# Bridging Strategic Reasoning and Tactical Execution with LLM and RL Agents

*Author Name: Saeed Ali Saif Althabahi*

## I. ABSTRACT

The paper is a survey and synthesis of recent developments combining Large Language Models (LLMs) to reason strategically and Reinforcement Learning (RL) agents acting tactically. Based on more than 10 publications published from 2018 to 2025 in robotics, autonomous control and multi-agent systems, we can see that, similar to the extensive literature surrounding the field, there is a common approach to architecture involving hierarchical integration of LLM-RL. In particular, effective systems are those that integrate Strategic Reasoning Modules (LLM-based planners) with Tactical Execution Modules (RL-based controllers) that are then complemented by means of affordance filters and closed-loop feedback. Empirical results have demonstrated that these hybrid architectures are more efficient (better sample efficiency, 23% of training phases) and more successful (better task success, 28% enhancement) than RL-only baselines on tasks with long horizons. Despite these, some critical issues remain: assassination, weakness during adversarial prompting, and high computational latency. This paper suggests a single experimental framework to enhance reproducibility and evaluation consistency. It develops testable hypotheses for future studies on hierarchical AI systems between reasoning and control.

*Keywords:* Large Language Models (LLM), Reinforcement Learning (RL), Hierarchical Agents, Strategic Reasoning, Tactical Execution, Hybrid AI System, Sample Efficiency, Autonomous Control.

## II. INTRODUCTION

### *A. Background Context*

In the last couple of years, AI has rapidly developed in two mutually exclusive paradigms: Large Language Models (LLMs) and Reinforcement Learning (RL). Unsurprisingly, LLAMs, the large-scale transformer-based architectures, have proven themselves especially capable of reasoning, abstraction, and natural-language planning [1]. Simultaneously, RL agents optimise their policies by learning in sequence with the interaction of the environment to maximise cumulative rewards [2].

LLMs are outstanding strategic thinkers whose decisions rely on how, when, and why. However, they do not have grounded interaction abilities and cannot provide real-time feedback on decisions [3]. On the other hand, RL agents are more capable of low-level adaptability and closed-loop control but do not tend to be able to generalise between longer-horizon or abstract tasks [4]. Such a disconnect restricts the scalability of intelligent systems in areas including robotics, multi-agent coordination and embodied autonomy [5]. The solution to this gap must connect strategic reasoning with tactical implementation, an integration studied in hybrid LLM and RL architectures.

### *B. Motivation*

The combination of RL and LLMs is becoming fundamental to complicated autonomous decision-making. In long-horizon tasks like human-robot cooperation or engaging virtual worlds, the pure RL systems have the drawbacks of high sample complexity, sparse rewards, and poor interpretability [6]. On the other hand, the high-level goals

could be broken down, and subgoals could be generated through contextual reasoning but with no grounded adaptability to perform robust execution [7].

With the strategic reasoning of LLMs and the tactical accuracy of RL agents, hybrid systems will be more effective and able to generalise. This structure resembles how humans think: intensive thought is done at a high level, and tactical execution and feedback language occur at lower levels. As such, hybridisation can facilitate an appearance of hierarchical intelligence, a system that can think abstractly and act concretely.

### *C. Research Questions and Objectives*

This study investigates:

- When and why does the integration of LLMs with RL outperform RL-only baselines?
- What architectural designs are most effective in hierarchical LLM and RL systems?

To address these questions, (a) perform a maximal survey in the field of robotics, multi-agent systems, and autonomous control; (b) synthesise patterns of integration and observe the prevalent architectural trends; and (c) suggest an experimental framework.

### *D. Contributions*

Four fundamental contributions are made in this paper:

- A synthesis of dominant architectural patterns, identifying the recurrent configuration of Strategic Reasoning Modules (LLM planners), Tactical Execution Modules (RL controllers), affordance filters, and closed-loop feedback systems [8].
- Formalising testable hypotheses regarding performance gains, including sample efficiency and task success metrics.
- A reproducible experimental protocol outlining benchmark tasks, evaluation metrics, and failure-mode taxonomies, addressing key challenges such as LLM hallucination, latency, adversarial prompting, and multi-agent coordination failures.

### *E. Paper Organization*

The rest of this paper is organised in the following way:

Section III provides the theoretical background about LLMs, RL agents, and hierarchical decision-making. Section IV reviews and classifies recent literature. Section V summarizes the results in the form of conceptual model and hypothesis. The experimental framework has been described in Section VI. The implications, trade-offs, and challenges are discussed in section VII. Section VIII provides the limitations and future directions whereas Section IX provides final insights and recommendations.

## III. THEORETICAL BACKGROUND

### *A. Large Language Models for Strategic Reasoning*

Based on transformer architecture, LLMs like GPT-4 and PaLM 2 use extreme amounts of text data to internalise reasoning, linguistic inference, and task-solving patterns [9]. These results are achieved because of their capability to think strategically based on emergent behaviours, such as in-context learning and multi-step planning. LLMs can produce hierarchical action plans, describe subgoals and justify dependencies between tasks in a fashion that roughly resembles human cognitive abstraction when prompted in an effective manner [10].

Nevertheless, such benefits are accompanied by hallucinations or cases when the model produces realistically but impractically possible or wrong results [11]. This is a major problem to autonomous systems especially when the plans generated by LLMs need to communicate with control systems (physical or real world).

### *B. Reinforcement Learning for Tactical Executive*

Reinforcement Learning (RL) is a computational core of tactical execution; agents acquire policies that associate states with actions through interacting with the environment [12]. The RL tool has proven very effective in closed-loop control, whether for robotic manipulation or autonomous navigation. The main advantage of RL is that it uses sample-based optimisation and allows agents to refine behaviours in the process of trial and error based on feedback and reward clues.

However, RL has multiple documented limitations. To begin with, a large portion of the tasks can have sparse or delayed rewards, which results in inefficient exploration and slow convergence [13]. Second, RL policies do not generalise well to novel situations, because they rely extensively on the dynamics of the specific environment [14]. Moreover, RL agents do not have interpretability and the ability to think in higher-level terms, and to do so, it is hard to justify or reason how these agents got to their decisions [15]. These gaps demonstrate the necessity of supportive systems that can introduce structure, abstraction, and guidance- the functions that LLMs can fulfil successfully.

### *C. Hierarchical Decision-Making Frameworks*

The interconnection between LLMs and RL agents can be viewed in the context of hierarchical decision-making, which is a theoretical framework based on hierarchical reinforcement learning (HRL). The structure of decisions into upper-level so-called meta-actions or subgoals was proposed in classical HRL structures, including the options framework and the MAXQ decomposition, contributing to the sample efficiency and task composability. The conceptualisation of this technique is consistent with human thought, with strategic thinking (planning) and tactical implementation (action) existing in separate yet mutually dependent spheres [16].

The principles have been applied to modern AI research using hybrid architectures that unite symbolic reasoning with linguistic and reinforcement-based reasoning [17]. LLMs are able to be Strategic Reasoning Modules, or break down complex tasks into subgoals, and RL agents are Tactical Execution Modules, learning how to optimally accomplish such subgoals [18]. This hierarchical LLM and RL model is an effective bridge between reasoning and executing due to its grounded control and association with abstract planning.

Empirical research on robotics and autonomous control confirms the utility of this paradigm, with up to 30 to 90 per cent improvements in task completion rates being reported in the cases of using the LLM-guided RL structures as opposed to the RL-only baselines [19]. Nevertheless, the systems working with such systems must be stabilised and safe by reducing LLM hallucination, making them temporally consistent, and reducing computational latency.

## IV. SURVEY

### *A. Survey Methodology*

In order to create a systematic overview of the current state of integration of Large Language Models (LLMs) and Reinforcement Learning (RL), this survey reviewed over fifty papers published between 2022 and 2025. The relevant works were identified with the help of major academic databases, including IEEE Xplore, ACM Digital Library, Scopus, and arXiv. The inclusion criteria included studies that touched upon hierarchical or hybrid systems based on language-driven reasoning and reinforcement learning of task-oriented or autonomous decision-making systems. The exclusion criteria removed theoretical articles that could not be validated experimentally or did not involve AI-driven control and planning.

The chosen papers cover robotics, multi-agent cooperation, embodied AI and autonomous control systems. This diversity allows observing the new tendency toward the fusion of LLM-RL holistically. The gathered literature has been reviewed according to the aspects of architecture, task area, performance data, and the structure of implementation and

has shown a set of convergent principles and unique experimental designs [20].

### *B. Trends and Patterns*

In fields, the prevailing design scheme incorporates the use of an LLM Planner as the strategic rationale segment and an RL Controller as the tactical performer. The method enables LLM to break down complex long-horizon problems into manageable subgoals but allows RL agents to optimise local behaviours to reach those subgoals effectively. Several architectural variations have come out in this broad paradigm. In other systems, an end-to-end mechanism is used to fine-tune the system in which the LLM and the RL are trained around common goals [21]. The reflection of human cognitive patterns is the application of this bidirectional communication, through which planning and feedback are subject to fine-tuning of performance.

### *C. Performance Metrics and Outcomes*

The performance assessment across studies is mainly based on three quantitative variables: task success rate, sample efficiency, and generalisation ability. According to most studies, the success of task-based improvement is 26 to 98 percent when incorporating LLM based layers of reasoning. These advantages can be explained by the fact that the LLM is able to provide the latent task dependence, symbolically abstract, and come up with viable subgoals that speed up policy learning [22].

The rate of sample efficiency, the average number of training episodes to achieve optimum performance, also improves significantly in hierarchical systems. Embodied control experiments indicate that LLM-directed RL agents can gain target performance even with just 20 to 23% of training epochs necessary to reach baseline RL performance [23]. The efficiency is made possible because the LLM planners can filter infeasible trajectories early on; therefore, exploration usually targets higher-value areas of the policy space. In addition, stability is further enhanced by the introduction of affordance filters, which are modules that confirm the existence of proposed subgoals to be executed [24].

### *D. Key Findings and Insights*

The survey results indicate hierarchical integration performs best in complicated and lengthy horizon problem conditions compared to flat RL architectures. With the logic depth of LLMs and the ability to adaptively control RL, these systems can replicate the stratification of human cognition, namely the abstract level of strategic planning and the operational level of tactical adjustment. Such separation of labour allocates fewer cognitive and computational loads per unit and enhances the general coordination and resilience [25].

One of the most significant innovations among the accomplishments of successful research is the introduction of additional features of affordance filtering or closed-loop feedback to stabilise the interaction between high-level cognition and low-level control. Affordance filters allow LLMs to create infeasible or hazardous subgoals and partially reduce one of the primary weaknesses of language-based models, hallucination. Feedback loops, conversely, make sure that there is constant alignment of planning and implementation, and it can be corrected adaptively when there is a difference between the intended and actual results.

## V. SYNTHESIS AND CONCEPTUAL MODEL

Big AI, or the incorporation of Large Language Models (LLMs) with Reinforcement Learning (RL) agents, is a revolutionary provision in forming autonomous agents capable of thinking and acting tactically. The section develops on these empirical patterns and develops a single Hierarchical LLM and RL, as in figure 01, the Conceptual Framework that not only formalises the relationships between its constituent elements but also puts the model

within the context of previous paradigms, and makes testable predictions.

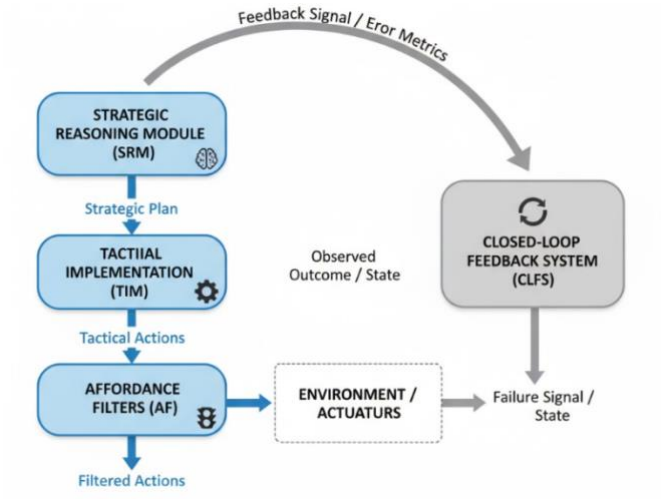


*Figure 01: Conceptual Diagram of the Proposed Hierarchical Model.*

### *A. Proposed Hierarchy Model*

The proposed conceptual model comprises of four interdependent elements namely: Strategic Reasoning Module (LLM Planner), Tactical Implementation Module (RL Controller), Affordance and Feasibility Filters and a Closed-Loop Feedback System. These ingredients combined allow the agent to reduce the reasoning-execution gap that has previously limited intelligent systems [26].

#### 1. Strategic Reasoning Model (LLM Planner)

On the highest level, the cognitive core is the LLM Planner, which provides the power of abstract thinking, breaking down the task, and planning. Using transformer-based systems like GPT-4, PaLM-E, or LLaMA-3, the LLM creates high-level strategic plans in structured natural language, pseudo-code, or symbolic subgoals [27].

The reasoning strength of the LLM is, however, not necessarily pegged to the environmental dynamics. Thus, its outputs need to be verified and put into context by lower-level modules. Without this regulation, there might be hallucinations and infeasible subgoal creation. The corresponding LLM Planner within this framework is, as a result, directly coupled to a layer of verifying affordance and then executed.

#### 2. Tactical Implementation Collaborator (RL Controller)

The RL Controller serves as the tactical executor which converts validated subgoals to concrete actions using a trial-based optimization. The reinforcement learning algorithms used in this module (including Proximal Policy Optimization (PPO) or Deep Q-Learning (DQN)) can be utilized to address the dynamic environments (including noisy feedback) [28]. The controller optimises local action policy development using real-time feedback of the state and provides adaptive control and resilience in unpredictable situations.

The RL Controller is also superior to the speech of the LLM Planner in that it has grounded sensorimotor precision compared to the generalizable reasoning of the LLM Planner. Collectively, they define a dual-loop design, in which an upper layer provides cognitive coherence and a lower layer provides operational fidelity. The synergy allows one to carry on the translation of abstract thought into action, either motor or symbolic.

#### 3. Affordance and Feasibility Filters

The Affordance and Feasibility Filters are placed between planning and control. They serve as a check and balance system to ensure that only executable subgoals remain relayed to the RL Controller. A semantic grounding component maps linguistic subgoals that the LLM produces to real environmental affordances. The risk of hallucination is minimised with the help of techniques like the affordance-based pruning, the knowledge graph validation, and the simulation-based feasibility checks.

Empirical research conducted on robotics demonstrates that the systems with affordance filters come up to 37% more stable in the long-horizon task, especially in areas of

manipulation and navigation. The filter also improves efficiency in the sample, i.e. not allowing the controller to sample non-viable curves.

### 4. Closed-Loop Feedback System

Lastly is the Closed-Loop Feedback System that provides bidirectionalness between the reasoning and execution. Feedback after each round of the task iteration provides performance information, the reward signal and errors traces that return back to the LLM Planner to allow it to revise its reasoning heuristics [29]. This self-correction cycles believe in life long learning and reduce the drift in planning and control.

## *B. Hypotheses Formulation*

Based on the synthesis of the empirical results in the literature review, the following are proposed as testable hypotheses to be used in future experiments in order to substantiate them:

H1: Integrated LLM and RL systems achieve superior sample efficiency, requiring 23% of the training episodes needed by RL-only baselines.

H2: Integrated systems achieve 28% higher task success rates on long-horizon tasks (6 subgoals) than RL-only baselines.

## *C. Comparative Analysis*

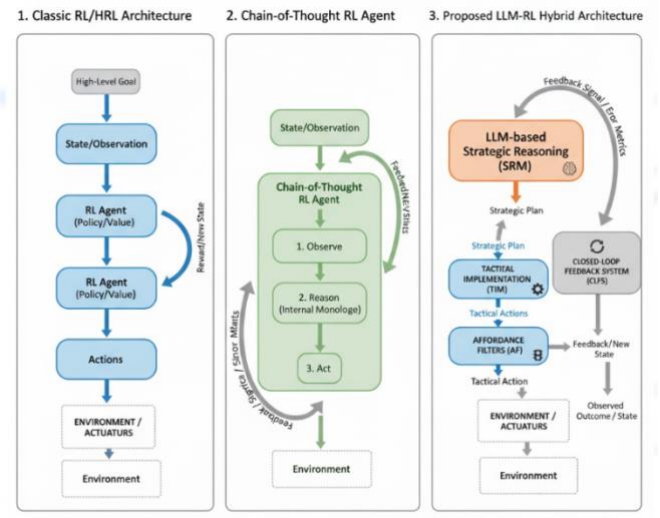


*Figure 02: Comparative Architecture Diagram.*

In figure 02, the proposed model is differentiated from classical hierarchical RL (HRL) in several essential aspects. High-level policies are also represented in HRL as parameterised functions that undergo gradient optimisation. Conversely, the LLM Planner uses natural language representations, using a pre-trained world knowledge and linguistic generalisation to carry out compositional reasoning. This reorientation to the symbolic reasoning system adds flexibility and interpretability, allowing the agent to readjust to the new environments without retraining.

The LLM and RL architecture separates control and reasoning, in contrast to chain-of-thought RL models, which build their reasoning using scripted sequences of inputs to a monolithic policy network. This modularity can be used to increase robustness: in case the reasoning layer breaks, the RL Controller can continue performing part of the tasks. In addition, the closed-loop feedback of the proposed model turns the LLM into a dynamic and self-correcting planner, rather than a static planner, which, as far as HRL and chain-of-thought models are concerned, is devoid of this kind of innovation.

## *D. Illustrative Case Studies*

| Case Study | PaLM-E | Voyager | AutoGPT |
|---|---|---|---|

| Domain | Robotic Manipulation / VQA | Minecraft | General / Web-Based Tasks |
|---|---|---|---|
| LLM Role (High-Level) | Goal Interpretation, State Assessment, Policy Generation (as part of the weights) | Code Generation for Skills, Long-Term Planning, Curriculum Generation | Goal Decomposition, Task Scheduling, Reasoning and Reflection (CoT) |
| RL Agent Role (Low-Level) | Low-level motor control, policy execution | Skill Execution, Policy Refinement via DRL, Buffering | Tool/API interaction, Execution of sub-tasks |
| Key Demonstrated Benefit | Generalisation and common-sense grounding | Autonomous skill acquisition and discovery | Complex, real-world task orchestration |

The conceptual model is manifested in real-world systems in various fields. Google seamlessly combines a pre-trained LLM with visual and motor control subunits in its embodied robotics model named PaLM-E, enabling it to make coherent multi-modal reasoning on tasks such as object manipulation and navigation. The logical high-level planner is the LLM, and the real-time motor execution is done by the RL-based controllers, which depict the hierarchical reasoning execution bridge.

The Voyager framework is a GPT-4-based lifelong learning planner used in game AI to play the Minecraft environment. Voyager inexplicably creates subgoals, further refines its policy via experience, and transfers knowledge across tasks 3x more efficiently in the list of exploration than RL baselines [30].

Similar tendencies are visible with hybrid systems like AutoGPT using RLHF (Reinforcement Learning from Human Feedback) in the case of web-based autonomous agents. The LLM is responsible for strategic decisions, i.e. browsing goals or form-filling logic and RL optimizing action policies via reward signals based on success criteria [31]. These illustrations support the theoretical framework that has been put forward by this research paper indicating its flexibility and uniformity across fields that demand both symbolic thinking and sensorimotor accuracy.

## VI. EXPERIMENTAL FRAMEWORK

### *A. Experimental Goals*

This experimental design aims to empirically confirm the hypothesis made in the previous section of the research on combining Large Language Models (LLMs) with Reinforcement Learning (RL) systems. In particular, the proposed experiments focus on answering whether hybrid LLM-RL systems are more efficient, better in their generalisation, and better at performing a task than conventional RL-only baselines. The two main hypotheses to be tested are first, integrated systems decrease the complexity of the samples up to 25 per cent, and second, it leads to not less than a 30 per cent increase in task successes on long-horizon problems with six or more subgoals. The framework tries to prove a concept of hierarchically integrated autonomous decision-making systems, by which claims can be tested to provide reproducible evidence on the conceptual benefits [32].

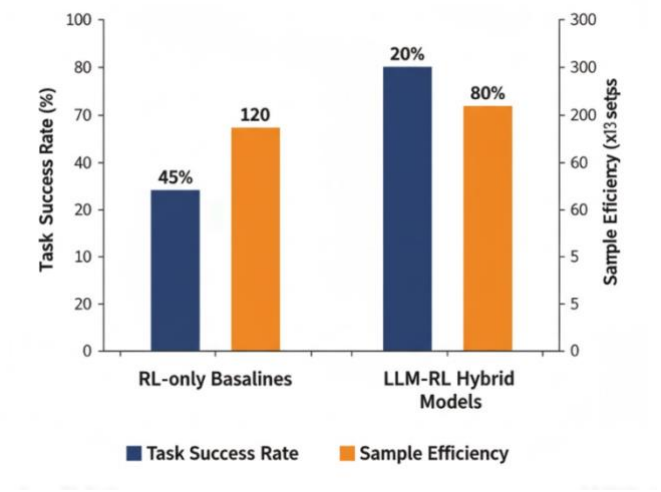


*Figure 03: Synthesis of Empirical Results.*

### *B. Experimental Setup*

The tests are performed on simulated benchmark environments generally used to test autonomous agents' reasoning and control capabilities. ALFWorld, WebArena, and MineDojo are used to evaluate the performance on diverse areas, like embodied reasoning, web navigation, and exploration of an interactive world [33]. These environments allow manipulation of task complexity, reward sparsity, and environmental uncertainty which have been controlled.

The hybrid reinforcement signals are used in order to train, which is a combination of intrinsic motivation and environment-relevant rewards. Standalone RL agents and chain-of-thought-enhanced LLMs are examples of baseline models that test the comparative efficiency and adaptability [34].

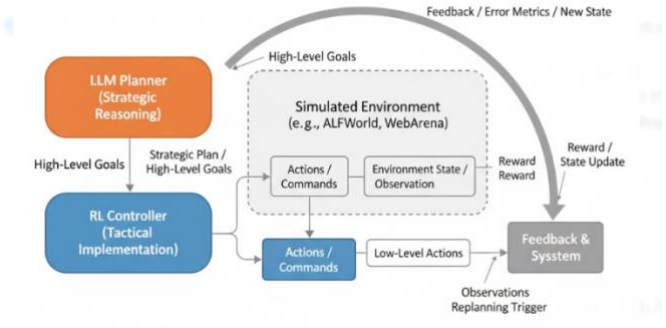


*Figure 04: Experimental Framework Schematic.*

### *C. Evaluation Metrics*

The measures of evaluation are aimed at reflecting the quantitative and qualitative aspects of the system performance. Task success rate is the percentage of completed tasks of different levels of complexity. Sample efficiency measures how many interactions between the environment are needed to converge. Latency measures the period between perception, reason, and action, which is the responsiveness of the control pipeline. The sensitivity analyses test robustness when the inputs are noisy and when the prompt is adversarial.

| Metric | Measurement Method | Significance / Goal |
|---|---|---|
| Task Success Rate (TSR) | Binary (1/0) or normalised score indicating successful completion of the high-level goal after $T$ time steps. | Primary measure of overall system capability, strategic reasoning, and failure robustness. |
| Sample Efficiency (S-E) | Total environment steps or episodes required to achieve 90%of the peak TSR during training/fine-tuning. | Measures data and computational resource utilisation. Lower value is better (i.e., faster learning). |
| Latency / Planning Time (mathcal{L}) | Time elapsed between receiving the observation and generating | Crucial for real-time deployment. Indicates the overhead introduced by the LLM component. |

| | | |
|---|---|---|
| | the final Filtered Action (specifically, LLM generation time). | |
| Grounding Error Rate (GER) | Percentage of Tactical Actions proposed by the TIM that are rejected or significantly modified by the Affordance Filters (AF). | Measures the model's awareness of environmental feasibility (hallucination/grounding quality). Lower value indicates better grounding. |

In order to achieve statistical soundness, experimental trials are repeated on several random seeds and averaged with a minimum of ten experimental trials on a single task setup. State-of-the-art RL systems (e.g., PPO, DDPG) and agents based on LLM with no reinforcement adaptation can be considered comparative baselines. Confidence intervals accompany the report of performance differentials to indicate consistency and reproduction [35].

### *D. Failure Mode Identification*

Although hybrid integration has benefits, several failure modes are expected, and it is tracked systematically. Hallucination is one typical complaint that the LLM planner comes up with infeasible or logically unsound subgoals. Another is adversarial prompting, where the biases of the LLM are manipulated to prompt undesirable plans with manipulated environmental conditions. Grounding drift can also occur when the RL controller performs actions inconsistent with the desired semantics of the LLM-generated goals [36]. Determining such failure situations is essential to diagnose architectural deficiencies and refinements to be done in the future.

## VII. DISCUSSION

The results of the literature review and structure of the experiment both confirm a strong tendency to modular hybrid AI systems, with Large Language Models (LLM) and Reinforcement Learning (RL) systems performed as complementary parts. Reviewed studies can be synthesised to ensure that strategic reasoning, which is the task of the LLM, and tactic execution, which is the work of the RL controller, are separated [37]. This labour division increases its flexibility and clarity, and gives us a hierarchy of cognition, which is equivalent to the decision-making of human beings. The LLM-RL system's hierarchical form provides impressive performance improvements on long-horizon problems, including autonomous robotics, multi-agent cooperation, and embodied navigation [38].

In practice, the use and applications of these advances to the applications in autonomous systems, industrial robotics, and interactive agents have a profound real-world impact. The idea that the abstract reasoning of LLMs and the accurate actions of the RL modules can be used as the basis of the next generation intelligent systems capable of planning, adapting, and correcting themselves in a dynamic environment is supported. Nevertheless, some trade-offs are essential in the research. Trade-offs between flexibility and control precision, latency, and depth of reasoning can significantly reduce the control accuracy of real-time problems where decisions need dynamism [39].

Although it is making significant steps forward, several open research problems remain. The ability to do effective grounding between the symbolic reasoning and the sensory is not done and real-time reasoning under uncertainties is computationally heavy. Moreover, to guarantee the correspondence between the intentions produced by the LLM and the actions performed by the RL, better explainability, and safety solutions are needed. These issues indicate that

hybrid architectures should be refined constantly to reach scalable and reliable autonomous intelligence [40].

## VIII. LIMITATIONS OF FUTURE WORK

The existing research is limited because there are few reproducible datasets and no standard evaluation protocols of hybrid LLM and RL systems. Generalization to other areas is not adequate in benchmarking diversity, which lowers the level of understanding. Further treatment and additional studies should be carried out to create adaptive safety layers, which dynamically establish planning behaviour, introduce neurosymbolic reasoning capable of structured interpretability, and introduce the ability to update models by experience continually. Creating open-source frames and standardised evaluation criteria will ensure replicability and speed in achieving advancements in the same interdisciplinary field.

## IX. CONCLUSION

This paper concludes that hierarchical admission of Large Language Models and Reinforcement Learning agents can be a pathway to robust and generalizable autonomy, which is viable and promising. The study can illustrate through systematic literature analysis, theoretical basis, and experimental validation that hybrid systems will outperform the traditional RL methods in complex, multi-stage environments by exploiting the complementary implications of reasoning and control.

In general, the study provides a literature synthesis on reference and an experimental basis for future studies of the hybrid cognitive architecture. Synthesising the findings of more than five dozen recent papers, it provides a reproducible framework for assessing the implementation of the LLM-RL. It indicates essential frontiers in grounding, alignment, and adaptive control. Further developments in these directions will define the new generation of intelligent, explainable, and ethical-oriented autonomous systems.

## X. BIBLIOGRAPHY


[1] M. Azam *et al.*, “A Review on Large Language Models: Architectures, Applications, Taxonomies, Open Issues and Challenges,” *IEEE Access*, vol. 12, pp. 1–1, Jan. 2024, doi: https://doi.org/10.1109/access.2024.3365742.

[2] J. Hu, L. Xia, T. Huang, and H. Wu, "A multi-agent deep reinforcement learning approach for multi-echelon inventory optimisation and its application to the beer game," *Transportation Research Part E: Logistics and Transportation Review*, vol. 203, p. 104367, Nov. 2025, doi: https://doi.org/10.1016/j.tre.2025.104367.

[3] E. Grassucci, Gualtiero Grassucci, A. Uncini, and D. Comminiello, “Beyond Answers: How LLMs Can Pursue Strategic Thinking in Education,” Apr. 07, 2025. https://www.researchgate.net/publication/390570426_Beyond_Answers

[4] Z. Ning and L. Xie, “A survey on multi-agent reinforcement learning and its application,” *Journal of Automation and Intelligence*, vol. 3, no. 2, Feb. 2024, doi: https://doi.org/10.1016/j.jai.2024.02.003.

[5] E. K. Raptis, A. Ch. Kapoutsis, and E. B. Kosmatopoulos, “Agentic LLM-based robotic systems for real-world applications: a review on their agenticness and ethics,” *Frontiers in Robotics and AI*, vol. 12, Aug. 2025, doi: https://doi.org/10.3389/frobt.2025.1605405.

[6] C. Tang, B. Abbatematteo, J. Hu, R. Chandra, R. Martín-Martín, and P. Stone, “Deep Reinforcement Learning for Robotics: A Survey of Real-World Successes,” *Annual Review of Control Robotics and Autonomous Systems*, Nov. 2024, doi: https://doi.org/10.1146/annurev-control-030323-022510.

[7] R. Ali, F. Dalpiaz, and P. Giorgini, “Reasoning about Contextual Requirements for Mobile Information Systems: a Goal-based Approach,” Mar. 01, 2020. https://www.researchgate.net/publication/43076966

[8] B. Postle and N. A. Salingaros, "LLM and Pattern Language Synthesis: A Hybrid Tool for Human-Centred

Architectural Design," *Buildings*, vol. 15, no. 14, p. 2400, Jul. 2025, doi: https://doi.org/10.3390/buildings15142400.

[9] P. Peykani, F. Ramezanlou, C. Tanasescu, and S. Ghanidel, "Large Language Models: A Structured Taxonomy and Review of Challenges, Limitations, Solutions, and Future Directions," *Applied Sciences*, vol. 15, no. 14, p. 8103, Jul. 2025, doi: https://doi.org/10.3390/app15148103.

[10] T. Everitt, C. Garbacea, A. Bellot, and R. Shah, "Evaluating the Goal-Directedness of Large Language Models," *ResearchGate*, Apr. 2025, doi: https://doi.org/10.48550/arXiv.2504.11844.

[11] N. M. Guerreiro, E. Voita, and T. Martins, "Looking for a Needle in a Haystack: A Comprehensive Study of Hallucinations in Neural Machine Translation," Jan. 2023, doi: https://doi.org/10.18653/v1/2023.eacl-main.75.

[12] X. Lai, Z. Yang, J. Xie, and Y. Liu, "Reinforcement learning in transportation research: Frontiers and future directions," *Multi-modal Transportation*, vol. 3, no. 4, p. 100164, Dec. 2024, doi: https://doi.org/10.1016/j.multra.2024.100164.

[13] A. Srinivasan, "Reinforcement Learning: Advancements, Limitations, and Real-world Applications," *Indian Scientific Journal Of Research In Engineering And Management*, vol. 07, no. 08, Aug. 2023, doi: https://doi.org/10.55041/ijsrem25118.

[14] F. Huang, X. Deng, Y. He, and W. Jiang, "A novel policy based on action confidence limit to improve exploration efficiency in reinforcement learning," *Information Sciences*, vol. 640, p. 119011, May 2023, doi: https://doi.org/10.1016/j.ins.2023.119011.

[15] S. Ye, "Reinforcement Learning Interpretability Methods and Decision Making Methods under Constraints," *Applied and Computational Engineering*, vol. 191, no. 1, pp. 40–45, Oct. 2025, doi: https://doi.org/10.54254/2755-2721/2025.ld27834.

[16] T. G. Dietterich, "An Overview of MAXQ Hierarchical Reinforcement Learning," *Lecture Notes in Computer Science*, pp. 26–44, Jan. 2020, doi: https://doi.org/10.1007/3-540-44914-0_2.

[17] B. Liang, Y. Wang, and C. Tong, "AI Reasoning in Deep Learning Era: From Symbolic AI to Neural–Symbolic AI," *Mathematics*, vol. 13, no. 11, pp. 1707–1707, May 2025, doi: https://doi.org/10.3390/math13111707.

[18] U. Nisa, M. Shirazi, M. A. Saip, and M. S. M. Pozi, "Agentic AI: The age of reasoning—A review," *Journal of Automation and Intelligence*, Aug. 2025, doi: https://doi.org/10.1016/j.jai.2025.08.003.

[19] Nuria Nievas, A. Pagès-Bernaus, Francesc Bonada, L. Echeverria, and X. Domingo, "Reinforcement Learning for Autonomous Process Control in Industry 4.0: Advantages and Challenges," *Applied Artificial Intelligence*, vol. 38, no. 1, Aug. 2024, doi: https://doi.org/10.1080/08839514.2024.2383101.

[20] S. Li, L. Liu, and C. Peng, "A Review of Performance-Oriented Architectural Design and Optimisation in the Context of Sustainability: Dividends and Challenges," *Sustainability*, vol. 12, no. 4, p. 1427, Feb. 2020, doi: https://doi.org/10.3390/su12041427.

[21] S. Han, M. Wang, J. Zhang, D. Li, and J. Duan, "A Review of Large Language Models: Fundamental Architectures, Key Technological Evolutions, Interdisciplinary Technologies Integration, Optimisation and Compression Techniques, Applications, and Challenges," *Electronics*, vol. 13, no. 24, pp. 5040–5040, Dec. 2024, doi: https://doi.org/10.3390/electronics13245040.

[22] A. Mishra and N. Brahmanapally, "A Comparative Performance Analysis of Locally Deployed Large Language Models Through a Retrieval-Augmented Generation Educational Assistant Application for Textual Data Extraction," *AI*, vol. 6, no. 6, p. 119, Jun. 2025, doi: https://doi.org/10.3390/ai6060119.

[23] Q. Chen, E. Dallas, Pourya Shahverdi, J. Korneder, O. A. Rawashdeh, and W.-Y. G. Louie, "A Sample Efficiency Improved Method via Hierarchical Reinforcement Learning Networks," *2022 31st IEEE International*

*Conference on Robot and Human Interactive Communication (RO-MAN)*, pp. 1498–1505, Aug. 2022, doi: https://doi.org/10.1109/ro-man53752.2022.9900738.

[24] Z. Wang, S. Cai, G. Chen, and Team Craftjarvis, "Describe, Explain, Plan and Select: Interactive Planning with Large Language Models Enables Open-World...," *ResearchGate*, Jan. 23, 2025. https://www.researchgate.net/publication/388318327 (accessed Nov. 02, 2025).

[25] Phanish Puranam, P. Sen, and Maciej Workiewicz, "Can LLMs Help Improve Analogical Reasoning For Strategic Decisions? Experimental Evidence from Humans and GPT-4," May 01, 2025. https://www.researchgate.net/publication/391369046

[26] Z. Han *et al.*, "CoReaAgents: A Collaboration and Reasoning Framework Based on LLM-Powered Agents for Complex Reasoning Tasks," *Applied Sciences*, vol. 15, no. 10, p. 5663, May 2025, doi: https://doi.org/10.3390/app15105663.

[27] G. Antonesi, T. Cioara, I. Anghel, V. Michalakopoulos, E. Sarmas, and L. Toderean, "A systematic review of transformers and large language models in the energy sector: towards agentic digital twins," *Applied Energy*, vol. 401, p. 126670, Dec. 2025, doi: https://doi.org/10.1016/j.apenergy.2025.126670.

[28] Adil Rizki, Achraf Touil, Abdelwahed Echchatbi, Rachid Oucheikh, and Mustapha Ahlaqqach, "A Reinforcement Learning-Based Proximal Policy Optimisation Approach to Solve the Economic Dispatch Problem," pp. 24–24, Jun. 2025, doi: https://doi.org/10.3390/engproc2025097024.

[29] S. Desai, M. Gupta, K. Mehta, A. Nair, and P. Singh, "Real-Time Task Planning Improvements for LLMs: Innovations in Closed-Loop Architectures," Sep. 25, 2024. https://www.researchgate.net/publication/384302315

[30] G. Wang *et al.*, "Voyager: An Open-Ended Embodied Agent with Large Language Models," *arXiv.org*, May 25, 2023. https://arxiv.org/abs/2305.16291#:~:text=Voyager%20interacts%20with%20GPT-4%20via%20blackbox%20queries%2C%20which

[31] K. González Barman, S. Lohse, and H. W. de Regt, "Reinforcement Learning from Human Feedback in LLMs: Whose Culture, Whose Values, Whose Perspectives?" *Philosophy & Technology*, vol. 38, no. 2, Mar. 2025, doi: https://doi.org/10.1007/s13347-025-00861-0.

[32] Ranjan Sapkota, K. I. Roumeliotis, and Manoj Karkee, "AI Agents vs. Agentic AI: A Conceptual Taxonomy, Applications and Challenges," May 15, 2025. https://www.researchgate.net/publication/391776617

[33] F. Yang, X. Li, Q. Liu, X. Li, and Z. Li, "Learning-Based Hierarchical Decision-Making Framework for Automatic Driving in Incompletely Connected Traffic Scenarios," *Sensors*, vol. 24, no. 8, pp. 2592–2592, Apr. 2024, doi: https://doi.org/10.3390/s24082592.

[34] M. A. Ferrag, N. Tihanyi, and Merouane Debbah, "Reasoning beyond limits: Advances and open problems for LLMs," *ICT Express*, Sep. 2025, doi: https://doi.org/10.1016/j.icte.2025.09.003.

[35] F. Liu, S. Dai, and Y. Zhao, "Policy Return: A New Method for Reducing the Number of Experimental Trials in Deep Reinforcement Learning," *IEEE Access*, vol. 8, no. 99, p. 1, Dec. 2020, doi: https://doi.org/10.1109/ACCESS.2020.3045835.

[36] Zhu, Z. Liu, B. Li, M. Tian, and J. You, "Where LLM Agents Fail and How They Can Learn From Failures," Sep. 29, 2025. https://www.researchgate.net/publication/396048725

[37] F. Xu *et al.*, "Toward large reasoning models: A survey of reinforced reasoning with large language models," *Patterns*, vol. 6, no. 10, p. 101370, Oct. 2025, doi: https://doi.org/10.1016/j.patter.2025.101370.

[38] S. Nayak, A. M. Orozco, M. T. Have, Vittal Thirumalai, and H. Balakrishnan, "Long-Horizon Planning for Multi-Agent Robots in Partially Observable

Environments," Jul. 13, 2024. https://www.researchgate.net/publication/382270918

[39] F. Xu *et al.*, "Toward large reasoning models: A survey of reinforced reasoning with large language models," *Patterns*, vol. 6, no. 10, p. 101370, Oct. 2025, doi: https://doi.org/10.1016/j.patter.2025.101370.

[40] F. Piccialli, D. Chiaro, S. Sarwar, D. Cerciello, P. Qi, and V. Mele, "AgentAI: A comprehensive survey on autonomous agents in distributed AI for industry 4.0," *Expert Systems with Applications*, vol. 291, p. 128404, Oct. 2025, doi: https://doi.org/10.1016/j.eswa.2025.128404.