge for operational success.

Index Terms—Hierarchical Decision-Making, Large Language Models, Context-Aware Agents, Human-AI Collaboration

I. INTRODUCTION

A. Background & Motivation

ODERN autonomous systems are increasingly deployed in dynamic and high-stakes settings, ranging from industrial inspection to rescue efforts in the aftermath of man-made or natural disasters [1]–[4]. In these complex operational contexts, information about key environmental elements (e.g., navigable routes, safe operating zones, or key inspection points) can become outdated, misleading both human teams and autonomous systems. As conditions evolve over time, prior knowledge can deviate from the objective

¹Faculty of Engineering and Applied Science, Cranfield University, Cranfield MK43 0AL, UK, {d.panagopoulos, adolfo.perrusquia-guzman, weisi.guo}@cranfield.ac.uk. This work is funded by EPSRC iCASE with Thales UK (EP/X52475X/1)

reality on the ground. This rapid obsolescence introduces knowledge gaps between an agent's (i.e., human or robot) contemporary understanding of the world and the updated valuation of environmental features moments later, leading to poor situational awareness [5], [6].

Human stakeholders operating in these environments [7], [8], whether technical experts or individuals on-site, naturally gather situation-specific clues through direct observation. This continuous exposure offers localised bias that can plug knowledge gaps left by obsolete environmental information [9]–[11]. However, their verbalised input, often rich and composite in nature, is rarely integrated into planning or decision-making pipelines. Although current solutions operate under a humanon-the-loop paradigm, where a centralised command center supervises robotic decisions, they fall short in leveraging the real-time verbalised input from on-site stakeholders [12]. This reflects a tendency in autonomous systems design to neglect meaningful human-in-the-loop engagement in highstakes environments, even though methodologies already allow for human influence in the decision-making process [13]. Bridging this communication barrier, so that agents can effectively integrate multifaceted language-based descriptions into a structured learning mechanism, remains a critical challenge, highlighting the need for adaptive decision-making frameworks capable of leveraging real-time insights.

Reinforcement learning (RL) offers a promising approach to enable agents to iteratively refine their policies based on environmental feedback [14]. However, conventional RL approaches, which rely on conventional flat policies operating over entire state spaces, struggle to handle the complexity of environments with distant or interdependent goals. In goal-conditioned RL, for instance, accurately estimating state values for distant goals remains challenging, necessitating a structured approach that prioritises reaching intermediate objectives first. Developing a framework that embraces a hierarchical structure becomes essential, as it enables agents to plan policies over manageable subsets of the state space [15]. Through task decomposition, which mirrors human cognitive resource allocation [16], agents can better coordinate multiple interdependent tasks [17], where decisions in one task influence the success of others. This hierarchical approach, when enhanced with contextual insights from stakeholders, forms the foundation of our proposed solution.

In this article, we propose LUCIFER (Language Understanding and Context-Infused Framework for Exploration and Behavior Refinement) to address these

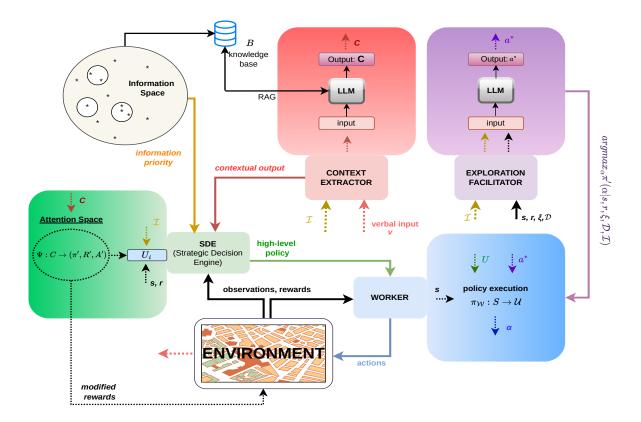


Fig. 1. Overview of LUCIFER's workflow. The Strategic Decision Engine (SDE) assigns tasks based on environmental observations and Information Space, while Worker agents interact with the Environment by executing primitive actions. The Context Extractor processes verbal inputs into structured representations, which refine decision-making via the Attention Space. The Exploration Facilitator leverages past interactions and learning status to guide exploration by making zero-shot predictions.

challenges. First, LUCIFER systematically integrates contextual knowledge by leveraging LLMs as extractors. This component transforms linguistic inputs from ground-level stakeholders into structured, actionable insights. These insights are then embedded into the agent's decisionmaking process via an attention space mechanism to ensure that agent behaviors are grounded in up-to-date contextual knowledge rather than obsolete priors. Second, to effectively handle temporal coupling and long-horizon task complexity, LUCIFER adopts a hierarchical policy structure operating at multiple temporal scales. This approach, mirroring how humans decompose tasks to manage limited cognitive resources, enables specialised agents to develop task-specific behaviors and operate across distinct tasks, while accounting for temporal coupling, where one decision can impact the success of another. Finally, recognising the challenges of randomly exploring such environments, LUCIFER employs LLMs in a complementary role as exploration facilitators. Here, we exploit the advanced reasoning capabilities of LLMs to make predictions during the action selection process of specialised agents and reduce reliance on traditional RL-driven trial-and-error exploration.

Although originally inspired by search and rescue (SAR) scenarios, where timely updates about hazardous zones or trapped survivors are critical, the framework is readily adaptable to a broad range of domains (e.g., manufacturing lines,

agricultural robotics, assistive robotics, environmental monitoring). The underlying principles of dynamic contextual knowledge integration and hierarchical coordination equally apply to other fields where autonomy and human expertise intersect. As an illustrative example, we demonstrate how LUCIFER supports multi-objective missions in a simulated SAR environment, highlighting notable gains in exploration efficiency and overall decision quality. The proposed framework is motivated by the analysis of UAV-supported SAR operations [18] showing that effective decisions rely on continuously updated knowledge maps that incorporate live field data from both autonomous agents and human stakeholders. Bearing that in mind, we emphasise the need for systems capable of continuous information integration, processing, and adaptation to ground truths in order to improve real-time, context-driven decision-making.

B. Contributions

In summary, this article makes the following contributions:

 We introduce LUCIFER, a scalable, hierarchical domain-agnostic framework that enables the systematic fusion of situational knowledge into autonomous decision-making systems. This addresses a fundamental challenge particularly in human-AI collaboration by providing a generalisable approach to leveraging domainrelevant cues in dynamic environments.

- 2) We demonstrate the dual functionality of off-the-shelf LLMs highlighting their versatility in emulating humanlike extraction process and exhibiting zero-shot, trainingfree predictive capabilities.
- 3) We develop an attention space mechanism that bridges the gap between LLM-processed human insights with RL, providing a structured approach to representing contextual bias integration.
- 4) We compare different LLM variants using a SAR environment as a testbed and demonstrate that LUCIFER outperforms conventional single-policy baselines. Our results provide valuable insights into the practical implementation of LLM-assisted intelligent systems.

Algorithm 1 LLM as Context Extractor

```
Require: verbal input V, knowledge base B, Information Space \mathcal I
```

```
Ensure: spatially insights C
 1: Initialize LLM with knowledge base B: L_B: V \to C
 2: Set C \leftarrow \emptyset
 3: for all v_i \in V do
       Extract key entities E_i from v_i using L_B
 4:
       for all e_i \in E_i do
 5:
          Infer contextual meaning of e_j based on \mathcal{I}
 6:
          Assign category label cat_j using \mathcal{I} priorities
 7:
          Structure contextual representation: c_i = (e_i, cat_i)
 8:
 9:
          Add c_i to C
```

10: end for11: end for

12: return C

Algorithm 2 LLM as Exploration Facilitator

Require: state s_t , trajectory buffer ξ , memory buffer \mathcal{D} , action space A

Ensure: selected action a^*

- 1: Encode inputs: s_t , ξ , and $\mathcal D$ into prompt format
- 2: Query LLM with the encoded prompt
- 3: Receive action distribution $P(a \mid s_t, \xi, \mathcal{D})$
- 4: Select action $a^* = \arg \max_{a \in A} P(a \mid s_t, \xi, \mathcal{D})$
- 5: **return** a^*

II. RELATED WORK

The integration of LLMs into computational systems has opened important directions in current AI research, especially since the advert of ChatGPT and similar large-scale models era. This section reviews key paradigms of LLM combined with learning schemes and their role in environmental information processing. We build upon the comprehensive taxonomies and analyses established in recent survey works [19]–[22].

A. Integration Paradigms of LLMs and RL

Following Cao et al.'s [21] taxonomy, LLMs can serve multiple roles in RL systems. One prominent role is as reward designers, where they have the potential to design or shape reward functions. For instance, [23] and [24] utilise LLMs to generate nuanced reward signals aligned with human

preferences. Similarly, based on the characteristics of the environment, task, or agent's behavior, existing works leverage LLMs to design reward functions [25]–[30]. Extending this, [31] employs LLMs to empower credit assignment, bridging a critical challenge in episodic RL tasks.

3

Due to the credit assignment problem, LLMs can break down complex tasks into structured sub-tasks, acting as high-level planners [32]–[34]. For example [35], [36] demonstrate how LLMs can dynamically select appropriate sub-goals during task execution. Recent works such as [37] emphasise using LLMs to generate intrinsic goals to promote open-ended exploration for RL agents. Similarly, the work in [38] highlights interactive planning systems where LLMs generate, explain, and prioritise sub-goals to allow RL agents to handle diverse tasks seamlessly. In urban driving settings, [39] illustrates a novel collaborative automated policy training workflow, which consists of multiple LLM agents for curriculum RL.

A fundamental distinction established by Luketina et al. [19] is between language-conditional and language-assisted RL. In the former scheme agents directly interact with the environment through language instructions [40], [41]. Conversely, in the latter, language serves as a medium of communicating domain knowledge without being integral to the core task [42]–[44]. Our work aligns with the language-assisted RL paradigm. Here, language is used to convey broader descriptive information about the environment's structure, properties, and dynamics, rather than just surface-level instructions about what actions an agent should take.

B. Environmental Information Processing

Another crucial role of LLMs when integrated with RL systems, highlighted by Cao et al. [21], is serving as information processors. This bridges the gap between descriptive environmental information and structured agent inputs through feature representation extraction or language translation. In feature representation extraction, it is shown how frozen or fine-tuned LLMs can extract meaningful representations from environment observations [45]-[48]. In language translation, environmental and task instruction information is grounded into formal task-specific language to reduce learning complexity [49]–[51]. Recent studies also demonstrate the use of generative models for direct entity and relation extraction from natural language inputs. Models like REBEL [52] and InstructUIE [53] reformulate extraction as a generative task, enabling models to identify key elements and classify them into relevant categories. These works, however, typically operate in isolation from downstream control systems, and are rarely embedded into interactive agents that act upon the extracted knowledge. In contrast, LLMs have demonstrated potential in translating abstract ethical principles into actionable behaviors such as avoiding socially sensitive areas during autonomous navigation [54]. LUCIFER extends this direction by grounding context into spatially structured constraints that adapts agent behavior in real time. To the best of our knowledge, no existing studies have explored the direct integration of LLMdriven entity classification into a hierarchical control loop, where language serves as an input channel for contextual

4

information and the extracted knowledge continuously shapes agent behavior.

Our proposed framework aligns naturally with the taxonomy proposed in [55], which categorises informed machine learning approaches along three dimensions: knowledge source, knowledge representation, and knowledge integration. Knowledge source in LUCIFER originates from human stakeholders who possess domain-relevant contextual biases over the operational environment. Knowledge representation is encoded through LLMs, further refined via a Retrieval-Augmented Generation (RAG) pipeline [56], that turns linguistic verbal inputs into actionable, spatially-relevant insights. Finally, knowledge integration is facilitated by an attention space mechanism that supports decision-making and guides exploration to address the challenges typically associated with exploratory processes.

III. METHODOLOGY

LUCIFER extends the framework introduced in [57], by introducing a novel dual-role for LLMs enabling both context extraction and exploration guidance and enhancing the attention-space mechanism to shape policy, reward, and action space. Additionally, it includes a comprehensive benchmarking of multiple LLMs. These advancements transform the original framework into a domain-agnostic and scalable system capable of generalising across diverse operational domains that require structured task execution and real-time adaptation, while offering a more robust and detailed evaluation of its core components (see Fig. 1).

A. Markov Decision Process Configuration

The foundation of our proposed framework, LUCIFER, is a Markov Decision Process (MDP) designed to model the hierarchical decision-making problem. We define the MDP as a tuple $\langle S, A, T, R, \gamma \rangle$. Here, S is a finite set of states and A is a finite set of primitive actions, $T: S \times A \times S \to [0,1]$ is the state transition probability function, $R: S \times A \to \mathbb{R}$ is the reward function, $\gamma \in [0,1]$ is the discount factor balancing immediate and future rewards.

B. Hierarchical Task Decomposition

To address the complexity of long-horizon tasks, the framework employs a hierarchical structure that decomposes complex decision-making problems into simpler, distinct subproblems [15]. In our hierarchical two-layer structure, a high-level planner assigns structured tasks (i.e., goals), while specialised workers produce primitive actions within the environment to execute them. This ensures effective policy switching between intermediate goals that must be achieved en route to the final goal.

Formally, let $\mathcal{U}=\{U_1,U_2,\ldots,U_n\}$ represent the set of tasks, where each task U_i is delegated to a worker $w_i\in\mathcal{W}$ responsible for executing it. Specifically, at the high-level, a Strategic Decision Engine (SDE) produces tasks $U_i\in\mathcal{U}$ that specify desired changes in state observations. The SDE selects tasks by either sampling from its policy $\pi_{\mathcal{U}}:S\to\mathcal{U}$ or using a pre-defined task transition process. At the low-level,

a Worker module is activated, where each worker operates under a policy $\pi_{\mathcal{W}}: S \to A$ tailored to its assigned task U_i . Task transitions are governed by a termination condition $\beta: S \to \{0,1\}$, defined for each task U_i . Specifically, $\beta(s)=1$ if the current state $s \in S$ satisfies the completion criteria for the active task U_i (e.g., completing a target state or sub-goal), and $\beta(s)=0$ otherwise. This ensures that a worker continues executing actions under $\pi_{\mathcal{W}}$ until the intermediate goal is met, at which point the SDE advances to the next task U_{i+1} . This process repeats until the final goal is achieved, aligning the hierarchical execution with the underlying MDP dynamics.

C. Information Space

The Information Space serves as a reference map to the SDE, ensuring that the agent operates within a structured, mission-relevant knowledge domain, particularly in environments where contextual information is critical. It is defined as a collection $\mathcal{I} = \{I_1, I_2, \dots, I_m\}$, where each I_j represents a distinct category, type, or unit of information pertinent to the agent's operational objectives. These elements such as mission priorities, constraints, and safety considerations define the scope of meaningful information for the system, thus guiding its perception, reasoning, and actions.

D. LLM as Context Extractors

The Context Extractor component leverages LLMs to convert verbal inputs into structured representations that can be directly utilised within the decision-making pipeline. This transformation is essential, as raw linguistic inputs often lack the structured format required by autonomous systems. The LLM processes these inputs, extracts key entities, and maps them to predefined categories within the Information Space \mathcal{I} . The resulting structured output, denoted as C, informs and shapes the agent's situational understanding (see Section III-F).

Formally, given a verbal input space $V=\{v_1,\ldots,v_m\}$, the LLM acts as a transformation function $L_B:V\to C$ enhanced by a RAG pipeline. This pipeline augments the LLM's domain-specific expertise through integration with a knowledge base B, enabling context-aware interpretation without the need for extensive retraining or fine-tuning. For each input $v_i\in V$, the LLM identifies a set of relevant entities E_i and infers their contextual meaning based on the structured defined in Information Space \mathcal{I} . Each extracted entity e_j is then classified into a category cat_j according to \mathcal{I} . This process, outlined in Algorithm 1, ensures that the agent receives precise, spatially-grounded insights $C=\{c_1,\ldots,c_n\}$ aligned with its operational objectives and optimised for downstream decision-making.

E. LLM as Exploration Facilitator

The Exploration Facilitator uses LLMs to improve exploration by predicting plausible actions at decision points, leveraging both short-term learning behavior and long-term experience. Rather than relying solely on trial-and-error learning, this approach supplements conventional action selection

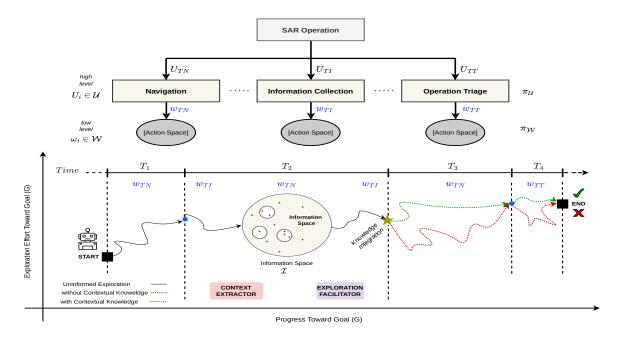


Fig. 2. Illustration of LUCIFER in a post-earthquake SAR Case Study: The high-level planner decomposes the mission into three sub-tasks—navigation U_{TN} , information collection U_{TI} , and triage U_{TT} —executed by specialised workers. At T_2 , worker w_{TN} navigates to key locations, while w_{TI} gathers critical data, dictated by the Information Space \mathcal{I} . An LLM-based Context Extractor processes human inputs to provide structured insights (\star) , and a separate LLM-based Exploration Facilitator guides w_{TI} 's action selection process. These components reduce the agent's reliance on inefficient, high-effort exploratory trajectories (red) and instead enable optimised decision-making pathways (green), enhancing progress toward mission goals.

strategies, such as ε -greedy exploration, with informed, zero-shot predictions.

Formally, the agent maintains two types of buffers. The trajectory buffer ξ captures short-term, episodic behavior by storing transitions of the form $\{(s_t, a_t, s_{t+1}, r_t)\}$ within a single episode. This buffer is reset at the start of each new episode and provides the LLM with a local view of recent agent behavior and outcomes. In contrast, the memory buffer \mathcal{D} is a persistent, cross-episodic record that accumulates structured summaries of past experiences across tasks. It captures past interactions at semantically meaningful locations (e.g., visited regions, task-relevant states), along with associated action attempts and their success or failure. Both buffers are maintained and encoded as structured, text-based representations. This design choice let the LLM leverage its reasoning capabilities as it is enabled to better reason over the agent's cumulative experience and reuse historical knowledge across episodes. At each exploratory decision point, the agent queries the LLM using the current state s_t , the recent trajectory ξ , and the long-term memory \mathcal{D} . The LLM returns a predicted action with the highest probability $a^* = \arg \max_{a \in A} P(a \mid s_t, \xi, \mathcal{D})$. This zero-shot, training-free predictive mechanism, shown in Algorithm 2, reframes exploration as a structured reasoning task, where the LLM fuses immediate context with longterm patterns to guide the agent toward informative and goalaligned behaviors.

F. Attention Space

The attention space mechanism—distinct fundamentally from transformer-based self-attention that computes token

```
Algorithm 3 Attention Space Mechanism for Policy Shaping
Require: spatially insights C, learned table Q(s, a), Attention
    Space \Psi(s, a), critical states \mathcal{S}_{crit}
Ensure: updated Q'(s, a)
 1: Set \Psi(s,a) \leftarrow 0, Q'(s,a) \leftarrow Q(s,a), \forall (s,a)
 2: for all c_i \in C do
       Extract relevant state s_i and category cat_i from c_i
       for all a \in A(s) leading to s_i do
 4:
          Compute \Psi(s, a) using Eq. (2)
 5:
          Overwrite Q'(s, a) \leftarrow \Psi(s, a)
 6:
 7:
       end for
 8: end for
 9: return Q'(s,a)
```

```
Algorithm 4 Attention Space Mechanism for Reward Shaping
Require: reward function R(s, a), spatially insights C, dis-
    count factor \gamma, potential function \Phi(s)
Ensure: final shaped reward function R''(s, a)
 1: Initialise \Phi_{\Psi}(s) \leftarrow 0, \forall s
 2: for all c_i \in C do
 3:
       Extract relevant state s_i and category cat_i from c_i
 4:
       Update \Phi_{\Psi}(s_i) using Eq. (6)
       for all (s, a, s') where s' = s_i do
 5:
         Compute F(s, s') using Eq. (4)
 6:
         Compute R'(s, a) using Eq. (5)
 7:
         Compute R''(s, a) using Eq. (7)
 8:
 9:
       end for
10: end for
11: return R''(s,a)
```

Algorithm 5 Attention Space Mechanism for Action Space Adjustment

Require: action space A(s), spatially insights C

Ensure: updated action space A'(s)

- 1: Initialise $A'(s) \leftarrow A(s)$
- 2: for all $c_i \in C$ do
- 3: Extract relevant state s_j and category cat_j from c_j
- 4: Update A'(s) using Eq. (8)
- 5: end for
- 6: **return** A'(s)

relationships—serves as an intermediary between the structured output generated by the LLM (Section III-D) and downstream agents in the hierarchical framework. Unlike token-level attention in language models, this mechanism dynamically refines the agent's decision-making process by embedding contextual structured representation C into the learning process. Unlike arbitrary biasing, this influence is a form of heuristic refinement that leverages environmental structure, aligning with the notion of ecological rationality [58]. To do so, it influences key core components, including the policy (π) , reward function (R), and action space (A), as illustrated in Fig. 3. Formally, this transformation is defined as:

$$\Psi: C \to (\pi', R', A') \tag{1}$$

where C denotes the structured contextual insights (e.g., identified constraints, operational priorities, or task-relevant objectives), π' represents a context-aware policy that prioritises relevant actions through modified state-action mappings, R' is an adapted reward function that introduces contextual preferences, and A' reflects a modified action space that promotes or restricts certain behaviors based on situational relevance.

- 1) Policy Refinement/Shaping: Policy shaping adjusts the agent's decision-making by biasing Q-values of state-action pairs linked to contextually critical states. Let $\mathcal{S}_{crit} = \{s_u, s_d, s_o\}$ represent a set of generalised state categories derived from C:
 - s_u: Undesirable states (e.g., states violating constraints or safety thresholds)
 - s_d : **Desirable states** (e.g., states that fulfill intermediate objectives)
 - s_o: Critical objective states (e.g., states achieving mission-critical goals)

The attention space introduces biases into the Q-values as follows:

$$\Psi(s,a) = \begin{cases}
-\lambda_u, & \forall a \text{ leading to } s_u \\
\lambda_d, & \forall a \text{ leading to } s_d \\
\lambda_o, & \forall a \text{ leading to } s_o
\end{cases}$$
(2)

where λ_u , λ_d , $\lambda_o \in \mathbb{R}$ are scalar parameters that penalise or incentivise transitions towards s_u , s_d , and s_o , respectively. The modified Q-values induce a refined policy π' , where action preferences are systematically guided by contextual relevance through $\Psi(s,a)$.

2) Reward Shaping: Reward shaping enhances learning efficiency by modifying the reward signal using contextual and spatial information. The final shaped reward function R''(s,a) combines potential-based reward shaping (PBRS), a theoretically sound approach that preserves the optimal policy of the original MDP, with immediate context-sensitive reward adjustments. The potential function $\Phi(s)$ encodes spatial preferences over states and is adaptable based on the environment's operational goals. It can be formulated to either encourage proximity to desirable states (e.g., goals, points-of-interest) or discourage proximity to undesirable states (e.g., hazards, unsafe zones). It is defined as:

$$\Phi(s) = f\left(\min_{s_r \in S_{\text{ref}}} \operatorname{dist}(s, s_r)\right) \tag{3}$$

where $S_{\rm ref} \subset \mathcal{S}_{crit}$, and ${\rm dist}(s,s_r)$ is a domain-relevant distance metric (e.g., L1, L2 norm, etc.). The transformation function $f(\cdot)$ maps this distance into a scalar potential depending on whether proximity is rewarded or penalised. The potential-based shaping term F(s,s') is then calculated as the discounted difference in potential between the next state s' and the current state s:

$$F(s, s') = \gamma \Phi(s') - \Phi(s) \tag{4}$$

This term is added to the original reward to produce the shaped reward:

$$R'(s,a) = R(s,a) + F(s,s')$$
 (5)

To incorporate real-time contextual updates from the LLM, an additional immediate shaping signal is applied based on the semantic category of the current state:

$$\Phi_{\Psi}(s) = \begin{cases}
-\beta_u, & \text{if } s = s_u \\
\beta_d, & \text{if } s = s_d \\
\beta_o, & \text{if } s = s_o
\end{cases}$$
(6)

where β_u , β_d , $\beta_o \in \mathbb{R}$ are parameters that modulate penalties or incentives based on semantic state relevance.

The final shaped reward function combines both spatial and semantic shaping:

$$R''(s,a) = R'(s,a) + \Phi_{\Psi}(s) \tag{7}$$

3) Action Space Shaping/Biasing: Beyond policy and reward shaping, the attention space mechanism allows for dynamic toggling of the agent's action space A(s) based on contextual insights. Formally, the adjusted action space is defined as:

$$A'(s) = \begin{cases} A(s) \setminus \{a \mid a \to s_u\}, & \text{if } s_u \text{ detected} \\ \{a \mid a \to s_d, s_o\}, & \text{if } s_d \text{ or } s_o \text{ prioritised} \\ A(s) \cup \{a \mid a \to s_d, s_o\}, & \text{if expanding exploration} \\ A(s), & \text{otherwise} \end{cases}$$
(8)

First, if an undesirable state s_u is present, any action that leads the agent to it is removed from its available action space. Second, in some cases, depending on the problem settings, the

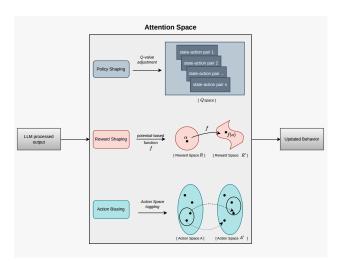


Fig. 3. Illustration of the attention space mechanism. The LLM-processed output (i.e., structured into \mathcal{C}) guides policy shaping, reward shaping, and action biasing. By embedding contextual insights into these key components, the agent's behavior is adaptively refined to reflect real-time environmental knowledge.

system designer may choose to further restrict the action space by keeping only the action that leads to desired states s_d or s_o . Third, the system designer, alternatively, can expand the action space by adding new actions that move the agent toward desired states s_d or s_o . For instance, in exploration phases, actions leading to s_d or s_o might be added to encourage systematic probing of under-explored regions. If none of these conditions apply, the agent's action space remains unchanged.

This adjustment effectively transforms the underlying MDP and necessitates consistent application across all aspects of learning. Crucially, the computation of TD targets must also respect this modified action space. For off-policy Q-learning, the standard update:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha [r_{t+1} + \gamma \max_{a' \in A} Q(s_{t+1}, a') - Q(s_t, a_t)]$$

Must be formulated to consider only valid actions in the maximisation operation:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha [r_{t+1} + \gamma \max_{a' \in A'} Q(s_{t+1}, a') - Q(s_t, a_t)]$$
(10)

Similarly, on-policy methods like SARSA must select a' from A' rather than A. Inconsistent application, where actions are restricted during selection but not during target computation, creates fundamental learning conflicts that prevent convergence, as the agent would simultaneously pursue different optimal policies during action selection versus value updating.

IV. EXPERIMENTAL SETUP

A. Problem Setting

To validate the adaptability of our proposed framework, we consider a post-earthquake urban SAR scenario. The mission involves locating and assisting potential victims, while continuously gathering information about environmental hazards

("HAZ") and safe zones ("POI"). Following the hierarchical decomposition introduced in Section III-B, the mission is divided into three specialised sub-tasks. Handled by worker w_{TN} , the navigation sub-task U_{TN} involves moving the agent to designated location using four directional actions. Carried out by worker w_{TI} , the information collection sub-task U_{TI} focuses on collecting critical situational data. This data is delivered directly via verbal inputs from human stakeholders, necessitating robust interpretation of their inputs. Executed by worker w_{TT} , the operational sub-task U_{TT} encompasses rescue activities, debris removal, and triage procedures. A high-level planner orchestrates the sequence of these subtasks, executed in an order that respects domain-specific dependencies captured by the Information Space \mathcal{I} . For example, the triage sub-task U_{TT} is deferred until sufficient environmental information is gathered, ensuring safety and operational readiness. Initially, the planner assigns w_{TN} to navigate toward designated collection points. Once reached, w_{TI} is activated to collect relevant information types. This cycle may repeat multiple times across different locations until the planner determines that conditions are met (dictated again by \mathcal{I}) to move and initiate the final triage phase. This process is illustrated in Fig. 2.

B. Simulated Environment

We employ a 2D gridworld environment implemented with OpenAI Gym to simulate a disaster area. Each grid cell may represent an empty space, obstacle, information point, hazard, safe zone, or victim location (final location). This setup (Fig. 4) allows for rapid training and fine-grained evaluation of LUCIFER's core modules, including hierarchical decisionmaking, LLM-guided reasoning, and attention-based shaping. Additionally, we develop a 3D ROS2 Gazebo simulation as a proof-of-concept to demonstrate real-world applicability. While no quantitative results are reported from this environment, it showcases the framework's potential for realistic robotic deployment. We assess environmental difficulty along two key dimensions. First, we define two levels of information complexity: **3-info**, where the agent must collect three types of information, and 6-info, where six distinct information types must be collected before initiating rescue efforts. Second, we control the reward density by using both sparse and non**sparse** reward configurations. In the sparse setting, rewards are only granted upon completing the final objective (e.g., saving the victim), whereas in the non-sparse variant, intermediate rewards guide exploration and collection behavior. All combinations of these settings are tested across the full set of agents presented in this paper. Quantitative results are reported in Table III. Code and data will be made publicly available upon acceptance at https://github.com/dimipan/journal_LLM_RL.

C. Implementation Details

The simulation environment is implemented in Python using OpenAI Gym. Evaluations were conducted on a system with an Intel i7-12700 CPU, 32 GB RAM, and an Nvidia RTX A2000 GPU (6 GB VRAM). We train RL agents using offpolicy Q-learning with learning rate $\alpha=0.1$, discount factor

 $\gamma = 0.99$, and a decaying ε -greedy exploration strategy. Each experiment is run for 1000 episodes, averaged over 50 independent trials for statistical reliability. The state space includes the agent's position, number of information elements collected, and victim rescue status. The action space consists of system 34 actions in total, distributed across three taskspecific workers in the hierarchical framework. For navigation 4 directional movement actions, for information collection 26 actions, and for triage 4 rescue-related actions). We compare a total of 13 pre-trained LLMs integrated via a local RAG setup (Ollama, Chroma vector store, and LangChain). These models are assessed on their ability to interpret verbal inputs and generate structured outputs as described in Section III-D. The output consists of identified locations and their classified categories (e.g., POI or HAZ) paired with spatial coordinates, which are matched against the domain knowledge base B to support downstream decision-making. A subset of 5 LLMs is further evaluated in their role as exploration facilitators (Section III-E), specifically in guiding worker w_{TI} during information collection by generating zero-shot action predictions. For agents employing the attention-based shaping mechanism, we employ Gemma2 (9B) as the context extractor across all tested scenarios. As exploration facilitators, Hermes3 (8B) is deployed in the 3-info configuration and Llama 3.1 (8B) is used in the 6-info setting.

TABLE I PERFORMANCE COMPARISON OF LLM MODELS AS CONTEXT EXTRACTORS

Model	Acc. (%)	Resp. Time (s)	Loc. Errors	Class. Errors	Hall. Errors	Success Rate (%)
Gemma2 (9B)*	100.0	15.3	0.0	0.0	0.0	100.0
Llama3:instruct (8B)*	91.4	10.1	0.5	0.4	0.0	62.5
Hermes3 (8B)*	97.0	10.7	0.1	0.4	0.0	78.8
Dolphin3 (8B)*	95.3	9.5	0.0	0.9	0.0	68.8
Llama3.1 (8B)*	97.3	13.1	0.1	0.1	0.0	81.3
Tulu3 (8B)*	98.5	9.8	0.0	0.3	0.0	86.3
Qwen2.5 (7B)*	91.3	9.2	0.5	0.4	0.0	63.8
Mistral (7B)*	93.0	8.2	0.4	0.0	0.0	58.8
Zephyr (7B)	94.5	12.6	0.2	0.6	0.0	60.0
deepseek-r1 (7B)	95.1	21.6	0.3	0.2	0.0	73.0
Gemma3 (4B)*	93.9	4.9	0.4	0.4	0.0	60.0
Llama3.2 (3B)*	92.7	2.7	0.3	0.3	0.0	47.5
Qwen2.5 (3B)*	73.6	2.3	1.6	0.9	0.0	37.5

^{*}Models marked with an asterisk produced relevant responses adhering to the instructed output structure across all 80 test runs.

V. RESULTS AND ANALYSIS

A. Comparative Performance of LLMs

How do different LLMs perform as Context Extractors for SAR-related inputs? We evaluate the effectiveness of various LLMs in processing verbal inputs relevant to SAR tasks, focusing on their ability to convert natural language into actionable, structured representations. Table I presents a comparative analysis of model performance across a standardised test set of 14 verbal inputs, comprised of 7 simple and 7 complex scenarios, where complexity is defined by the number of locations referenced per input. Each model is evaluated over 80 independent runs. Accuracy (Acc.) is computed via a point-based scoring system: each correctly identified location earns 1

point, and each correct classification ("POI" vs. "HAZ") earns an additional point, with a maximum of 2 points per location. We track three distinct error types: *Location Errors*, indicating missed or incorrectly identified locations; *Classification Errors*, indicating correct locations misclassified as "POI" or "HAZ"; *Hallucination Errors* indicating identified locations not present in the original input. A model achieves a *Success Rate* of 100% only if it scores 100% accuracy and produces zero errors in all categories. This metric ensures correctness and reliability that is crucial for safety-critical applications like SAR. Response time is also recorded to evaluate real-time feasibility.

Among all models tested, *Gemma2* (9B) delivers the best overall performance, achieving 100.0% accuracy, zero errors, and a 100.0% success rate. However, this comes at the cost of the second longest response time, averaging 15.3 seconds. Notably, none of the models produced hallucination errors, highlighting their reliability in avoiding fabricated or misleading content, which is vital for such high-stakes decision-making.

- 1) 8B Models: The 8B-parameter models exhibit varied performance. Tulu3 (8B) leads this group with 98.5% accuracy, no location errors, 0.3 classification errors, and an 86.3% success rate, paired with a relatively fast response time of 9.8 seconds. Llama3.1 (8B) follows closely with 97.3% accuracy, minimal errors (0.1 location errors and 0.1 classification errors), and an 81.3% success rate, though its response time is longer at 13.1 seconds. Hermes3 (8B) achieves 97.0% accuracy, with 0.1 location errors and 0.4 classification errors, resulting in a 78.8% success rate and a response time of 10.7 seconds. Dolphin3 (8B) records 95.3% accuracy, no location errors, but a higher 0.9 classification errors, leading to a 68.8% success rate and a response time of 9.5 seconds. Llama3: instruct (8B) has the lowest accuracy in this group at 91.4%, with 0.5 location errors and 0.4 classification errors, yielding a 62.5% success rate and a response time of 10.1 seconds.
- 2) 7B Models: These models generally show slightly lower accuracy and success rates compared to the 8B models. In particular, deepseek-r1 (7B) stands out with 95.1% accuracy (based on 74 successful runs out of 80), 0.3 location errors, 0.2 classification errors, and a 73.0% success rate, but it has the longest response time in this category at 21.6 seconds. Zephyr (7B) achieves 94.5% accuracy (based on 70 successful runs out of 80), with 0.2 location errors and 0.6 classification errors, resulting in a 60.0% success rate and a response time of 12.6 seconds. Mistral (7B) records 93.0% accuracy, 0.4 location errors, no classification errors, and a 58.8% success rate, with a fast response time of 8.2 seconds. Qwen2.5 (7B) has 91.3% accuracy, 0.5 location errors, 0.4 classification errors, and a 63.8% success rate, with a response time of 9.2 seconds.
- 3) Smaller Models (4B and 3B): Smaller models demonstrate trade-offs between accuracy and response time. As shown, Gemma3 (4B) achieves 93.9% accuracy, with 0.4 location errors and 0.4 classification errors, a 60.0% success rate, and a fast response time of 4.9 seconds. Llama3.2 (3B) records 92.7% accuracy, 0.3 location errors, 0.3 classification errors, and a 47.5% success rate, with a very fast response time of 2.7 seconds. Qwen2.5 (3B) has the lowest performance, with

73.6% accuracy, 1.6 location errors, 0.9 classification errors, and a 37.5% success rate, but it offers the fastest response time at 2.3 seconds.

TABLE II PERFORMANCE COMPARISON OF LLM MODELS AS EXPLORATION FACILITATORS

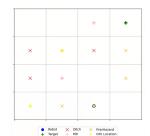
Model	Setting	Non Sparse		Sparse		
		Acc. (%)	Time (s)	Acc. (%)	Time (s)	
Gemma2 (9B)	3-info	96.8±3.9	1.6±0.12	72.8±6.4	1.6±0.07	
	6-info	86.5±6.8	1.8±0.06	69.8±8.2	1.7±0.09	
Llama3.1 (8B)	3-info	99.7±0.3	1.0±0.06	98.8±1.4	1.0±0.03	
	6-info	98.9 ±1.2	0.9±0.04	96.4 ±2.1	1.1±0.04	
Hermes3 (8B)	3-info 6-info	99.8 ±0.2 97.4±1.4	0.8±0.05 1.0±0.04	99.7 ±0.4 95.7±1.3	0.9±0.03 0.9±0.05	
Tulu3 (8B)	3-info	98.7±0.9	0.9±0.14	96.7±3.5	0.9±0.05	
	6-info	93.4±5.6	1.1±0.02	81.0±18.4	1.0±0.06	
Qwen2.5 (7B)	3-info	99.3±0.6	0.9±0.08	96.1±1.6	0.8±0.03	
	6-info	96.6±1.4	0.9±0.04	93.4±3.0	1.0±0.04	

How do different LLMs perform as Exploration Facilitators? Table II presents the performance of selected LLMs in their role as Exploration Facilitators, specifically guiding the w_{TI} (information collection) worker in the HierQ-LLM agent across both sparse and non-sparse reward settings, and under 3-info and 6-info configurations. Accuracy and response time are reported for each condition. Three key trends emerge from this evaluation. First, under non-sparse 3-info conditions, all models achieve over 95% accuracy. This suggests that in less complex environments with dense feedback, exploration guidance is straightforward and well-handled by all evaluated models. Second, Gemma2 (9B) shows unexpected performance degradation under both non-sparse 6-info and sparse settings. Accuracy drops to 86.5% in non-sparse 6-info, 72.8% in sparse 3-info, and 69.8% in sparse 6-info. While *Tulu3* (8B) also sees a drop in sparse 6-info (81.0%), it still outperforms Gemma2 under these conditions. This indicates potential limitations in Gemma2's generalization to longer-horizon or low-feedback scenarios despite its strong performance as a context extractor. Third, Hermes3 (8B) achieves the highest accuracy under the 3-info configuration, with near-perfect performance in both sparse (99.7%) and non-sparse (99.8%) settings, whereas *Llama3.1* (8B) leads under the more complex 6-info settings, with top accuracy in both sparse (96.4%) and non-sparse (98.9%) environments. These patterns suggest Hermes3 is best suited for short-horizon guidance, while Llama3.1 generalises better to longer, more complex missions. Overall, these results highlight that while most models perform well in easy scenarios, sparsity and task length expose important differences in robustness and generalisation.

B. Comparison of Shaping Methods

To quantitatively evaluate agent performance under varying levels of complexity and uncertainty, we employ five core metrics that capture both task effectiveness and decision

quality. Mission Success Rate (MSR): The percentage of episodes in which the agent completes the full mission by collecting all required information types and executing the correct rescue action at the designated target location. Information Collection Success Rate (ICSR): The percentage of episodes where the agent successfully collects all required information types, regardless of whether the final rescue action is completed. **Predictor Success Rate (PSR)**: For agents using an LLM to infer the correct information type, this metric reflects the accuracy of those predictions. For agents not using an LLM, it is defined as the accuracy of collecting the correct information through random action sampling. Mission Success Without Collisions (MSWC): A safety-oriented variant of MSR, counting only missions completed without encountering hazardous states. Average Reward (AR): The mean cumulative reward obtained per episode, indicating overall task efficiency. Table III summarises performance across both sparse and non-sparse reward environments and under 3-info and 6-info task configurations. Results show that flat agents (Q, Q-PS, Q-RS, Q-AS), even when enhanced with LLM-based context extraction, consistently exhibit low MSR—remaining below 15% across all scenarios. This demonstrates the limitations of non-hierarchical architectures in managing complex, multi-phase tasks. Hierarchical agents outperform their flat counterparts, demonstrating higher MSR, ICSR, and MSWC scores across the board. Shaping techniques, especially policy shaping (PS) and action shaping (AS), further boost mission success and safety. As expected, moving from 3-info to 6-info configurations results in noticeable performance degradation, especially in sparse environments. This trend confirms the increased difficulty of longer-horizon missions requiring greater contextual understanding and planning. Nonetheless, shaped hierarchical agents maintain comparatively high performance across all metrics, showcasing their adaptability and robustness in uncertain, high-stakes scenarios. A detailed discussion and interpretation of these performance trends is provided in Section VI.



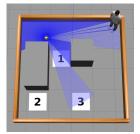


Fig. 4. 2D and 3D environments used for 3-info configuration.

VI. DISCUSSION & LIMITATIONS

Taken together, the results validate the effectiveness of combining hierarchical task structure with targeted shaping mechanisms, particularly in sparse and complex environments where flat agents consistently underperform. Hierarchical decomposition, when paired with shaping, supports more effecient and safe behavior. Beyond these broad trends, closer inspection of agent-specific outcomes reveals how different components of LUCIFER affect performance. Below, we unpack these effects

TABLE III

COMPARISON OF AGENTS UNDER NON-SPARSE (LEFT) AND SPARSE (RIGHT) ENVIRONMENTS IN 3 AND 6 INFORMATION SETTING. THE BEST IS MARKED IN **BOLD**, AND THE SECOND-BEST IN UNDERLINE.

	Non-Sparse Environment							
Agent	Setting	MSR (%)	ICSR (%)	PSR (%)	MSWC (%)	AR		
Q	3-info 6-info	0	1.2 1.1	$\begin{array}{c} \leq 5.0 \\ \leq 5.0 \end{array}$	0	-4.9 -4.6		
Q-PS	3-info 6-info	9.2 0	43.8 1.8	$\begin{array}{l} \leq 5.0 \\ \leq 5.0 \end{array}$	9.2 0	-0.7 -3.7		
Q-RS	3-info 6-info	13.9 0	41.0 1.5	$\begin{array}{l} \leq 5.0 \\ \leq 5.0 \end{array}$	0.9 0	4.3 -5.7		
Q-AS	3-info 6-info	10.8 0	45.9 2.4	$\begin{array}{l} \leq 5.0 \\ \leq 5.0 \end{array}$	10.8 0	1.4 -3.9		
HierQ	3-info 6-info	59.4 51.6	63.7 54.4	$\leq 5.0 \\ \leq 5.0$	0.5 0.7	55.0 56.7		
HierQ-LLM	3-info 6-info	60.4 53.6	<u>65.1</u> <u>56.9</u>	99.8 98.9	0.5 0.8	55.8 58.8		
HierQ-PS	3-info 6-info	<u>62.0</u> 52.4	63.3 54.4	$\begin{array}{c} \leq 5.0 \\ \leq 5.0 \end{array}$	62.0 52.4	<u>57.6</u> 56.9		
HierQ-RS	3-info 6-info	58.1 51.3	63.2 54.7	$\begin{array}{l} \leq 5.0 \\ \leq 5.0 \end{array}$	0.8 1.2	53.1 56.3		
HierQ-AS	3-info 6-info	57.8 51.3	63.2 54.4	$\begin{array}{l} \leq 5.0 \\ \leq 5.0 \end{array}$	57.8 51.3	51.8 55.3		
HierQ-LLM -PS	3-info 6-info	83.4 64.8	84.9 66.9	99.8 98.9	83.4 64.8	87.6 75.6		

	Sparse Environment							
Agent	Setting	MSR (%)	ICSR (%)	PSR (%)	MSWC (%)	AR		
Q	3-info 6-info	0	0	$\begin{array}{l} \leq 5.0 \\ \leq 5.0 \end{array}$	0 0	-32.4 -22.8		
Q-PS	3-info 6-info	0 0	0	$\begin{array}{l} \leq 5.0 \\ \leq 5.0 \end{array}$	0 0	-32.4 -22.8		
Q-RS	3-info 6-info	0	0	≤ 5.0 ≤ 5.0	0	-32.4 -22.7		
Q-AS	3-info 6-info	0	0	$\begin{array}{l} \leq 5.0 \\ \leq 5.0 \end{array}$	0	-32.4 -22.7		
HierQ	3-info 6-info	55.7 37.4	57.1 39.6	$\begin{array}{l} \leq 5.0 \\ \leq 5.0 \end{array}$	0.5 0.7	31.2 5.8		
HierQ-LLM	3-info 6-info	54.8 35.7	<u>57.7</u> 37.8	99.7 96.4	0.3 0.8	29.1 2.7		
HierQ-PS	3-info 6-info	<u>56.5</u> <u>38.2</u>	57.5 39.4	≤ 5.0 ≤ 5.0	<u>56.5</u> <u>38.2</u>	31.8 6.3		
HierQ-RS	3-info 6-info	54.5 37.2	57.2 39.9	$\begin{array}{l} \leq 5.0 \\ \leq 5.0 \end{array}$	0.6 1.0	29.8 5.3		
HierQ-AS	3-info 6-info	54.9 37.9	57.6 40.1	≤ 5.0 ≤ 5.0	54.9 37.9	29.3 5.5		
HierQ-LLM -PS	3-info 6-info	75.1 44.7	76.0 46.0	99.7 96.4	75.1 44.7	53.3 12.6		

and discuss key limitations, along with directions for future improvement.

A. Interpretation of Shaping and LLM Effects

A closer inspection of Table III reveals several key insights into the effects of shaping mechanisms and LLM integration.

Shaping promotes safety: For flat agents utilising policy or action shaping (e.g., Q-PS, Q-AS), successful missions are always completed safely as shown by MSWC equaling MSR in the non-sparse 3-info condition. This suggests that even with limited success overall, it inherently promotes policy safety. For hierarchical agents, this safety pattern generalises across all configurations, demonstrating the synergy between structured decomposition and shaping. Exploitation drives success: The consistently high MSR achieved by HierQ-PS across all settings may be attributed to its ability to quickly switch to exploitation once all required information is collected. With a shaped Q-table, the agent can complete the mission efficiently and reduce unnecessary exploration and exposure to risk. LLM Integration is bottlenecked by RL components: While HierQ-LLM performs competitively, it may not outperform across the board as expected. This is because the LLM only controls the information collection worker (w_{TI}) ; navigation (w_{TN}) and rescue (w_{TT}) are still governed by standard RL policies. Thus, poor navigation performance can limit the effectiveness of even perfect LLM-driven guidance. Formally, replacing the RL-based action selection process for a single worker affects only a subset of the entire mission pipeline. In larger environments or tasks requiring more diverse information types, we expect the exploration facilitator's impact to grow. Hybrid methods perform best: Finally, as expected, HierQ-LLM with policy shaping (HierQ-LLM-PS) yields the best overall performance across all configurations showing that attention space and LLM-guided reasoning complement each other.

B. Complete or Sufficient Information

A central design choice in LUCIFER is transitioning from exploration to exploitation once sufficient information, as defined by the Information Space, has been gathered. This switching to a "do" mode, reflecting the behavior of a rescuer who, after a preliminary survey, shifts from scouting to executing a plan, supports rapid mission execution, which is a critical requirement in real-world SAR contexts. However, in practice, information may be incomplete or dynamic. Blindly switching to exploitation based on early inputs could yield suboptimal outcomes if new or contradictory data emerges mid-mission. Future work could explore adaptive mechanisms that allow the agent to reassess confidence or re-enter information-gathering mode when contextual uncertainty increases.

C. Learning Scalability

Our evaluation employs tabular RL agents to make shaping effects interpretable and easy to implement. Although reward and action space shaping techniques are readily transferable to deep reinforcement learning (DRL) settings as they operate on universal RL components, policy shaping is more challenging due to the absence of explicit Q-tables in function-approximated policies. Future work could explore embedding

LLM-derived signals into policy networks. Second, while LUCIFER employs a fixed task decomposition based on domain knowledge, learning high-level policies dynamically, such as through a Semi-Markov Decision Process (SMDP), could enhance adaptability by enabling the agent to select tasks autonomously.

D. Dependence on LLMs

LUCIFER's context extractor and exploration facilitator rely on the quality of LLM outputs, which introduces risks. Even with RAG-based augmentation (e.g., with domain manuals or maps), LLMs may hallucinate, misclassify, or suffer from retrieval mismatches. For instance, retrieving data from a flood manual in a wildfire scenario could destabilise behavior. Mitigating this risk requires improving domain adaptation and retrieval accuracy. Crucially, our use of LLMs aligns with prior work advocating for heuristic use of LLM outputs rather than direct policy substitution [33]. LUCIFER extends this principle by incorporating LLM-derived knowledge as a guiding heuristic through an attention space to ensure that planning remains adaptive rather than rigidly dictated by the LLM. Additionally, in our framework, the context extractor is integrated into the agent's reasoning and decision-making pipeline. However, it can easily be adapted as a standalone module to automate the translation of complex linguistic reports into structured insights. This could reduce cognitive load on human commanders in SAR command-and-control centers. Future work could explicitly investigate and quantify this operational benefit.

E. Human Language Variability & Communication Gaps

Finally, from the human-in-the-loop perspective, a further limitation arises from the nature of human language [59]. Communication in high-pressure situations is often vague or unclear, yet current implementation assumes a reasonable level of clarity in verbal inputs to function effectively. During controlled proof-of-concept experiments, this assumption held allowing us to test the core hypothesis: that spatial insights derived from human linguistic descriptions can steer an agent more effectively than purely autonomous exploration. However, real-world deployments may violate this assumption requiring agents to proactively query human stakeholders to elicit missing or ambiguous information. Developing interactive dialogue mechanisms, driven by uncertainty or task relevance, would allow the agent to close knowledge gaps in real time reframing the RL objective to include informationseeking behaviors. Addressing these limitations represents a promising next step toward more adaptive, scalable, and human-aware learning systems.

VII. CONCLUSION

In summary, we introduce LUCIFER, a domain-agnostic framework that lays the foundation for decision-making systems capable of integrating human knowledge while maintaining algorithmic rigor. It represents a substantial advancement in the synthesis of human and machine intelligence, resulting

in an integrated system that is computationally robust and inherently human-centric. Unlike traditional systems that treat humans as passive supervisors, our proposed framework positions stakeholders as active knowledge providers, underscoring the value of human informational processing within computational systems. By unifying dual-role LLMs with a hierarchical architecture, the framework supports context-aware reasoning and behavior in high-stakes environments. Looking ahead, extending the system to support interactive, bidirectional communication between humans and agents presents a compelling direction for future research and a paradigm shift in how intelligent systems are designed and deployed into real-world operations.

REFERENCES

- [1] R. Murphy, J. Casper, J. Hyams, M. Micire, and B. Minten, "Mobility and sensing demands in usar," in 2000 26th Annual Conference of the IEEE Industrial Electronics Society. IECON 2000. 2000 IEEE International Conference on Industrial Electronics, Control and Instrumentation. 21st Century Technologies, vol. 1. IEEE, 2000, pp. 138–142.
- [2] G.-J. M. Kruijff, M. Janíček, S. Keshavdas, B. Larochelle, H. Zender, N. J. Smets, T. Mioch, M. A. Neerincx, J. V. Diggelen, F. Colas et al., "Experience in system design for human-robot teaming in urban search and rescue," in Field and Service Robotics: Results of the 8th International Conference. Springer, 2014, pp. 111–125.
- [3] A. Pirinen, A. Samuelsson, J. Backsund, and K. Åström, "Aerial view localization with reinforcement learning: Towards emulating search-andrescue," arXiv preprint arXiv:2209.03694, 2022.
- [4] C. Gruffeille, A. Perrusquía, A. Tsourdos, and W. Guo, "Disaster area coverage optimisation using reinforcement learning," in 2024 International Conference on Unmanned Aircraft Systems (ICUAS). IEEE, 2024, pp. 61–67.
- [5] S. J. Russell and P. Norvig, Artificial intelligence: a modern approach. Pearson, 2016.
- [6] L. P. Kaelbling, M. L. Littman, and A. R. Cassandra, "Planning and acting in partially observable stochastic domains," *Artificial intelligence*, vol. 101, no. 1-2, pp. 99–134, 1998.
- [7] A.-F. Saeed and N. Kasim, "Role of stakeholders in mitigating disaster prevalence: Theoretical perspective," in *MATEC Web of Conferences*, vol. 266. EDP Sciences, 2019, p. 03008.
- [8] A. V. Ter-Mkrtchyan and A. L. Franklin, "Stakeholder analysis in the context of natural disaster mitigation: The case of flooding in three us cities," *Sustainability*, vol. 15, no. 20, p. 14945, 2023.
- [9] G. Oulahen, B. Vogel, and C. Gouett-Hanna, "Quick response disaster research: Opportunities and challenges for a new funding program," *International Journal of Disaster Risk Science*, vol. 11, pp. 568–577, 2020.
- [10] A. Ganji and S. Miles, "Toward human-centered simulation modeling for critical infrastructure disaster recovery planning," in 2018 IEEE Global Humanitarian Technology Conference (GHTC). IEEE, 2018, pp. 1–8.
- [11] D. Erokhin and N. Komendantova, "Understanding human behavior response to disasters," in Oxford Research Encyclopedia of Natural Hazard Science, 2024.
- [12] S. Nahavandi, "Trusted autonomy between humans and robots: Toward human-on-the-loop in robotics and autonomous systems," *IEEE Systems, Man, and Cybernetics Magazine*, vol. 3, no. 1, pp. 10–17, 2017.
- [13] 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.
- [14] R. S. Sutton, A. G. Barto et al., Reinforcement learning: An introduction. MIT press Cambridge, 1998, vol. 1, no. 1.
- [15] 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.
- [16] C. G. Correa, M. K. Ho, F. Callaway, and T. L. Griffiths, "Resourcerational task decomposition to minimize planning costs," arXiv preprint arXiv:2007.13862, 2020.
- [17] H. Osooli, "A multi-robot task assignment framework for search and rescue with heterogeneous teams," Master's thesis, University of Massachusetts Lowell, 2024.

- [18] S. Hart, V. Steane, and M. Chattington, "Prospective decision modelling of uncrewed aerial vehicle operators to inform design recommendations for future systems," in 2024 IEEE 4th International Conference on Human-Machine Systems (ICHMS). IEEE, 2024, pp. 1–6.
- [19] 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," 2019. [Online]. Available: https://arxiv.org/abs/1906.03926
- [20] 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," arXiv preprint arXiv:2402.01874, 2024.
- [21] Y. Cao, H. Zhao, Y. Cheng, T. Shu, Y. Chen, G. Liu, G. Liang, J. Zhao, J. Yan, and Y. Li, "Survey on large language model-enhanced reinforcement learning: Concept, taxonomy, and methods," *IEEE Transactions* on Neural Networks and Learning Systems, 2024.
- [22] A. R. Laleh and M. N. Ahmadabadi, "A survey on enhancing reinforcement learning in complex environments: Insights from human and Ilm feedback," arXiv preprint arXiv:2411.13410, 2024.
- [23] M. Kwon, S. M. Xie, K. Bullard, and D. Sadigh, "Reward design with language models," arXiv preprint arXiv:2303.00001, 2023.
- [24] Y. J. Ma, W. Liang, G. Wang, D.-A. Huang, O. Bastani, D. Jayaraman, Y. Zhu, L. Fan, and A. Anandkumar, "Eureka: Humanlevel reward design via coding large language models," arXiv preprint arXiv:2310.12931, 2023.
- [25] T. Xie, S. Zhao, C. H. Wu, Y. Liu, Q. Luo, V. Zhong, Y. Yang, and T. Yu, "Text2reward: Reward shaping with language models for reinforcement learning," arXiv preprint arXiv:2309.11489, 2023.
- [26] J. Song, Z. Zhou, J. Liu, C. Fang, Z. Shu, and L. Ma, "Self-refined large language model as automated reward function designer for deep reinforcement learning in robotics," arXiv preprint arXiv:2309.06687, 2023.
- [27] K. Chu, X. Zhao, C. Weber, M. Li, and S. Wermter, "Accelerating reinforcement learning of robotic manipulations via feedback from large language models," arXiv preprint arXiv:2311.02379, 2023.
- [28] A. Adeniji, A. Xie, C. Sferrazza, Y. Seo, S. James, and P. Abbeel, "Language reward modulation for pretraining reinforcement learning," arXiv preprint arXiv:2308.12270, 2023.
- [29] H. Li, X. Yang, Z. Wang, X. Zhu, J. Zhou, Y. Qiao, X. Wang, H. Li, L. Lu, and J. Dai, "Auto mc-reward: Automated dense reward design with large language models for minecraft," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 16426–16435.
- [30] M. Klissarov, P. D'Oro, S. Sodhani, R. Raileanu, P.-L. Bacon, P. Vincent, A. Zhang, and M. Henaff, "Motif: Intrinsic motivation from artificial intelligence feedback," arXiv preprint arXiv:2310.00166, 2023.
- [31] Y. Qu, Y. Jiang, B. Wang, Y. Mao, C. Wang, C. Liu, and X. Ji, "Latent reward: Llm-empowered credit assignment in episodic reinforcement learning," arXiv preprint arXiv:2412.11120, 2024.
- [32] C. H. Song, J. Wu, C. Washington, B. M. Sadler, W.-L. Chao, and Y. Su, "Llm-planner: Few-shot grounded planning for embodied agents with large language models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 2998–3009.
- [33] D. Shah, M. R. Equi, B. Osiński, F. Xia, B. Ichter, and S. Levine, "Navigation with large language models: Semantic guesswork as a heuristic for planning," in *Conference on Robot Learning*. PMLR, 2023, pp. 2683–2699.
- [34] B. Hu, C. Zhao, P. Zhang, Z. Zhou, Y. Yang, Z. Xu, and B. Liu, "Enabling intelligent interactions between an agent and an Ilm: A reinforcement learning approach," arXiv preprint arXiv:2306.03604, 2023
- [35] R. Yang, J. Chen, Y. Zhang, S. Yuan, A. Chen, K. Richardson, Y. Xiao, and D. Yang, "Selfgoal: Your language agents already know how to achieve high-level goals," arXiv preprint arXiv:2406.04784, 2024.
- [36] C. Colas, L. Teodorescu, P.-Y. Oudeyer, X. Yuan, and M.-A. Côté, "Augmenting autotelic agents with large language models," in *Conference on Lifelong Learning Agents*. PMLR, 2023, pp. 205–226.
- [37] G. Pourcel, T. Carta, G. Kovač, and P.-Y. Oudeyer, "Autotelic Ilm-based exploration for goal-conditioned rl," in *Intrinsically Motivated Open*ended Learning Workshop at NeurIPS 2024, 2024.
- [38] Z. Wang, S. Cai, G. Chen, A. Liu, X. Ma, and Y. Liang, "Describe, explain, plan and select: Interactive planning with large language models enables open-world multi-task agents," arXiv preprint arXiv:2302.01560, 2023.
- [39] Z. Peng, Y. Wang, X. Han, L. Zheng, and J. Ma, "Learningflow: Automated policy learning workflow for urban driving with large language models," arXiv preprint arXiv:2501.05057, 2025.

- [40] C. Lynch, A. Wahid, J. Tompson, T. Ding, J. Betker, R. Baruch, T. Armstrong, and P. Florence, "Interactive language: Talking to robots in real time," *IEEE Robotics and Automation Letters*, 2023.
- [41] H. Liu, L. Lee, K. Lee, and P. Abbeel, "Instruction-following agents with multimodal transformer," arXiv preprint arXiv:2210.13431, 2022.
- [42] Z. Huang, Z. Sheng, C. Ma, and S. Chen, "Human as ai mentor: Enhanced human-in-the-loop reinforcement learning for safe and efficient autonomous driving," *Communications in Transportation Research*, vol. 4, p. 100127, 2024.
- [43] Y. Du, O. Watkins, Z. Wang, C. Colas, T. Darrell, P. Abbeel, A. Gupta, and J. Andreas, "Guiding pretraining in reinforcement learning with large language models," in *International Conference on Machine Learning*. PMLR, 2023, pp. 8657–8677.
- [44] T. Carta, C. Romac, T. Wolf, S. Lamprier, O. Sigaud, and P.-Y. Oudeyer, "Grounding large language models in interactive environments with online reinforcement learning," in *International Conference on Machine Learning*. PMLR, 2023, pp. 3676–3713.
- [45] F. Paischer, T. Adler, V. Patil, A. Bitto-Nemling, M. Holzleitner, S. Lehner, H. Eghbal-Zadeh, and S. Hochreiter, "History compression via language models in reinforcement learning," in *International Conference on Machine Learning*. PMLR, 2022, pp. 17156–17185.
- [46] F. Paischer, T. Adler, M. Hofmarcher, and S. Hochreiter, "Semantic helm: A human-readable memory for reinforcement learning," Advances in Neural Information Processing Systems, vol. 36, 2024.
- [47] W. K. Kim, S. Kim, H. Woo et al., "Efficient policy adaptation with contrastive prompt ensemble for embodied agents," Advances in Neural Information Processing Systems, vol. 36, 2024.
- [48] R. P. Poudel, H. Pandya, S. Liwicki, and R. Cipolla, "Recore: Regularized contrastive representation learning of world model," in *Proceedings* of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 22904–22913.
- [49] J.-C. Pang, X.-Y. Yang, S.-H. Yang, and Y. Yu, "Natural language-conditioned reinforcement learning with inside-out task language development and translation," arXiv preprint arXiv:2302.09368, 2023.
- [50] S. Basavatia, K. Murugesan, and S. Ratnakar, "Starling: Self-supervised training of text-based reinforcement learning agent with large language models," arXiv preprint arXiv:2406.05872, 2024.
- [51] B. A. Spiegel, Z. Yang, W. Jurayj, B. Bachmann, S. Tellex, and G. Konidaris, "Informing reinforcement learning agents by grounding language to markov decision processes," in Workshop on Training Agents with Foundation Models at RLC 2024, 2024.
- [52] P.-L. H. Cabot and R. Navigli, "Rebel: Relation extraction by end-to-end language generation," in *Findings of the Association for Computational Linguistics: EMNLP 2021*, 2021, pp. 2370–2381.
- [53] X. Wang, W. Zhou, C. Zu, H. Xia, T. Chen, Y. Zhang, R. Zheng, J. Ye, Q. Zhang, T. Gui et al., "Instructuie: Multi-task instruction tuning for unified information extraction," arXiv preprint arXiv:2304.08085, 2023.
- [54] Y. Tang, L. Moffat, W. Guo, C. May-Chahal, J. Deville, and A. Tsourdos, "Encoding social & ethical values in autonomous navigation: Philosophies behind an interactive online demonstration," in *Proceedings of the* Second International Symposium on Trustworthy Autonomous Systems, 2024, pp. 1–9.
- [55] L. Von Rueden, S. Mayer, K. Beckh, B. Georgiev, S. Giesselbach, R. Heese, B. Kirsch, J. Pfrommer, A. Pick, R. Ramamurthy et al., "Informed machine learning-a taxonomy and survey of integrating prior knowledge into learning systems," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 1, pp. 614–633, 2021.
- [56] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel et al., "Retrievalaugmented generation for knowledge-intensive nlp tasks," Advances in neural information processing systems, vol. 33, pp. 9459–9474, 2020.
- [57] D. Panagopoulos, A. Perrusquia, and W. Guo, "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). IEEE, 2024, pp. 1175–1180.
- [58] F. Lieder and T. L. Griffiths, "Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources," *Behavioral and brain sciences*, vol. 43, p. e1, 2020.
- [59] S. T. Piantadosi, H. Tily, and E. Gibson, "The communicative function of ambiguity in language," *Cognition*, vol. 122, no. 3, pp. 280–291, 2012.