ISSN: 3079-6342

Uncertainty-Aware Reinforcement Learning for Robust Decision Making in LLM-Agent Collaboration

Maximilian R. Müller[1], Tobias F. Weber[2], Anna K. Schneider[3*]

[1,2,3] Department of Informatics, University of Heidelberg, 69120 Heidelberg, Germany

***Corresponding author:** a.schneider@uni-heidelberg.de

Abstract

Decision-making processes involving LLM agents are often susceptible to uncertainty arising from ambiguous inputs and stochastic generation behavior. To address this issue, this study proposes an uncertainty-aware reinforcement learning framework that incorporates Bayesian reward estimation and entropy-regularized policy updates. The approach is validated on a benchmark consisting of 9,600 tasks with varying levels of input ambiguity, including incomplete instructions and conflicting objectives. Results indicate that the proposed method reduces error propagation by 28.7% and improves decision robustness, with task success rates increasing from 68.9% to 80.2%. Additionally, calibration metrics such as expected calibration error (ECE) decrease by 19.5%, demonstrating improved reliability in agent outputs. The framework provides a systematic solution for enhancing robustness in LLM-based collaborative systems under uncertainty.

Keywords

Uncertainty modeling; Reinforcement learning; LLM agents; Robust decision-making; Bayesian optimization

1. Introduction

Large language models (LLMs) are no longer limited to single-turn text generation and are increasingly deployed as interactive agents capable of planning, action execution, tool use, and sequential decision-making in open environments. This transition has been supported by early alignment studies showing that model behavior can be improved through instruction tuning and reinforcement learning from human feedback, as well as by agent frameworks that connect language models with tools, external memory, and task environments [1]. Recent research has further shown that explicitly structured post-retrieval inference can improve the organization, traceability, and reliability of intermediate reasoning processes, highlighting the importance of structured decision pathways for controllable model behavior in complex settings [2]. Survey studies also note that LLM agents are increasingly viewed not merely as generators of text, but as decision-making systems whose effectiveness depends on long-horizon planning, environment interaction, and performance stability under changing conditions [3]. This shift has broadened the scope of LLM research from response quality alone to the broader problem of how language-driven systems can maintain robust behavior

ISSN: 3079-6342

across sequential and uncertain tasks. An important extension of this development is multi-agent collaboration. Rather than relying on a single model to solve an entire problem, multi-agent systems distribute subtasks across multiple agents with different responsibilities, allowing decision-making to be decomposed into smaller and potentially more manageable components. Early frameworks such as CAMEL and AutoGen showed that role assignment and structured communication can improve task decomposition, coordination, and cooperation efficiency [4]. Subsequent studies extended this idea by organizing agents into workflow-based systems in which different agents handle specific subtasks within a structured process [5]. Dynamic collaboration has also received increasing attention, with some frameworks allowing agent selection and communication patterns to be adjusted during execution rather than being fixed in advance [6]. Recent surveys summarize that role design, communication protocols, and coordination control remain central challenges in multi-agent LLM systems [7]. These findings suggest that collaboration offers clear advantages for complex tasks, but they also indicate that effective cooperation depends heavily on how interaction structures are designed and controlled. Despite these advances, reliability remains a major concern in LLM-agent collaboration. In practical tasks, uncertainty may arise from ambiguous instructions, incomplete observations, conflicting goals, noisy external feedback, and intrinsic randomness in model generation. Studies on model calibration have shown that LLM outputs may appear highly confident even when they are incorrect, which makes it difficult to trust downstream decisions based only on surface plausibility [8]. More recent work on uncertainty estimation suggests that LLM-based systems introduce additional sources of instability, including variability in reasoning paths, sensitivity to prompt conditions, and inconsistency across multi-step generation processes [9]. These problems become more serious in multi-agent settings because individual errors do not remain isolated. An early misinterpretation or overconfident decision by one agent may influence later reasoning, alter communication patterns, and propagate through the collaborative process until the final outcome is affected. As a result, uncertainty in multi-agent systems is not only a property of isolated outputs, but also a dynamic factor that shapes coordination quality over time. Recent studies have therefore begun to examine uncertainty during the decision process rather than only at the final output stage. In tool-based agent settings, existing work has shown that prompt selection and execution paths can be distorted by miscalibration, and that better calibration improves downstream task performance [10]. For multi-step agents, studies on uncertainty propagation have reported that uncertainty accumulates along the decision trajectory and cannot be adequately measured from the final step alone [11]. Step-level calibration methods

ISSN: 3079-6342

have similarly shown that correcting intermediate errors can improve performance in long-horizon tasks by preventing local failures from becoming trajectory-level deviations [12]. Together, these findings indicate that uncertainty should be treated as an evolving property of the full decision process rather than as a post hoc confidence issue. This perspective is particularly relevant for collaborative agent systems, where multi-step dependencies and inter-agent interactions create more opportunities for local uncertainty to amplify into broader coordination failure. Reinforcement learning (RL) provides a suitable methodological foundation for addressing this problem because it optimizes decisions over full trajectories rather than isolated outputs. Recent studies show a clear transition from prompt-dominated control strategies toward policy learning based on repeated interaction with environments [13]. By using reward signals to guide behavior, RL offers a natural way to improve adaptation, long-horizon planning, and decision consistency in interactive tasks. Even so, current RL-based approaches still face several limitations in LLM-agent settings. Many environments provide sparse, delayed, or unstable reward signals, which makes training difficult and often reduces policy reliability. A number of existing methods also depend on relatively limited datasets, expert demonstrations, or narrow task settings, which restricts generalization to new environments and more ambiguous conditions [14]. Related studies have shown that models trained on narrow experience often fail to maintain stable performance when goals shift, observations become incomplete, or decision contexts become more uncertain. These observations suggest that effective learning in LLM-agent systems requires not only policy optimization, but also reward designs and update mechanisms that explicitly account for uncertainty. Benchmark studies support the same conclusion. LLM-Coordination showed that current models still struggle in tasks requiring shared reasoning, synchronized decision-making, and stable cooperation [15]. MultiAgentBench extended this evaluation to a wider range of task types and found that final task accuracy alone does not adequately capture collaboration quality. Existing surveys further note that many current benchmarks remain limited in scale, environmental variation, or task diversity, making it difficult to evaluate robustness under uncertain or dynamically changing conditions. This limitation is nontrivial because coordination failures in realistic systems are often caused not by a single catastrophic decision, but by the accumulation of small errors across multiple steps and interacting agents. When uncertainty is not explicitly represented or controlled, early deviations can spread through communication and influence later decisions in ways that are difficult to detect and correct. Several important gaps therefore remain in the current literature. Many multi-agent systems emphasize coordination performance but do not explicitly model uncertainty, which

ISSN: 3079-6342

weakens robustness when inputs are ambiguous, information is incomplete, or agent goals are not perfectly aligned. Existing calibration methods mainly focus on single outputs or local intermediate steps, while uncertainty-aware policy learning in multi-agent settings remains relatively underexplored [16]. Reinforcement learning has improved adaptability in interactive environments, yet only limited work combines policy learning with formal uncertainty estimation in settings that vary in ambiguity, conflict, and sequential dependency [17]. More broadly, current studies often treat uncertainty analysis, coordination control, and policy optimization as separate problems, which makes it difficult to develop a unified account of robust decision-making in collaborative LLM systems. These limitations indicate the need for a framework in which uncertainty is incorporated directly into the learning process and used to guide multi-agent decision behavior throughout the trajectory. To address these issues, this study proposes an uncertainty-aware reinforcement learning framework for robust decision-making in LLM-agent collaboration. The proposed method combines Bayesian reward estimation with entropy-based policy updates so that learning takes into account not only expected outcomes, but also uncertainty arising during the decision process. In contrast to approaches that assess confidence only after an answer is produced, the proposed framework treats uncertainty as a trajectory-level signal that influences policy improvement, coordination behavior, and error control during interaction. This design aims to reduce error propagation, improve calibration quality, and strengthen robustness in collaborative decision-making under ambiguous and conflicting conditions. The framework is evaluated on 9,600 tasks with varying levels of input ambiguity and goal conflict, allowing a more systematic analysis of how uncertainty affects multi-agent behavior across different settings. The significance of this study lies in both methodology and application. From a methodological perspective, it provides a unified way to connect uncertainty estimation with reinforcement learning for collaborative LLM agents. From an application perspective, it offers a more robust decision framework for multi-agent systems operating in environments where ambiguity, conflict, and sequential dependence are unavoidable. By integrating uncertainty directly into policy learning, this work seeks to advance the development of more reliable, stable, and interpretable LLM-agent collaboration.

ISSN: 3079-6342

2.Materials and Methods

2.1 Sample and Study Setting

The study used a synthetic multi-agent decision dataset designed to represent uncertainty in collaborative tasks. A total of 9,600 task instances were generated and divided into three levels of input ambiguity: low, medium, and high. Each task included incomplete instructions, missing information, or conflicting objectives to simulate uncertain conditions. The environment was defined as a step-based interaction process, where multiple language model agents operated with partial information. Each task involved 3–5 agents with fixed roles, including planner, verifier, and executor. Task difficulty was adjusted by changing the number of constraints, the level of ambiguity, and the number of interaction steps.

2.2 Experimental Design and Control Setup

The experiments included one proposed method and two baseline settings. The proposed method used uncertainty-aware reinforcement learning with Bayesian reward estimation and entropy-based policy updates. The first baseline used fixed prompts without policy learning. The second baseline used reinforcement learning with standard reward signals but without uncertainty modeling. All methods were tested under the same task conditions. Each experiment was repeated five times with different random seeds to reduce random effects. Performance was evaluated using task success rate, error propagation rate, and calibration metrics.

2.3 Measurement Methods and Quality Control

Task success was defined as the proportion of tasks that met all constraints and objectives. Error propagation was measured as the number of incorrect intermediate decisions that led to final failure. Calibration performance was evaluated using expected calibration error, which measures the difference between predicted confidence and actual accuracy. To ensure stable results, all experiments used the same model settings, including decoding parameters and input format. Prompts were kept consistent across all runs. Detailed logs were recorded for each step, including agent actions and intermediate states. Invalid or incomplete runs were removed based on predefined rules.

2.4 Data Processing and Model Formulation

The collected data were processed to calculate performance and calibration metrics. Task success rate was calculated as:

ISSN: 3079-6342

$$SR=\frac{N_{success}}{N_{total}}$$

Where $N_{success}$ is the number of successful tasks and N_{total} is the total number of tasks. Expected calibration error was calculated as:

$$ECE=\sum_{k=1}^{K}\frac{|B_k|}{N}|acc(B_k)\text{-}conf(B_k)|$$

Where B_k is the set of predictions in bin k, $acc(B_k)$ is the average accuracy, and $conf(B_k)$ is the average confidence. Policy learning used an entropy-based objective to support stable decisions under uncertainty:

$$L(\theta)=E[R_t]+\lambda H(\pi_\theta)$$

Where R_t is the reward, H is the policy entropy, and λ is a weighting factor. All metrics were normalized before comparison.

2.5 Statistical Analysis

Statistical analysis was used to compare results across methods and uncertainty levels. Mean values and standard deviations were calculated for all metrics across repeated runs. Differences between methods were tested using independent two-sample t-tests with a significance level of $p<0.05$. Effect size was measured using Cohen's d. In addition, 95% confidence intervals were calculated to assess result stability. Standard statistical tools were used, and assumptions of normal distribution and equal variance were checked before testing.

3.Results and Discussion

3.1 Overall performance under uncertain task conditions

The proposed framework showed better overall performance across the full benchmark. The task success rate increased from 68.9% to 80.2%, and error propagation decreased by 28.7%. Expected calibration error also decreased by 19.5%, which shows better agreement between model confidence and actual decision quality. These results indicate that uncertainty-aware policy learning improved both final task outcomes and the reliability of intermediate decisions. This result is consistent with recent studies showing that multi-agent LLM systems become less stable when task conditions are unclear and when cooperation depends on repeated information exchange. Recent review studies have also pointed out that role design and coordination structure remain key factors in collaborative LLM performance [18,19]. A related agent workflow is shown in Fig.1, where planning, tool use, and reflection are assigned to separate modules within a shared reasoning process.

ISSN: 3079-6342

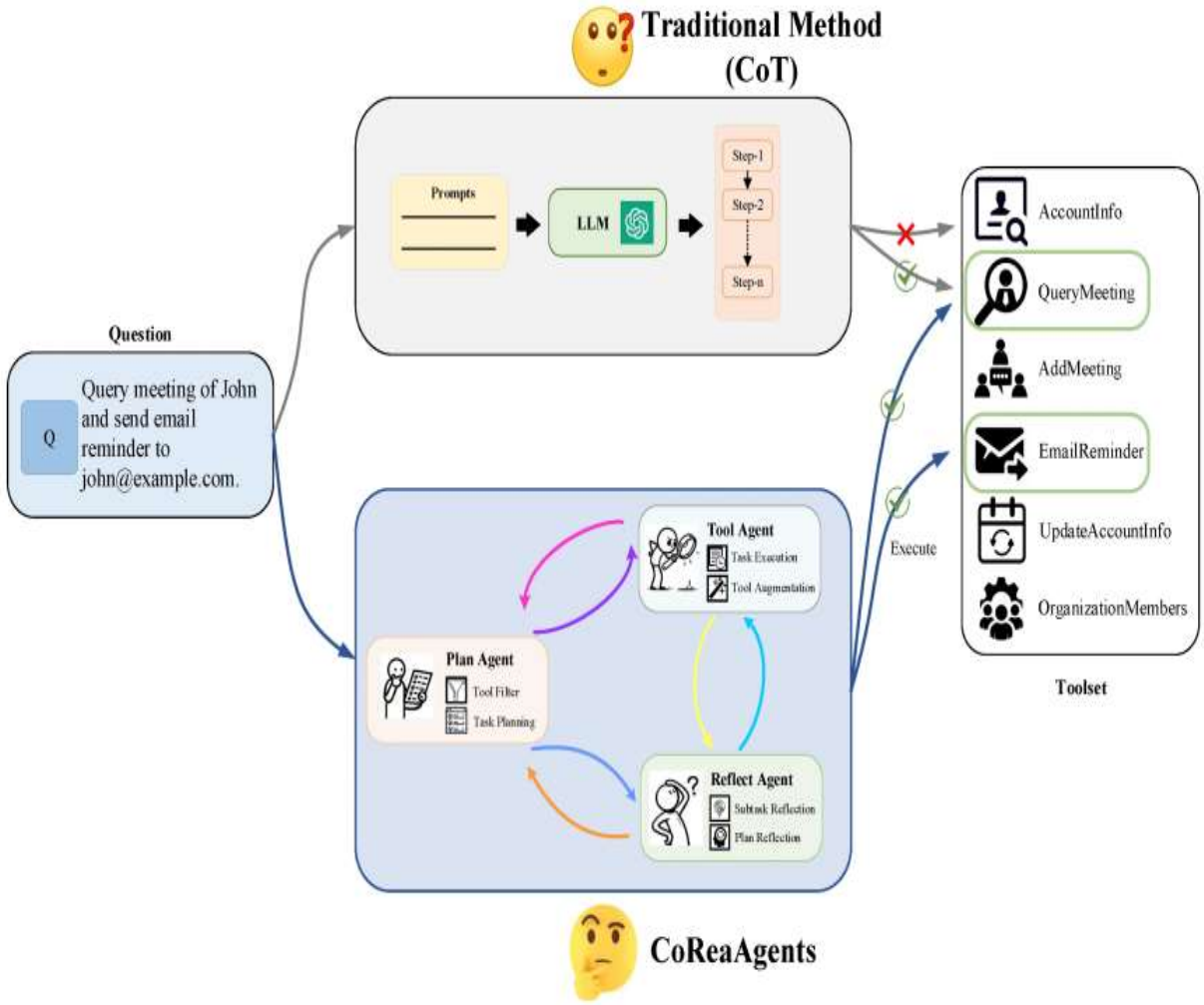
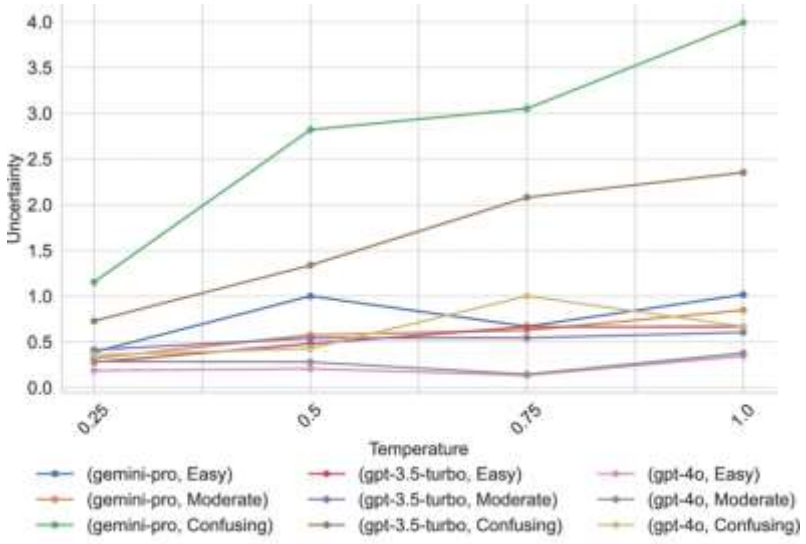


Figure 1 Multi-agent collaboration framework with role-based modules for planning, tool use, and reflection.

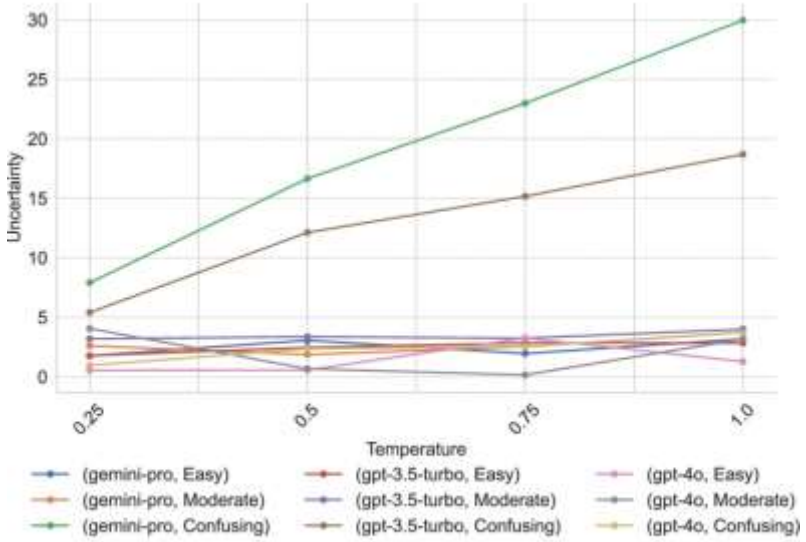
3.2 Effects of ambiguity level on decision quality

Performance changed with the level of input ambiguity. The improvement of the proposed method was largest in tasks with incomplete instructions and conflicting goals, while the gain was smaller in low-ambiguity settings. This pattern is reasonable because unclear inputs increase the chance that an early wrong action will affect later steps. In such cases, standard prompt-based systems often produce responses that seem reasonable at the local level but are not consistent with the overall task. The present findings support recent work showing that uncertainty in LLM systems is closely related to prompt complexity and response variation [20,21]. A comparable pattern is illustrated in Fig.2, where uncertainty values increase with prompt difficulty and generation temperature across several language models.

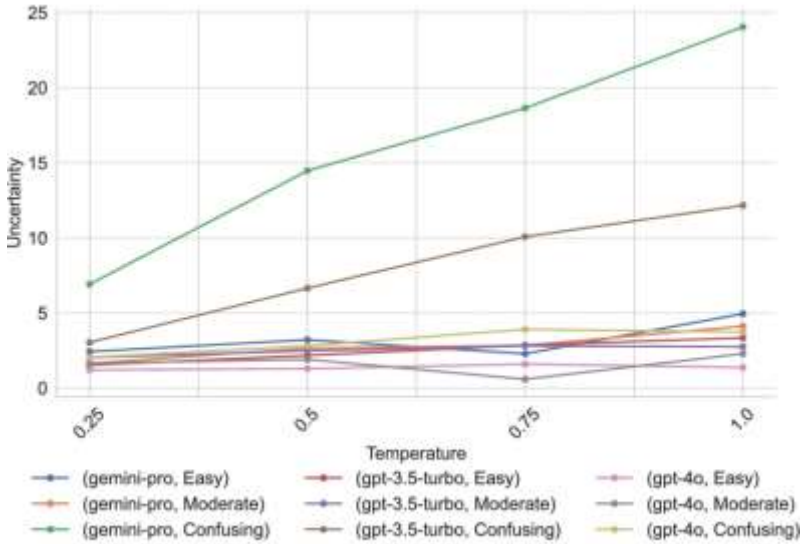
ISSN: 3079-6342



(a) PCA



(b) Isomap



(c) MDS

Figure 2 Changes in model uncertainty under different prompt types and temperature settings.

ISSN: 3079-6342

3.3 Robustness and calibration behavior

The decrease in expected calibration error is important because robust collaboration depends on more than task completion alone. A system may reach the correct answer but still show poor reliability if intermediate confidence does not match true correctness. In the present study, the lower calibration error indicates that the proposed framework produced more reliable decision signals during collaboration. This result agrees with recent studies showing that uncertainty in LLM systems should be examined across the full decision process rather than only at the final output stage. Research on process-level uncertainty has shown that ambiguity, decoding randomness, and variation in reasoning paths can build up over multiple steps and reduce system reliability [22]. The current results extend this view by showing that reinforcement learning can reduce this problem when reward estimation includes uncertainty information directly.

3.4 Comparison with existing studies and remaining limitations

Compared with recent studies, the present framework has two practical strengths. It was tested on 9,600 tasks with different ambiguity levels, which provides broader coverage than many recent studies that focus on a single workflow or a limited set of prompts. It also evaluated both task performance and calibration quality, which gives a more complete view of system robustness. However, several limitations remain. The benchmark is synthetic, and real decision settings may include noisier inputs, longer dependencies, and more complex conflicts among agents. In addition, the current design used fixed agent roles, so it did not test dynamic role switching during execution. Recent reviews of LLM-agent systems have identified uncertainty handling, adaptive coordination, and scalable evaluation as open issues, and the present study still falls within these broader limits. Future work should test the framework in real interactive environments and examine whether the same uncertainty-aware strategy remains effective when tool use, memory load, and team structure become more complex.

4.Conclusion

This study evaluated an uncertainty-aware reinforcement learning framework for multi-agent large language model systems under ambiguous and conflicting task conditions. The results show clear improvements in task success, error propagation, and calibration performance when uncertainty is included in the learning process. The use of Bayesian reward estimation together with entropy-based policy updates helps agents make more stable decisions and reduces the effect of early errors during multi-step interaction. This approach differs from standard reinforcement learning methods that rely only on expected rewards, and it provides

ISSN: 3079-6342

a more reliable way to guide agent behavior in uncertain environments. The findings suggest that uncertainty-aware policy learning can improve both decision quality and consistency in collaborative systems. The framework has potential use in applications that require reliable decision-making under uncertainty, such as automated planning, negotiation systems, and multi-agent coordination tasks. It can also support cases where inputs are incomplete or goals are not fully aligned. However, some limitations remain. The experiments were conducted on a synthetic benchmark, and real-world environments may include more noise and longer dependencies. In addition, the current design uses fixed agent roles and does not consider dynamic role adjustment during execution. Future work should test the method in real applications and explore more flexible agent structures to improve robustness and generalization.

References

[1] Kaufmann, T., Weng, P., Bengs, V., & Hüllermeier, E. (2024). A survey of reinforcement learning from human feedback. Transactions on Machine Learning Research.

[2] Xu, D., Liu, H., Qiu, D., & Ma, Q. (2026). Structured Modeling and Representation Methods for Post-Retrieval Inference Processes in Large Video Language Models.

[3] Aratchige, R. M., & Ilmini, W. M. K. S. (2025). Llms working in harmony: A survey on the technological aspects of building effective llm-based multi agent systems. arXiv preprint arXiv:2504.01963.

[4] Qiu, Y. (2024). Estimation of tail risk measures in finance: Approaches to extreme value mixture modeling. arXiv preprint arXiv:2407.05933.

[5] Bluethgen, C., Van Veen, D., Truhn, D., Kather, J. N., Moor, M., Polacin, M., ... & Nooralahzadeh, F. (2025). Agentic Systems in Radiology: Design, Applications, Evaluation, and Challenges. arXiv preprint arXiv:2510.09404.

[6] Krishnan, N. (2025). Advancing multi-agent systems through model context protocol: Architecture, implementation, and applications. arXiv preprint arXiv:2504.21030.

[7] Liu, S., Liu, X., & Feng, H. (2025, November). Research on AI-Driven Visual Design and Immersive Interactive Experiences Based on Multimodal Cognition and User. In Proceedings of the 2025 International Conference on Digital Society and Intelligent Computing (pp. 734-740).

[8] Steyvers, M., Tejeda, H., Kumar, A., Belem, C., Karny, S., Hu, X., ... & Smyth, P. (2024). The calibration gap between model and human confidence in large language models. arXiv preprint arXiv:2401.13835.

[9] Yue, L., Xu, D., Qiu, D., Shi, Y., Xu, S., & Shah, M. (2025, December). Sequential Cooperative Multi-Agent Online Learning and Adaptive Coordination Control in Dynamic and Uncertain

ISSN: 3079-6342

Environments. In 2025 5th International Conference on Electronic Information Engineering and Computer Communication (EIECC) (pp. 692-697). IEEE.

[10] Alansari, A., & Luqman, H. (2025). Large language models hallucination: A comprehensive survey. arXiv preprint arXiv:2510.06265.

[11] Gao, G., Ma, X., Lu, C., & Gao, R. (2026). Reliability Analysis and Application Research of SMS Communication Systems in Medical Notification Scenarios.

[12] Lee, S., Kim, B., & Lee, H. (2026). Mitigating Cognitive Inertia in Large Reasoning Models via Latent Spike Steering. arXiv preprint arXiv:2601.22484.

[13] Xu, D., Chen, H., & Gui, H. (2026). Unified Online Estimation Method for SOC, SOH, and Power Capacity Considering Safety Boundary Consistency in Battery Management Systems.

[14] Prados, A., Garrido, S., & Barber, R. (2024). Learning and generalization of task-parameterized skills through few human demonstrations. Engineering Applications of Artificial Intelligence, 133, 108310.

[15] Wang, Y., Chen, J., Wang, Y., & Yin, X. (2026). Application of Obtainable Biological Agent Characteristics in Efficacy Stratification of Oral Anti-Obesity Drugs.

[16] Shah, M. I. A., Barrett, E., & Mason, K. (2025). Uncertainty-aware knowledge transformers for peer-to-peer energy trading with multi-agent reinforcement learning. arXiv preprint arXiv:2507.16796.

[17] Zhang, Y., Gu, W., & Wang, J. (2026). Construction of Wind Farm Asset Health Index Based on Multi-Dimensional Indicators and Analytic Hierarchy Process and Its Correlation with Operational Performance. Authorea Preprints.

[18] Agashe, S., Fan, Y., Reyna, A., & Wang, X. E. (2025, April). Llm-coordination: evaluating and analyzing multi-agent coordination abilities in large language models. In Findings of the Association for Computational Linguistics: NAACL 2025 (pp. 8038-8057).

[19] Jiao, Y., Zhao, B., Wang, A., & Shi, T. (2026). Construction and Empirical Study of a Modularized Teaching System for Art Courses Based on a Unified Training Pathway.

[20] Lainwright, N., & Pemberton, M. (2024). Assessing the response strategies of large language models under uncertainty: A comparative study using prompt engineering. Preprint posted online on August, 1.

[21] Jiao, Y., Wang, A., Zhao, B., & Shi, T. (2026). The Impact of Visual Language Strategies in Public Art Creation on Community Spatial Perception and Public Behavior.

[22] El Kodssi, I., Sbai, H., & Ait Mansour, N. (2026). PRISMA: Physically-Aware Reasoning and Intelligent Semantic Mining Architecture for IoT Process Discovery Using Deep Learning and Large Language Models. IEEE Access.