ISSN: 3079-6342

Entropy-Based Anomaly Detection for Memory Integrity Preservation in Multi-Agent Systems

Lukas Meier[1], Camille Morel[2], Adrian Schmid[3]*

Department of Computer Science, ETH Zurich, 8092 Zurich, Switzerland

Corresponding author: adrian.schmid@ethz.ch

Abstract

Memory poisoning often alters statistical properties of shared state representations. This study proposes an entropy-based monitoring mechanism to detect anomalous memory distributions in collaborative agent environments. Shannon entropy and conditional entropy metrics are computed over memory state vectors at each synchronization step. A divergence threshold based on Kullback–Leibler distance identifies suspicious memory updates. Experiments were conducted on distributed planning simulations with 150 agents and controlled poisoning injection rates from 5% to 30%. Detection precision reached 93.1% and recall 88.7% at a divergence threshold of 0.27. Early intervention reduced system-wide contamination by 48.3% compared with no monitoring. Entropy-driven detection provides a lightweight and scalable solution for preserving memory integrity.

Keywords

Memory poisoning; entropy detection; anomaly monitoring; multi-agent collaboration; statistical divergence; distributed systems

1. Introduction

Multi-agent systems (MAS) are increasingly used in tasks that require coordinated perception, distributed planning, cooperative control, and autonomous decision-making [1]. In these systems, agents often rely on shared memory, synchronized state representations, or common information buffers to exchange observations, maintain task consistency, and support joint actions [2]. The reliability of such shared information has a direct influence on overall system performance, particularly in large-scale and dynamic environments where agents repeatedly update and reuse common states [3]. Once the shared memory layer is corrupted, the effect may not remain local. Instead, it can propagate through repeated synchronization, influence normal agents over time, and gradually weaken the stability and reliability of the entire collaborative process [4]. Recent studies have shown that poisoned memory in collaborative agent environments may persist across multiple interaction rounds and continue to affect later decisions even after the initial attack source is no longer active [5]. This risk makes memory integrity a critical issue in secure multi-agent collaboration. Existing studies on MAS security have mainly focused on communication attacks, false-data injection, denial-of-service attacks, adversarial observations, and resilient consensus control [6,7]. A large body of work has developed mechanisms to detect abnormal signals, secure communication links, and preserve coordination under hostile external conditions [8]. These studies have greatly improved the understanding of secure cooperation in distributed systems. Even so, their primary concern is usually external inputs, communication channels, sensor signals, or control commands [9]. In practical multi-agent environments, agents do not depend only on real-time interaction. They also use stored state summaries, synchronized memory structures,

ISSN: 3079-6342

historical records, and intermediate task representations to support later decisions [10]. When these internal memory states are poisoned, the abnormality may remain hidden for several synchronization cycles before visible system degradation appears. This delayed effect makes memory poisoning harder to identify than direct communication interference, because the corrupted information can be repeatedly reused and amplified within normal collaboration dynamics [11]. Related research on poisoning attacks and backdoor threats in intelligent agents provides further evidence that internal representations are important attack targets. In reinforcement learning and multi-agent learning, poisoning can alter rewards, training samples, local observations, or value estimates, which may lead to biased policies, unstable convergence, or sustained performance loss [12]. In LLM-based multi-agent systems, malicious prompts, hidden backdoors, and manipulated reasoning traces have also been shown to influence intermediate memory states and later collaborative decisions. These findings suggest that the vulnerability of intelligent agents is not limited to external data streams. The internal memory layer itself can become a channel through which harmful information accumulates and spreads. However, much of the existing work still emphasizes attack construction, adversarial effectiveness, or robustness improvement during training. Comparatively less attention has been given to the online monitoring of shared memory after deployment, especially in settings where poisoning propagates gradually through multi-agent synchronization rather than causing an immediate and obvious failure. This gap is important because memory poisoning in collaborative systems has several characteristics that distinguish it from conventional attack models. Its impact is often cumulative rather than instantaneous. The abnormal signal may be statistically weak at the beginning, yet repeated synchronization can magnify the deviation and eventually influence many otherwise normal agents [13]. In addition, the propagation process is closely tied to the structure of shared state updates, which means that harmful memory changes may appear as subtle distributional shifts rather than as obvious communication anomalies. Defense strategies that depend heavily on specific learning architectures, predefined attack assumptions, or task-dependent supervision may therefore be difficult to generalize across different collaborative systems. Another limitation in the literature is that defense performance is often judged mainly by final task accuracy, success rate, or control stability. These end-point measures are useful, but they do not fully answer whether early statistical changes in shared memory can provide warning signals before large-scale contamination occurs. This issue becomes even more important in large distributed environments, where delayed detection can allow poisoned updates to spread widely before any corrective action is triggered. Anomaly detection methods based on statistical and information-theoretic measures offer a promising direction for addressing this problem [14]. Entropy-based methods are particularly attractive because they characterize uncertainty, structural disorder, and dependency change in complex data without requiring strong assumptions about every variable in the system. Shannon entropy is widely used to describe disorder in a state distribution, while conditional entropy can capture changes in dependency structure between related variables. Kullback-Leibler divergence is also a common tool for measuring the discrepancy between an observed distribution and an expected reference distribution [15,16]. These measures are computationally efficient, interpretable, and suitable for online update, which makes them well matched to large-scale distributed environments where system states evolve continuously. Although information-based detection has been applied in traffic analysis, out-of-distribution monitoring, fault diagnosis, and general anomaly detection, its use for protecting shared memory during multi-agent synchronization remains limited. In particular, there is still a lack of work that treats the shared memory layer itself as the primary object of online security monitoring. Several unresolved issues therefore remain in the current literature. Memory poisoning under shared multi-agent state synchronization has not yet been examined with sufficient depth, especially

ISSN: 3079-6342

from the perspective of propagation dynamics and early warning. Many available defenses are closely tied to particular attack settings, model structures, or training schemes, which restricts their use in broader collaborative environments [17]. The statistical behavior of poisoned memory updates before visible system failure is also not well understood, leaving open the question of whether low-cost online indicators can detect contamination in time. In addition, many related experiments are conducted under relatively limited settings or moderate system scales, which makes it difficult to judge how well existing methods perform when the number of agents and synchronization frequency increase. These gaps indicate the need for a detection framework that is lightweight, online, model-agnostic, and directly targeted at the shared memory layer where contamination can accumulate over time. This study proposes an entropy-based anomaly detection method for memory integrity preservation in multi-agent systems. The core assumption is that poisoned memory updates alter the statistical distribution and dependency structure of shared state vectors before severe behavioral failure becomes visible. Based on this assumption, the proposed framework monitors shared memory at each synchronization step by computing Shannon entropy and conditional entropy to capture uncertainty variation and structural change. A divergence criterion derived from Kullback-Leibler distance is then used to identify suspicious updates and trigger early intervention. Unlike many existing approaches that focus mainly on communication signals, external inputs, or robustness at the training stage, the proposed method directly targets the shared memory layer where corrupted information can accumulate and propagate across agents. The significance of this work lies in two aspects. From a methodological perspective, it introduces an online and statistically grounded mechanism for detecting abnormal memory evolution in collaborative agent environments. From an application perspective, it provides a practical way to reduce system-wide contamination risk in distributed MAS without requiring detailed knowledge of a specific attack model or a particular learning architecture. The study therefore aims to determine whether entropy-based monitoring can detect poisoned memory updates accurately and in a timely manner, and whether such early detection can effectively limit the propagation of contamination in large multi-agent environments.

2.Materials and Methods

2.1. Sample and Study Environment

The experiments were carried out in a simulated distributed planning environment designed for collaborative multi-agent decision-making under repeated memory synchronization. A total of 150 agents were included in the system, and each agent maintained a local memory state vector that was updated through interaction and periodic synchronization with neighboring or shared system states. The study focused on memory integrity under adversarial disturbance, with particular attention to the statistical changes caused by poisoned updates. The simulation environment was built to reflect a large-scale cooperative setting in which agents exchanged task-related state information over multiple rounds. Poisoning injection rates were set at 5%, 10%, 20%, and 30% to represent different levels of attack intensity. For each condition, multiple runs were conducted under the same planning rules to reduce random effects and to ensure that the observed changes were caused by memory corruption rather than by normal variation in system behavior.

ISSN: 3079-6342

2.2. Experimental Design and Control Setting

The study used a controlled comparative design with one normal condition and several poisoning conditions. In the control setting, all agents exchanged clean memory updates during synchronization, and no malicious perturbation was introduced. In the experimental settings, a fixed proportion of memory updates was deliberately altered before synchronization to simulate poisoning behavior. The purpose of this design was to compare the statistical behavior of shared memory under normal and abnormal conditions and to test whether entropy-based monitoring could separate these two cases in a stable way. The selected poisoning rates covered mild, moderate, and severe disturbance levels, which made it possible to examine how detection performance changed with attack strength. The comparison between monitored and unmonitored systems was also included to assess whether early anomaly detection could reduce the spread of corrupted memory across the agent population. This design provided a direct basis for evaluating both detection accuracy and intervention effect.

2.3. Measurement Procedure and Quality Control

At each synchronization step, the memory state vectors collected from the agent group were converted into probability distributions for statistical analysis. Shannon entropy was used to describe the uncertainty of the current memory distribution, while conditional entropy was used to measure changes in dependence between synchronized memory states. Kullback-Leibler divergence was then calculated between the observed memory distribution and the reference distribution to determine whether the current update should be marked as suspicious. To keep the measurements consistent, all state vectors were normalized before entropy and divergence values were computed. The same synchronization interval, poisoning rule, and detection threshold search range were applied across all trials. Repeated experiments were performed under each poisoning level, and the reported results were obtained from the aggregated outputs rather than from a single run. In addition, abnormal values caused by numerical instability in very small probabilities were controlled by adding a small smoothing constant during probability estimation. These steps helped maintain the stability and repeatability of the measurement process.

2.4. Data Processing and Model Formulation

The raw memory state vectors were first standardized and transformed into discrete probability distributions at each synchronization step. After that, entropy-based indicators were calculated for each update cycle, and the resulting values were compared across normal and poisoned conditions. Detection performance was evaluated by precision, recall, and contamination reduction rate. Shannon entropy was defined as

$$H(X)=-\sum_{i=1}^{n} p(x_i)\log p(x_i),$$

where $p(x_i)$ is the probability of the i-th memory state and n is the number of discrete states. To measure the deviation between the observed distribution P and the reference distribution Q, Kullback-Leibler divergence was calculated as

$$D_{KL}(P\|Q)=\sum_{i=1}^{n} p(x_i)\log\frac{p(x_i)}{q(x_i)}.$$

ISSN: 3079-6342

A memory update was marked as anomalous when the divergence value exceeded the preset threshold. In addition, detection precision and recall were calculated as

$$\text{Precision}=\frac{TP}{TP+FP},\quad \text{Recall}=\frac{TP}{TP+FN},$$

where TP, FP, and FN denote true positives, false positives, and false negatives, respectively. These indicators were used to compare the method across different poisoning levels and threshold settings.

2.5. Evaluation Strategy and Statistical Analysis

The method was evaluated from two aspects: anomaly detection performance and system-level protection effect. Detection performance was measured by precision and recall under different poisoning injection rates and divergence thresholds. System-level protection effect was assessed by comparing the spread of corrupted memory in systems with monitoring and systems without monitoring. The threshold value was selected by examining the trade-off between missed detection and false alarm over repeated trials, and the setting that produced the best overall balance was used for the main analysis. To reduce the influence of random fluctuations, each experimental condition was repeated several times, and the mean values were reported. Changes in contamination level after early intervention were expressed as percentage reduction relative to the unmonitored condition. This evaluation strategy made it possible to judge not only whether the method could detect poisoned memory updates, but also whether it could limit the wider impact of poisoning on collaborative agent behavior.

3.Results and Discussion

3.1. Detection performance at different poisoning levels

The proposed method showed stable detection performance under all poisoning conditions. When the divergence threshold was set to 0.27, the precision reached 93.1% and the recall reached 88.7%. These results show that most poisoned memory updates were correctly identified, while the number of false alarms remained low. The results also showed a clear link between poisoning rate and detection difficulty. At low injection rates, the changes in shared memory were relatively small, and some poisoned updates were still close to normal synchronization patterns. As the poisoning rate increased, the gap between normal and corrupted memory states became clearer, and detection became more reliable. This result indicates that the proposed method can capture gradual changes in memory structure, rather than only strong disturbances [18,19]. Unlike many anomaly detection studies that focus on traffic data, output labels, or time-series signals, this method works directly on synchronized memory distributions. This feature makes it more suitable for collaborative multi-agent systems, where contamination usually spreads through internal state exchange rather than a single external input channel (Fig. 1).

ISSN: 3079-6342

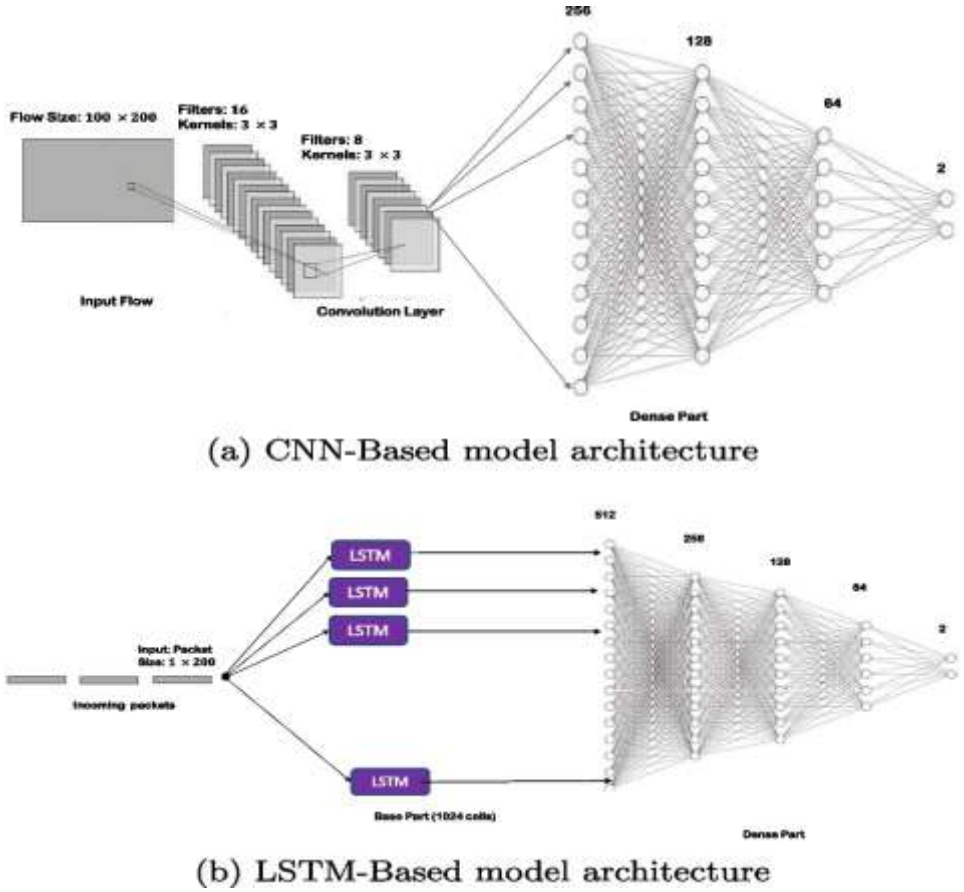

Figure 1. Detection performance of the entropy-based method under different memory poisoning rates.

3.2. Effect of early intervention on contamination spread

Early intervention clearly reduced the spread of corrupted memory. Compared with the system without monitoring, the monitored system reduced overall contamination by 48.3%. This result shows that suspicious updates can be intercepted during synchronization before they affect a large number of agents. The effect was most obvious under moderate poisoning levels. In this range, the attack was strong enough to change the statistical pattern of memory states, but not yet strong enough to dominate the whole shared state space. Under these conditions, the monitoring rule worked as an effective filter and blocked repeated propagation of harmful memory vectors. This finding is different from many earlier studies, which mainly discuss detection accuracy. In the present study, the main benefit was not only better identification of abnormal updates, but also better control of later contamination. In other words, the monitoring step changed the later system state by stopping the spread of corrupted memory at an early stage [20,21]. This effect is important in multi-agent planning because even a limited number of poisoned updates may influence many later decisions through repeated synchronization (Fig. 2).

ISSN: 3079-6342

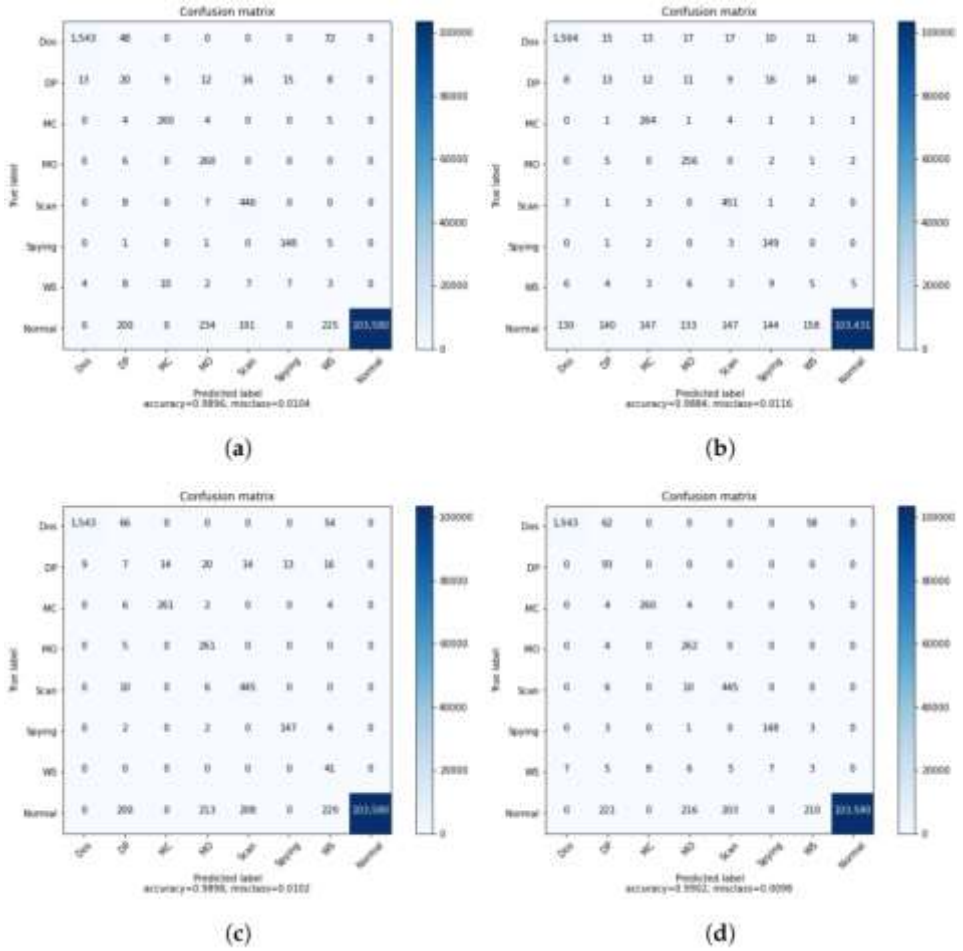

Figure 2:Effect of early intervention on the spread of corrupted memory in the multi-agent system.

3.3. Role of entropy and divergence measures

The combined use of Shannon entropy, conditional entropy, and Kullback-Leibler divergence provided a good balance between detection ability and computational cost. Shannon entropy described the overall uncertainty of the current memory distribution. Conditional entropy reflected changes in the dependency structure between synchronized states. Kullback-Leibler divergence measured how far the current update moved away from the expected distribution. These three indicators played different roles and complemented each other. Entropy alone could reflect a rise in uncertainty, but it could not always separate normal fluctuation from poisoning. Divergence alone was more direct, but its stability improved when the background uncertainty of the memory state was also considered. The results showed that the combined scheme worked well in repeated synchronization settings, where poisoned updates did not always appear as obvious outliers in the raw state space. Compared with deep anomaly detection models that rely on large network structures, the present method is easier to apply and easier to explain. This is useful in large distributed systems, where online monitoring needs to remain simple and efficient, and where interpretable statistical evidence is often preferred over black-box scores [22,23].

ISSN: 3079-6342

3.4. Comparison with earlier studies and study limitations

Compared with earlier work, this study offers two main advances. First, it moves the focus from communication-level anomaly detection to memory-level anomaly detection. This difference is important because a collaborative multi-agent system may still appear to function normally even when its shared memory has already been biased. Second, the method is designed for online use during synchronization, which gives it practical value for systems that require continuous integrity checking. Earlier studies on anomaly detection in distributed systems often report good classification performance, but many of them depend on more complex learning models or are tested on external attack data rather than poisoned internal memory states. The present results show that a simpler statistical method can still achieve high precision and recall, while also reducing later contamination. At the same time, this study has several limitations. The experiments were carried out in a controlled simulation setting with fixed poisoning rates and a fixed number of 150 agents. Real systems may involve changing communication structures, unequal trust between agents, and more adaptive attack behavior. In addition, the current framework uses a fixed divergence threshold. Future work may consider adaptive threshold selection under changing task conditions. Even with these limits, the results suggest that entropy-based memory monitoring is a useful way to protect collaborative systems from hidden memory corruption.

4.Conclusion

This study proposed an entropy-based method to protect memory integrity in multi-agent systems under poisoning attacks. By monitoring Shannon entropy, conditional entropy, and Kullback-Leibler divergence during memory synchronization, the method detected abnormal memory updates with high precision and recall, and it also reduced the spread of corrupted memory through early intervention. The main value of this work is that it moves attention from external communication anomalies to the internal reliability of shared memory, which plays an important role in collaborative agent systems but has received less attention in previous studies. The results show that simple statistical measures can provide clear and reliable signals for online monitoring without adding high computational cost. This makes the method useful for distributed planning and other cooperative tasks that depend on stable memory exchange among many agents. The study also shows good potential for use in secure autonomous systems, distributed robotics, and other large collaborative platforms where hidden memory corruption may affect later decisions. At the same time, the present results were obtained in a controlled simulation environment with fixed poisoning rates and a fixed number of agents. Further work is still needed under more complex conditions, such as dynamic network structures, adaptive attack strategies, and different memory update rules. Future studies should also examine adaptive threshold selection and test the method in real multi-agent applications. Overall, the findings suggest that entropy-based memory monitoring is a practical and scalable way to improve the reliability of collaborative intelligent systems.

References

[1] Gu, X., Yang, J., Tian, X., & Liu, M. (2025). Research on the Construction of a Human-Machine Collaborative Anti-Money Laundering System and Its Efficiency and Accuracy Enhancement in Suspicious Transaction Identification.

[2] Maisto, D., Donnarumma, F., & Pezzulo, G. (2023). Interactive inference: A multi-agent model of cooperative joint actions. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 54(2), 704-715.

ISSN: 3079-6342

[3] Yang, Y., Leuze, C., Hargreaves, B., Daniel, B., & Baik, F. (2025). EasyREG: Easy Depth-Based Markerless Registration and Tracking using Augmented Reality Device for Surgical Guidance. arXiv preprint arXiv:2504.09498.

[4] Acha, S. N. (2025). Cooperative Intelligent Control Through Reliable Trusted Communication: Enhancing Reliability and Data Pooling in Multi-Agent Systems (MAS) (Doctoral dissertation, North Carolina Agricultural and Technical State University).

[5] Liu, H., Xu, D., Ma, Q., Xu, S., & Qiu, D. (2026). Memory Poisoning Propagation and Repair Mechanism in Multi-Agent Collaborative Environments.

[6] Aslam, M. M., Ahmed, Z., Du, L., Hassan, M. Z., Ali, S., & Nasir, M. (2022). An overview of recent advances of resilient consensus for multiagent systems under attacks. Computational intelligence and neuroscience, 2022(1), 6732343.

[7] Ghods, A. A., & Doostmohammadian, M. (2025). Resilient consensus-based target tracking under false data injection attacks in multi-agent networks. Signals, 6(3), 44.

[8] Bai, W., Wu, Q., Wu, K., & Lu, K. (2024). Exploring the Influence of Prompts in LLMs for Security-Related Tasks. In Workshop on Artificial Intelligence System with Confidential Computing (AISCC 2024)(San Diego, CA). USA. https://dx. doi. org/10.14722/aiscc.

[9] Emami, M., Bayat, A., Tafazolli, R., & Quddus, A. (2024). A survey on haptics: Communication, sensing and feedback. IEEE Communications Surveys & Tutorials, 27(3), 2006-2050.

[10] Ma, Q., Yue, L., Xu, S., Shi, Y., & Liu, H. (2026). Web Agent Agentic Reinforcement Learning Decision Model Under Multi-Cost and Failure Risk Constraints.

[11] Qiu, D., Xu, D., & Yue, L. (2025, December). Reinforcement Learning-Augmented LLM Agents for Collaborative Decision Making and Performance Optimization. In 2025 7th International Conference on Frontier Technologies of Information and Computer (ICFTIC) (pp. 1337-1342). IEEE.

[12] Putla, H., Patibandla, C., Singh, K. P., & Nagabhushan, P. (2024). A pilot study of observation poisoning on selective reincarnation in multi-agent reinforcement learning. Neural Processing Letters, 56(3), 161.

[13] Lee, U., Kim, H., Kim, M., Oh, G., Joo, P., Park, A., ... & Mashour, G. A. (2025). Proximity to explosive synchronization determines network collapse and recovery trajectories in neural and economic crises. Proceedings of the National Academy of Sciences, 122(44), e2505434122.

[14] Du, Y. (2025). Research on Deep Learning Models for Forecasting Cross-Border Trade Demand Driven by Multi-Source Time-Series Data. Journal of Science, Innovation & Social Impact, 1(2), 63-70.

[15] Bonnici, V. (2024). A maximum value for the Kullback–Leibler divergence between quantized distributions. Information, 15(9), 547.

ISSN: 3079-6342

[16] Liu, S., Feng, H., & Liu, X. (2025). A Study on the Mechanism of Generative Design Tools' Impact on Visual Language Reconstruction: An Interactive Analysis of Semantic Mapping and User Cognition. Authorea Preprints.

[17] Kaushal, V., & Sharma, S. (2025). Securing the collective intelligence: a comprehensive review of federated learning security attacks and defensive strategies. Knowledge and Information Systems, 67(4), 3099-3137.

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

[19] Paulsen, J. D., & Keim, N. C. (2025). Mechanical memories in solids, from disorder to design. Annual Review of Condensed Matter Physics, 16(1), 61-81.

[20] Chen, H., Li, J., Ma, X., & Mao, Y. (2025, June). Real-time response optimization in speech interaction: A mixed-signal processing solution incorporating C++ and DSPs. In 2025 7th International Conference on Artificial Intelligence Technologies and Applications (ICAITA) (pp. 110-114). IEEE.

[21] Jangam, S. K., & Muntala, P. S. R. P. (2023). Challenges and Solutions for Managing Errors in Distributed Batch Processing Systems and Data Pipelines. International Journal of Emerging Research in Engineering and Technology, 4(4), 65-79.

[22] Li, T., Liu, S., Hong, E., & Xia, J. (2025). Human Resource Optimization in the Hospitality Industry Big Data Forecasting and Cross-Cultural Engagement.

[23] Esna-Ashari, M. (2025). Beyond the black box: a review of quantitative metrics for neural network interpretability and their practical implications. International journal of sustainable applied science and engineering, 2(1), 1-24.