2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

Intrinsic Language-Guided Exploration for Complex Long-Horizon Robotic Manipulation Tasks

Eleftherios Triantafyllidis[1], Filippos Christianos[1] and Zhibin Li[2]

***Abstract*— Current reinforcement learning algorithms struggle in sparse and complex environments, most notably in long-horizon manipulation tasks entailing a plethora of different sequences. In this work, we propose the Intrinsically Guided Exploration from Large Language Models (IGE-LLMs) framework. By leveraging LLMs as an assistive intrinsic reward, IGE-LLMs guides the exploratory process in reinforcement learning to address intricate long-horizon with sparse rewards robotic manipulation tasks. We evaluate our framework and related intrinsic learning methods in an environment challenged with exploration, and a complex robotic manipulation task challenged by both exploration and long-horizons. Results show IGE-LLMs (i) exhibit notably higher performance over related intrinsic methods and the direct use of LLMs in decision-making, (ii) can be combined and complement existing learning methods highlighting its modularity, (iii) are fairly insensitive to different intrinsic scaling parameters, and (iv) maintain robustness against increased levels of uncertainty and horizons.**

I. INTRODUCTION

Current deep Reinforcement Learning (RL) approaches are faced with significant challenges in complex and sparse environments, particularly in sequential long-horizon robotic tasks [1], [2], [3]. A challenge in RL is the need for exploration, where immediate feedback from the environment is not readily apparent and usually in the form of sparse rewards [4]. Overcoming these limitations with methods such as ϵ-greedy policies [5], or the introduction of noise in the actions space [6] can be inefficient in sparse long-horizon tasks [3], [7]. Even hand-crafting specific dense rewards requires notable engineering efforts, limiting generalisation [8], [9].

A promising alternative for encouraging exploration in such cases is the introduction of intrinsic rewards (r^i) [4], [7], [10], [11], [12], [13]. While these methods enhance exploration in long-horizon, sparse-reward RL problems, these can still overemphasise noisy and non-relevant to the end-goal state transitions due to prediction errors [7], [14].

On the other hand, Large Language Models (LLMs) provide promising means of context and common-sense aware reasoning that could aid agents in otherwise complex settings, by emphasising more relevant state transitions based on the task at hand [8], [15], [16]. As such, the emergence of LLMs and their language-based instructions have shown promising applications in conjunction with deep learning [8], [15], [17], [18], [19] and in particular in the domain of robotics [15], [20]. Nevertheless, while LLMs hold promise, it's essential to consider their limitations, as their outputs may not always be reliable nor optimal [19], [21]. Moreover, current methods utilising LLMs are faced with challenges, ranging from grounding and constraining LLMs [19], [20], [21], limiting emergent behaviour by following specific manual guidelines [18], directly relying on goal suggestions that may deviate from the relevance of the task at hand [8] and often requiring significant prompt engineering efforts [22].

Instead, we propose IGE-LLMs (see Figure 1), a framework utilising LLMs as an assistive intrinsic proxy reward, alongside the traditional extrinsic in the conventional RL setting. In this way, we assert that even naturally occurring incorrect LLM replies [19], [21], will not cause the policy to learn sub-optimally as the replies are used for guidance, with the extrinsic reward still eventually being the main policy driver. We hypothesise that utilising LLMs with their context-aware reasoning, can mitigate the limitations of current intrinsic methods exploring possibly non-related to the task states [7], [14]. We further assert that the integration of LLMs with RL and the random exploration of the agent with its environment will compensate for the occasional LLM's inaccuracies [19], [21]. We show that this synergy bridges the gaps of each approach, rather than using these in isolation, which to the best of our knowledge, is a current gap.

The contributions of our work are summarised as follows:

- A modular approach of utilising LLMs as intrinsic assistance, fostering exploration in complex environments with existing RL settings and algorithms;
- An extensive comparison of state-of-the-art intrinsic methods and related works on utilising LLMs;
- A method capable of promoting exploration in complex, long-horizon sequential robotic manipulation tasks;
- A computationally efficient approach of incorporating LLMs in RL paradigms whereby recurring situations of visited states are derived from a generated dictionary.

In the remainder, we present the related work, elaborate on the technical details of IGE-LLMs and present the results.

II. RELATED WORK

We consider related work that utilise LLMs, with a focus on the RL decision making process and work on intrinsically-motivated exploration, encompassing robot learning.

A. Reinforcement Learning – Trials, Errors, and the Pursuit of Elusive Sparse Rewards

Pre-programming robots as embodied intelligence via analytical models is sub-optimal due to the oversimplification of

[1]Authors are with the School of Informatics, The University of Edinburgh, United Kingdom. {eleftherios.triantafyllidis, f.christianos}@ed.ac.uk

[2]Zhibin Li is with the Department of Computer Science, University College London, United Kingdom. alex.li@ucl.ac.uk

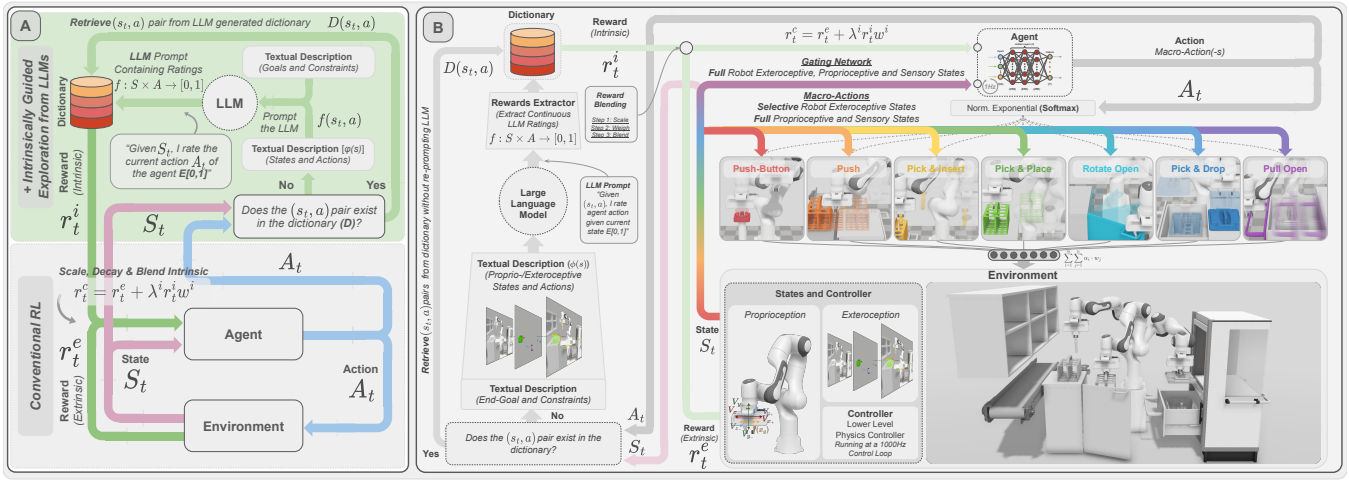

Fig. 1: Schematics illustrating the principles of our method. **(A)** The overview of IGE-LLMs. **(B)** IGE-LLMs on ROMAN's hierarchical architecture [3] for solving complex robotic manipulation tasks entailing sparse rewards and long horizons.

modelling real-world dynamics [3]. Deep learning algorithms [23], particularly RL, offer a promising alternative with their capacity to learn from interactions with the environment [24], [25], drawing inspiration from biology whereby even humans in their early lives must learn representations from unlabelled data [26], [27]. The common deep RL algorithms are the Proximal Policy Optimization (PPO) [28] and Soft Actor-Critic (SAC) [29]. While PPO is on-policy and generally less sample efficient than the off-policy SAC, we choose PPO in this work as it is less prone to instabilities and typically requires less hyperparameter tuning than SAC [28], [30].

In RL, agents aim to learn a set of policies by maximising their returns from the environment, known as extrinsic rewards. In general, when these extrinsic rewards are provided in an immediate or continuous manner, RL policies perform very well [3]. However, in many real-world robotic cases, these extrinsic rewards can be (i) **sparse** exacerbating the exploration-exploitation trade-off [26], [31], [32], (ii) **misaligned** [4] or (iii) require very careful **reward crafting** limiting generalisation [4], [8]. Unfortunately, for agents to randomly reach a sparse goal, especially in long-horizon sequential robotic manipulation tasks [1], [3], is highly unlikely due to the increased need for exploration [4], leading to resource-intensive learning sessions [33], [34].

B. Intrinsic Guidance – The Compass Navigating Through the Maze of Reinforcement Learning

To tackle environments with sparse rewards, a prominent category of exploration techniques used are intrinsic rewards [4], [7], [11], [14], [35]. Intrinsic rewards (r^i) are computed and added to the extrinsic term (r^e), such that the total (r^c) is $r^c = r^e + \lambda r^i$, with λ representing a scaling factor. Over time, r^i is decayed, by which point the agent should converge towards the extrinsic reward, transitioning from exploration to exploitation [4]. There are two main categories, (i) **count-based** and (ii) **prediction-based** intrinsic reward methods.

Count-based exploration methods are traditional intrinsic methods that incentivise agents to visit rarely encountered states, useful in small, discrete state spaces [35]. Nevertheless, these can be less effective in high-dimensional state spaces, commonly seen in complex robotic manipulation tasks entailing long horizon and sequential operation [3].

Recent prediction-based methods such as the Intrinsic Curiosity Module (ICM) [7] and the Random Network Distillation (RND) [11] have been proposed. ICM utilises prediction error, promoting the agent to explore new states, while RND generates intrinsic rewards by comparing feature representations of the agent with those of a fixed random network. Nonetheless, ICM can overemphasise noisy state transitions due to prediction error, potentially exploring non-relevant to a given task state transitions [7], [14]. On the other hand, RND's static random network might not capture evolving complexities in dynamically changing environments such as physics-based robotic interactions [3], [11], [14].

As it can be inferred, while intrinsic rewards can foster exploration, these may explore states beyond the relevance of the task at hand [4], [7], [11], [14], [35]. Perhaps, by merging the exploratory potential of intrinsic rewards in the form of contextual insights of LLMs, a more directed and efficient exploration strategy can be achieved.

C. Guiding Reinforcement Learning with LLMs

LLMs are capable of providing useful context as well as common-sense aware reasoning [8], [20]. Nevertheless, a primary limitation of LLMs is that their outputs can occasionally be inaccurate [19], [21]. The work of [18] utilised LLMs to extract information from game manuals via a QA extraction and reasoning module, whereby auxiliary rewards are inferred from "Yes/No" answers. While innovative, this inherently requires specific instructions, ultimately hard-wiring behaviours to the contents of the manual [18].

The work of [19], introduced grounded LLMs serving as the agent's policy which is progressively updated via online RL. However, LLM's estimations cannot be guaranteed to be accurate [21], potentially leading to non-optimal behaviour. In another work, [17] utilised LLMs to shape the behaviour of RL agents against user-specified objectives as textual

prompts so as to align the agent's behaviour with human-described objectives. While novel, this framework is evaluated on rather short horizons, only provides binary reward signals and finally replaces the traditional reward function with a proxy reward, rendering it reliant on the LLM's outputs which is prone to inaccuracies [19], [21]. The work of [8], presented a goal-oriented approach whereby LLMs are used as the means of encouraging exploration in RL by rewarding the attainment of LLM-suggested goals based on and reliant on state descriptions to generate exploration goals.

Instead, in our method and unlike [8], [17], [18], we utilise LLMs for the provision of a dense, continuous assistive intrinsic reward in environments naturally challenged by long-horizons and sparse extrinsic rewards. This intrinsic reward is decayed over time so as to not overemphasise plausible inaccuracies stemming from the LLM which can lead to bias [19], [21]. Instead, by decaying the intrinsic term, we assert that the agent will naturally converge to the extrinsic term, thus fostering emergent behaviour to achieve the end goal, while balancing exploitation and exploration. Moreover, our method facilitates learning efficiency in recurring situations whereby visited state-action pairs (s_t, a_t) can be derived from a generated dictionary, limiting the need for LLM inference.

III. METHODOLOGY

A. Technical Preliminaries

We first outline technical preliminaries including related count-based and prediction-based intrinsic rewards.

Markov Decision Process: We consider a Markov Decision Process (MDP) as a tuple (S, A, P, R, γ), whereby S and A represent the state and action spaces, respectively. $P(s'|s, a) = \Pr(s_{t+1} = s'|s_t = s, a_t = a)$ represents the transition function, specifying the probability of transitioning to the next state (s') given the current state (s) and the applied action (a). The extrinsic reward function given as $R(s, a, s')$ provides the reward to the agent received after transitioning from state s to state s' given action a. The $\gamma \in [0, 1]$ is the discount factor. The goal is to learn a policy $\pi(a|s)$ that defines the probability of taking action a in state s such that the expected discounted returns are maximised as $\mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t \mathcal{R}(s_t, a_t, s_{t+1}) \mid a_t \sim \pi(s_t)\right]$. To promote exploration, we can also define a function that complements the extrinsic reward function R with an intrinsic reward.

Count-Based Intrinsic Reward: With count-based intrinsic rewards, the agents are encouraged to visit states that have not frequently been encountered [4], [35], [36]. These rewards are inverse proportional to the visitations of already encountered states and are commonly represented as: $r_t^i = \frac{1}{\sqrt{N(s_t)}}$, whereby r_t^i denotes the intrinsic reward at time-step t and $N(s_t)$ the count of the encountered current state. Thereby, higher intrinsic rewards are given for states that have been visited less frequently, encouraging exploration.

Intrinsic Curiosity Module (ICM): With ICM, an intrinsic reward signal is introduced based on the ability of the agent to predict the consequence of its actions in a learned feature space. The intrinsic reward is formulated as $r_t^i = \alpha(\widehat{\phi}(s_{t+1}) - \phi(s_{t+1}))^2$, whereby α is a scaling factor. $\widehat{\phi}$ denotes the predicted representation of the next state given the current state and action, while ϕ represents the actual representation of the next state. In this way, the model attempts to learn to encode information affected by the action of the agent. Essentially, r^i represents the prediction error of the agent's estimate of the next state's feature representation. Thus, the agent is encouraged to perform actions that maximise the error, exploring novel areas [7].

Random Network Distillation (RND): RND introduces a prediction-based intrinsic reward signal that is derived from the ability of the learned network, referred to as the predictor, to mimic a randomly initialised network, referred to as the target network. The intrinsic reward r^i is given as $r_t^i = (\widehat{\phi}(s_{t+1}) - \phi(s_{t+1}))^2$, whereby $\widehat{\phi}$ and ϕ represent the predictor and target networks respectively for state s [11].

B. Intrinsically Guided Exploration from LLMs (IGE-LLMs)

We propose a novel way of utilising LLMs to encourage efficient exploration in RL, particularly in long-horizon sequential tasks with sparse rewards. Given a state s_t at time t, the LLM is tasked, via textual prompts, with evaluating the potential future rewards of the agent's actions $a \in A$. This evaluation is performed by a function $f : S \times A \to [0, 1]$, whereby S and A represent the state and action spaces respectively. The output of the action evaluation is represented as a continuous scalar value in $[0, 1]$. For every pair of states and actions (s_t, a), the LLM is tasked with assigning a rating $r^i = f(s_t, a)$, representing the desirability of performing action a at state s_t. Intuitively, f guides the agent towards actions that the LLM perceives as most beneficial given s_t.

This process can be repeated once for every new state, prompting the LLM to rate the available actions the agent can perform and build a dictionary for every seed. By creating a dictionary, $D : S \times A \to [0, 1]$, computational costs of running inference on the LLM or API-related costs can be minimised. Whenever the agent encounters a state s_t, it retrieves the corresponding intrinsic reward from the dictionary D. If the state-action pair (s_t, a) exists in D, hence previously encountered, the intrinsic reward is directly obtained as $r^i = D(s_t, a)$. Alternatively, if the state-action pair is not present in D, hence not yet encountered, the LLM is prompted to compute the intrinsic reward $r^i = f(s_t, a)$, which is stored as $D(s_t, a) = r^i$. Hence, the dictionary efficiently guides the agent's future actions in similar states.

By leveraging the existing RL paradigm, the total reward is given as: $r^c = r^e + \lambda^i \cdot r^i \cdot w^i$, representing the combined reward (r^c), entailing the sum of the extrinsic (r^e) and intrinsic (r^i) terms. The intrinsic term is furthermore controlled by a scaling factor (λ^i) as well as a linearly decaying weight (w^i) to guide the policy, without overfitting it to the intrinsic reward. In this way, the agent receives a decaying over-time dense intrinsic reward and gradually learns to converge towards the sparse extrinsic. We hypothesise this approach will encourage exploration in sparse reward environments entailing long horizons, commonly seen in complex robotic manipulation tasks [3]. Figure 1 depicts our method.

C. Apparatus and System Configuration

Combining learning algorithms with simulators mimicking real-world physics facilitates faster robotic learning and minimises the risks of hardware damage [24], [34], [37]. Hence, for a realistic learning-based robotic simulation, Unity3D (with PhysX) was used alongside the PyTorch-based ML-Agents toolkit [38], similarly to [3]. Moreover, the ROS# plugin alongside the physics frequency set to 1kHz, ensured accurate physical modelling and stability [34], [37]. The *GPT-4* LLM was employed, without any fine-tuning.

To further emulate a realistic robotic setup, we also utilise a vision system similar to [3], to detect the exteroceptive states and in particular the Objects of Interest (OIs) from the scene. The OIs are visualised in colour in Figure 1.b. The system implements an object detection module based on the VGG-16 backbone [39], initialised with pre-trained weights on the ImageNet dataset and fine-tuned with the OIs in the simulation. This system generated textual scene descriptions for the LLM, indicating the presence or absence of OIs.

IV. Extending the ROBotic MANipulation Network (ROMAN)

All methods were validated on a complex, long-horizon sequential robotic manipulation task, based on [3] and as depicted in Figure 1.b, Figure 2.a and Figure 3). In such settings, the attainment of a long-horizon sequential goal is contingent upon the successful completion of other sub-tasks; for instance, one cannot retrieve an item from a drawer without first opening it. We hypothesise that LLMs can be of notable benefit in such notably intricate long-horizon robotic manipulation tasks necessitating the correct sequential coordination and activation of a plethora of macro-actions [3]. To test our hypothesis, we extend the work known as ROMAN [3], an event-based Hybrid Hierarchical Learning (HHL) architecture composed of distinct specialising experts treated as **macro-actions**, commonly seen in real-life [40].

A. Main Environment

We concentrate on the environment where ROMAN was assessed on, see Figure 1.b. The environment consisted of a laboratory setting, with the primary objective of retrieving a vial, placing it in a rack, and moving it onto a conveyor belt. Within this setting, additional sub-tasks were derived, ensuring their interdependence and relevance [3], [40]. We maintain the macro-actions and evaluate the methodologies of this study hereinafter on the gating network and its ability to coordinate these to solve intricate long-horizon sequential robotic manipulation tasks. We hypothesise that our method, applied to ROMAN's gating network, will leverage the LLM's context-aware reasoning for enhanced exploration.

B. Internal Hybrid Learning Procedure

The expert Neural Network (NN)s, treated as macro-actions, were hybrid trained, combining Behavioural Cloning (BC) [41], RL (PPO) [28], and Generative Adversarial Imitation Learning (GAIL) [42]. Policies were warm-started with BC, thereafter updated using RL (PPO), with r^e and r^i from the environment and the GAIL discriminator respectively. Each NN received $N = 20$ demonstrations from keybindings, corresponding to velocity controlling the end-effector and its binary gripper state. For further details consult [3].

C. Expert NNs - States, Actions and Rewards

The proprioceptive states were identical for all experts. Exteroceptive states differed, dependent on each NN's specialising skill. This allowed each NN to focus on its own exteroceptive states relevant to its sub-task, similarly to how humans assess information relevance during motor tasks [43]. Experts shared identical actions, controlling the velocity of the robotic end-effector and the binary gripper state. A sparse, terminal reward $(+1)$ was given upon completion of their specific sub-task goal. For more details consult [3].

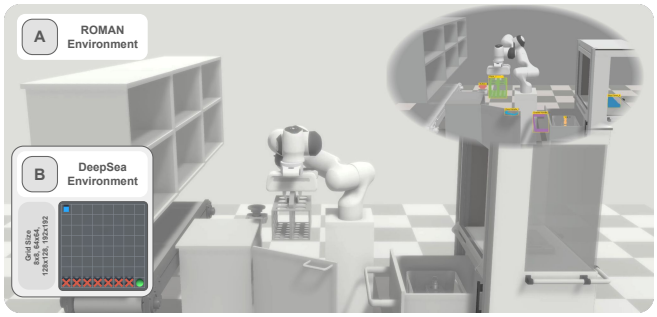

Fig. 2: Evaluation environments. **A:** ROMAN. **B:** DeepSea.

V. EVALUATION

We evaluate all methods in two environments (see Figure 2): a toy setting challenged by RL exploration and the main robotic environment challenged in both exploration and long-horizons. Each training seed utilising the LLM, corresponds to a different generated dictionary, ensuring randomness and diverse prompts. See [44] for detailed prompt examples. Lastly, a function $\phi(s)$ is employed to convert state vectors into textual descriptions for LLM input, given direct state vectors might be uninterpretable for LLMs [15]. For fairness with baselines, high-level descriptions are also converted and integrated as binary values in the state vector.

A. Illustrating the idea on a Toy Environment – DeepSea

We first experiment in a NxN grid-based environment – DeepSea; whereby the challenge of exploration in RL is targeted [45]. The agent starts at the top-left and aims to reach the goal located in the bottom-right grid. We evaluate increasingly complex grid sizes of $N \in \{8, 64, 128, 192\}$.

a) States, Actions and Rewards: The state space (s_t) included the agent's position $(x1, y1)$ and target goal $(x2, y2)$. Actions (a_t) corresponded to moving down $(x, y + 1)$ or diagonally down-right $(x + 1, y + 1)$. The goal yielded a +1 sparse reward; while other bottom tiles incurred a -1 penalty.

b) LLM Implementation: The LLM is provided a prompt at each time-step of the DeepSea environment. Similarly to the NNs having access to the positions of the agent and the goal, the LLM is provided identical information in the form of textual descriptions. The LLM is tasked to rate

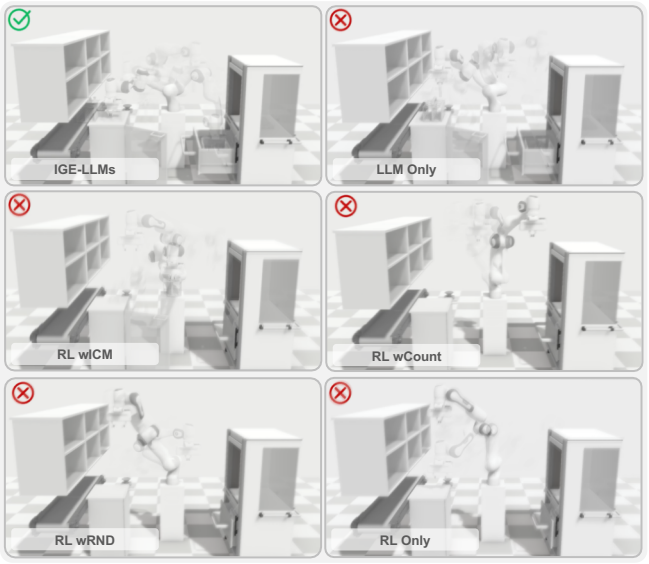

Fig. 3: IGE-LLMs and other methods on ROMAN's longest-horizon task, case seven, at $\sigma = \pm 0.5$cm noise.

$(r^i \in [0, 1])$ the two actions (a_t) of the agent, given the current position of the agent and the goal's constant position.

B. Complex Long-Horizon of Sparse Rewards Robotic Tasks

For the main environment, ROMAN, commonly seen complex robotic manipulation tasks are studied [3], [40].

a) States, Actions and Rewards: In this environment, the gating network was the primary NN used for the evaluation and its ability to orchestrate its macro-actions to achieve the long-horizon end goal. We incorporate a softmax function, $\sigma(\mathbf{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}$ to normalise the sum of weights of the seven macro-actions. The state space comprised the combined exteroceptive states of all its macro-actions in the hierarchy, allowing it to oversee the entire environment. Its proprioceptive states included the end-effector's position, velocity and gripper state. A sparse terminal $+1$ reward was provided only upon completing the sequential end-goal. Given the high complexity of the task, we posit that dense intrinsic rewards by the LLM with their context-aware reasoning can aid learning in long-horizon manipulation.

b) LLM Implementation: The LLM is provided with a prompt at each time-step equivalent to the frequency of the gating NN $(1Hz)$. The LLM is provided high-level textual descriptions corresponding to the binary states of the primary OIs in the scene, which is also integrated as binary values in the state vector to ensure fairness with baseline comparisons. At each time-step the LLM is tasked to rate $(r^i \in [0, 1])$ the seven macro actions (a_t), based on the textual description of the OIs, rating as such the state-action pairs (s_t, a).

VI. RESULTS

We first compare conventional RL utilising extrinsic rewards (**RL**), thereafter combining it with intrinsic methods, including prediction-based (**ICM**, **RND**), **count-based**, and our **IGE-LLM**. Next, we combine our method with existing intrinsic methods to demonstrate its modularity and robustness. To underscore the inherent limitations of occasional inaccuracies stemming from LLMs [19], [21], we also compare the LLM's direct output, without training. To this end, we also utilise state-of-the-art reasoning referred to as Chain of Thought (CoT) [22], due to its ability to elicit notably improved arithmetic, commonsense, and symbolic reasoning. For direct **LLM** and **LLM with CoT** outputs, we use the **argmax**. To incorporate randomness into the decision-making process with the argmax, when identically rated actions by the LLM are present in the dictionary, a random selection is made amongst those occurrences. Lastly, we perform a sensitivity analysis of our method and a robustness test of all trained models compared in this work. Smoothing and averaging are applied to Figure 4 for visualisation purposes. Consult the accompanying video for demonstrations of the models in the environments. The prompts provided to the LLM and its output $(r^i \in [0, 1])$, were zero-shot, while those incorporating CoT were one-shot, due to the inclusion of a contextual example.

A. Using LLMs as Intrinsic Assistance

Results for DeepSea and the main ROMAN environments are depicted in Figure 4. It is inferred that while existing intrinsic reward methods (r^i) are overall beneficial in promoting exploration compared to just the use of RL (r^e), these still struggle with increased dimensionality and horizons. RL+Count performs mostly well for DeepSea, yet shows notable drops in returns in ROMAN. RL+RND and RL+ICM both struggle with increased grid sizes in DeepSea and ROMAN, yet RL+ICM shows higher returns in ROMAN, still not close to the maximum attainable. Conversely, IGE-LLMs consistently performs well in DeepSea across all grid sizes and most importantly outperforms all other methods in ROMAN. Moreover, Figure 4 shows that our method can also be combined, and complements, existing prediction and count-based intrinsic methods, highlighting its modularity.

B. The Constraints of Direct Integration of LLMs

The direct integration of LLMs into the decision-making (see Figure 4), exhibit significant errors, even with advanced LLMs (*GPT-4*). The argmax results in Figure 4 show that the direct integration of the LLM, even when incorporating CoT [22] is inadequate, especially for in ROMAN's complex long-horizon tasks. It is concluded that directly relying on LLMs should be avoided, in line with [19], [21]. Instead, results show that utilising LLMs as an assistive intrinsic signal is preferable. In this way, the agent's innate interactions within its environment, as per the RL paradigm, can counteract the inherent limitations when directly relying on and integrating LLMs in the decision-making process.

C. Sensitivity Analysis

A common challenge when employing intrinsic-motivated exploration, is the sensitivity to the intrinsic scale λ [4]. Hence, to evaluate the sensitivity of our method, we evaluate different logarithmic spaced scaling parameters $\lambda^i \in$

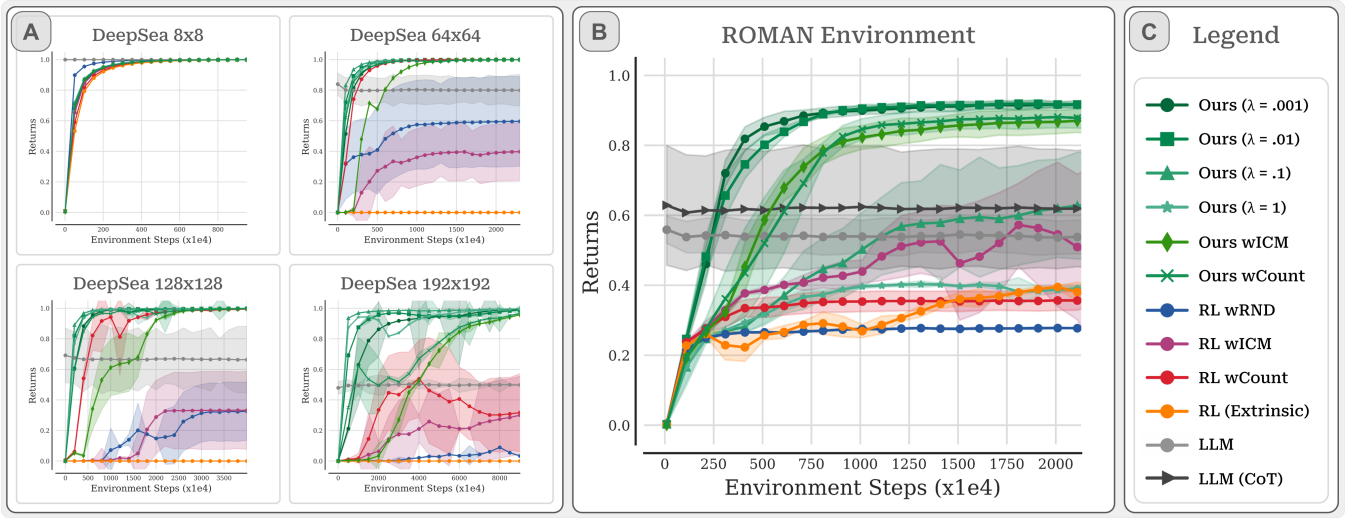

Fig. 4: Normalised evaluation returns. Shading depicts the standard deviation (σ) around the mean. **A** DeepSea, from $n = 5$ seeds for *8x8* and *64x64* and $n = 3$ seeds for *128x128* and *192x192*. **B** ROMAN from $n = 5$ seeds. **C** Legend.

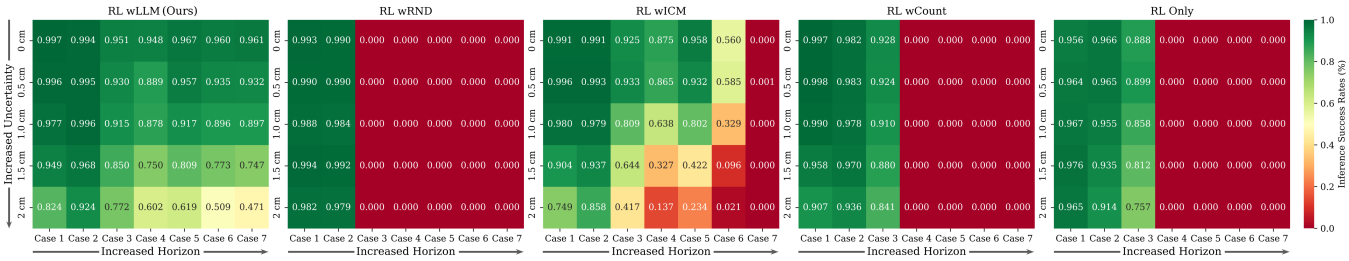

Fig. 5: Inference results for the ROMAN environment across five distinct models. The x-axis represents the task horizon, ascending from left to right, y-axis the exteroceptive noise, ascending from top to bottom. Each cell stems from 1K episodes.

$\{0.001, 0.01, 0.1, 1\}$. In this way, the robustness of our approach can be demonstrated from a fraction up to exceeding λ values to the extrinsic reward. From Figure 4 results show that IGE-LLMs exhibits robustness and is fairly insensitive to varying λ scales, yet with some observed influence in those exceeding r^e (i.e. $\lambda^i \in \{1\}$). The LLM and LLM wCoT results in Figure 4.b suggest that overemphasis on the LLM is discouraged due to its occasional inaccuracies [19], [21].

D. Inference Results - Robustness Test

To further evaluate our model's robustness, we add Gaussian noise to all exteroceptive states in ROMAN's gating NN, linked to positional observations. We test from 0cm to up to $\sigma = \pm 2$cm noise levels, in 0.5cm increments. From Figure 5, it is inferred that utilising r^i is overall beneficial compared to using just r^e. While RL wICM shows higher success rates than other methods, errors increase with longer horizons and uncertainty. RL, RL wRND and RL wCount appear to struggle with longer horizons, with Cases 4 to 7 being effectively unattainable. In contrast, IGE-LLMs, maintains high success rates across increasing noise and horizons, notably outperforming other methods. Also consult Figure 3.

VII. Discussion and Conclusion

A. Limitations and Future Work

While LLMs require textual descriptions, methods such as Contrastive Language-Image Pre-Training (CLIP) [46] could automate the description of exteroceptive states. Combining our method with [15] to bridge high-level instructions to robotic actions via LLMs, investigating different levels of weight decays (w^i) and a comparison between different LLM models, may prove beneficial. Nevertheless, results showed that even an advanced LLM (*GPT-4*) was overall inadequate when used in isolation, highlighting the value of using LLMs as mere guidance rather than directly as main policy drivers.

B. Conclusion

This work presented IGE-LLMs, a novel approach leveraging LLMs for intrinsic assistance, promoting exploration in RL tasks challenged by sparse rewards and long-horizons. Validated on a preliminary and subsequent intricate long-horizon robotic manipulation environment, results showed IGE-LLMs outperformed existing intrinsic methods and underlined the shortcomings of the direct use of LLMs in the decision making process. Furthermore, IGE-LLMs modularity was underscored by its ability to be combined and complement existing intrinsic methods, its insensitivity to most scaling parameters featured, and its consistent robustness against increased uncertainty and horizons highlighted. Collectively, these findings underline the value of IGE-LLMs for tackling the exploration and long-horizon challenges inherent in complex robotic manipulation tasks.

Acknowledgment

Supported by the EPSRC CDT in RAS (EP/L016834/1).

REFERENCES

[1] T. Davchev, O. O. Sushkov, J.-B. Regli, S. Schaal, Y. Aytar, M. Wulfmeier, and J. Scholz, "Wish you were here: Hindsight goal selection for long-horizon dexterous manipulation," in *International Conference on Learning Representations*, 2022. [Online]. Available: https://openreview.net/forum?id=FKp8-pIRo3y

[2] R. Fox, R. Berenstein, I. Stoica, and K. Goldberg, "Multi-task hierarchical imitation learning for home automation," in *2019 IEEE 15th International Conference on Automation Science and Engineering (CASE)*, 2019, pp. 1–8.

[3] E. Triantafyllidis, F. Acero, Z. Liu, and Z. Li, "Hybrid hierarchical learning for solving complex sequential tasks using the robotic manipulation network roman," *Nature Machine Intelligence*, vol. 5, no. 9, pp. 991–1005, Sep 2023. [Online]. Available: https://doi.org/10.1038/s42256-023-00709-2

[4] L. Schäfer, F. Christianos, J. P. Hanna, and S. V. Albrecht, "Decoupled reinforcement learning to stabilise intrinsically-motivated exploration," in *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, ser. AAMAS '22. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems, 2022, p. 1146–1154.

[5] C. J. C. H. Watkins, "Learning from delayed rewards," Ph.D. dissertation, King's College, Cambridge, 1989.

[6] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," in *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2016. [Online]. Available: http://arxiv.org/abs/1509.02971

[7] D. Pathak, P. Agrawal, A. A. Efros, and T. Darrell, "Curiosity-driven exploration by self-supervised prediction," in *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ser. ICML'17. JMLR.org, 2017, p. 2778–2787.

[8] 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," 2023.

[9] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané, "Concrete problems in ai safety," 2016.

[10] Y. Flet-Berliac, J. Ferret, O. Pietquin, P. Preux, and M. Geist, "Adversarially guided actor-critic," in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=_mQp5cr_iNy

[11] Y. Burda, H. Edwards, A. Storkey, and O. Klimov, "Exploration by random network distillation," in *International Conference on Learning Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=H1lJJnR5Ym

[12] P.-Y. Oudeyer, F. Kaplan, and V. V. Hafner, "Intrinsic motivation systems for autonomous mental development," *IEEE Transactions on Evolutionary Computation*, vol. 11, no. 2, pp. 265–286, 2007.

[13] N. Chentanez, A. Barto, and S. Singh, "Intrinsically motivated reinforcement learning," in *Advances in Neural Information Processing Systems*, L. Saul, Y. Weiss, and L. Bottou, Eds., vol. 17. MIT Press, 2004. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2004/file/4be5a36cbaca8ab9d2066debfe4e65c1-Paper.pdf

[14] A. A. Taiga, W. Fedus, M. C. Machado, A. Courville, and M. G. Bellemare, "On bonus based exploration methods in the arcade learning environment," in *International Conference on Learning Representations*, 2020. [Online]. Available: https://openreview.net/forum?id=BJewlyStDr

[15] W. Yu, N. Gileadi, C. Fu, S. Kirmani, K.-H. Lee, M. G. Arenas, H.-T. L. Chiang, T. Erez, L. Hasenclever, J. Humplik, B. Ichter, T. Xiao, P. Xu, A. Zeng, T. Zhang, N. Heess, D. Sadigh, J. Tan, Y. Tassa, and F. Xia, "Language to rewards for robotic skill synthesis," 2023.

[16] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[17] M. Kwon, S. M. Xie, K. Bullard, and D. Sadigh, "Reward design with language models," 2023.

[18] Y. Wu, Y. Fan, P. P. Liang, A. Azaria, Y. Li, and T. M. Mitchell, "Read and reap the rewards: Learning to play atari with the help of instruction manuals," 2023.

[19] 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," 2023.

[20] M. Ahn, A. Brohan, N. Brown, Y. Chebotar, O. Cortes, B. David, C. Finn, C. Fu, K. Gopalakrishnan, K. Hausman, A. Herzog, D. Ho, J. Hsu, J. Ibarz, B. Ichter, A. Irpan, E. Jang, R. J. Ruano, K. Jeffrey, S. Jesmonth, N. J. Joshi, R. Julian, D. Kalashnikov, Y. Kuang, K.-H. Lee, S. Levine, Y. Lu, L. Luu, C. Parada, P. Pastor, J. Quiambao, K. Rao, J. Rettinghouse, D. Reyes, P. Sermanet, N. Sievers, C. Tan, A. Toshev, V. Vanhoucke, F. Xia, T. Xiao, P. Xu, S. Xu, M. Yan, and A. Zeng, "Do as i can, not as i say: Grounding language in robotic affordances," 2022.

[21] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 1877–1901. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

[22] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, and D. Zhou, "Chain-of-thought prompting elicits reasoning in large language models," 2023.

[23] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015. [Online]. Available: https://doi.org/10.1038/nature14539

[24] M. A. Lee, Y. Zhu, K. Srinivasan, P. Shah, S. Savarese, L. Fei-Fei, A. Garg, and J. Bohg, "Making sense of vision and touch: Self-supervised learning of multimodal representations for contact-rich tasks," in *2019 International Conference on Robotics and Automation (ICRA)*, 2019, pp. 8943–8950.

[25] T. Zhang, Z. McCarthy, O. Jow, D. Lee, X. Chen, K. Goldberg, and P. Abbeel, "Deep imitation learning for complex manipulation tasks from virtual reality teleoperation," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, 2018, pp. 5628–5635.

[26] L. Zaadnoordijk, T. R. Besold, and R. Cusack, "Lessons from infant learning for unsupervised machine learning," *Nature Machine Intelligence*, vol. 4, no. 6, pp. 510–520, Jun 2022. [Online]. Available: https://doi.org/10.1038/s42256-022-00488-2

[27] A. Saxe, S. Nelli, and C. Summerfield, "If deep learning is the answer, what is the question?" *Nature Reviews Neuroscience*, vol. 22, no. 1, pp. 55–67, Jan 2021. [Online]. Available: https://doi.org/10.1038/s41583-020-00395-8

[28] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," 2017.

[29] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *Proceedings of the 35th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, J. Dy and A. Krause, Eds., vol. 80. PMLR, 10–15 Jul 2018, pp. 1861–1870. [Online]. Available: https://proceedings.mlr.press/v80/haarnoja18b.html

[30] S. Gu, T. Lillicrap, Z. Ghahramani, R. E. Turner, B. Schölkopf, and S. Levine, "Interpolated policy gradient: Merging on-policy and off-policy gradient estimation for deep reinforcement learning," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 3849–3858.

[31] N. Koganti, A. Rahman H. A. G., Y. Iwasawa, K. Nakayama, and Y. Matsuo, "Virtual reality as a user-friendly interface for learning from demonstrations," in *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*, ser. CHI EA '18. New York, NY, USA: Association for Computing Machinery, 2018, p. 1–4. [Online]. Available: https://doi.org/10.1145/3170427.3186500

[32] Y. Ding, C. Florensa, P. Abbeel, and M. Phielipp, "Goal-conditioned imitation learning," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32. Curran Associates, Inc., 2019. [Online]. Available: https://proceedings.neurips.cc/paper/2019/file/c8d3a760ebab631565f8509d84b3b3f1-Paper.pdf

[33] M. Thor and P. Manoonpong, "Versatile modular neural locomotion control with fast learning," *Nature Machine Intelligence*, vol. 4, no. 2, pp. 169–179, Feb 2022. [Online]. Available: https://doi.org/10.1038/s42256-022-00444-0

[34] C. Yang, K. Yuan, Q. Zhu, W. Yu, and Z. Li, "Multi-expert learning of adaptive legged locomotion," *Science Robotics*, vol. 5, no. 49, p. eabb2174, 2020. [Online]. Available: https://www.science.org/doi/abs/10.1126/scirobotics.abb2174

[35] M. G. Bellemare, S. Srinivasan, G. Ostrovski, T. Schaul, D. Saxton, and R. Munos, "Unifying count-based exploration and intrinsic motivation," 2016.

[36] G. Ostrovski, M. G. Bellemare, A. van den Oord, and R. Munos, "Count-based exploration with neural density models," in *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ser. ICML'17. JMLR.org, 2017, p. 2721–2730.

[37] E. Triantafyllidis, W. Hu, C. McGreavy, and Z. Li, "Metrics for 3d object pointing and manipulation in virtual reality: The introduction and validation of a novel approach in measuring human performance," *IEEE Robotics Automation Magazine*, pp. 2–17, 2021.

[38] A. Juliani, V.-P. Berges, E. Teng, A. Cohen, J. Harper, C. Elion, C. Goy, Y. Gao, H. Henry, M. Mattar, *et al.*, "Unity: A general platform for intelligent agents," *arXiv preprint arXiv:1809.02627*, 2018.

[39] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel, "Domain randomization for transferring deep neural networks from simulation to the real world," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2017, pp. 23–30.

[40] A. Billard and D. Kragic, "Trends and challenges in robot manipulation," *Science*, vol. 364, no. 6446, 2019. [Online]. Available: https://science.sciencemag.org/content/364/6446/eaat8414

[41] F. Torabi, G. Warnell, and P. Stone, "Behavioral cloning from observation," in *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, ser. IJCAI'18. AAAI Press, 2018, p. 4950–4957.

[42] J. Ho and S. Ermon, "Generative adversarial imitation learning," in *Advances in Neural Information Processing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, Eds., vol. 29. Curran Associates, Inc., 2016. [Online]. Available: https://proceedings.neurips.cc/paper/2016/file/cc7e2b878868cbae992d1fb743995d8f-Paper.pdf

[43] D. M. Wolpert, J. Diedrichsen, and J. R. Flanagan, "Principles of sensorimotor learning," *Nature Reviews Neuroscience*, vol. 12, no. 12, pp. 739–751, Dec 2011. [Online]. Available: https://doi.org/10.1038/nrn3112

[44] E. Triantafyllidis, F. Christianos, and Z. Li, "Intrinsic language-guided exploration for complex long-horizon robotic manipulation tasks," 2023.

[45] I. Osband, Y. Doron, M. Hessel, J. Aslanides, E. Sezener, A. Saraiva, K. McKinney, T. Lattimore, C. Szepesvari, S. Singh, B. V. Roy, R. Sutton, D. Silver, and H. V. Hasselt, "Behaviour suite for reinforcement learning," in *International Conference on Learning Representations*, 2020. [Online]. Available: https://openreview.net/forum?id=rygf-kSYwH

[46] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," 2021.