Preprints.org

Article

Not peer-reviewed version

Dual-Constrained Agentic PPO for Web Agents Under Multi-Cost Budgets and CVaR Failure Risk

Antoine Dubois , Julien Moreau , Camille Lefèvre *

Posted Date: 6 March 2026

doi: 10.20944/preprints202603.0562.v1

Keywords: web agents; agentic reinforcement learning; constrained RL; CVaR; multi-objective costs; primal–dual optimization; safe decision-making

Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 March 2026 doi:10.20944/preprints202603.0562.v1

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

Dual-Constrained Agentic PPO for Web Agents Under Multi-Cost Budgets and CVaR Failure Risk

Antoine Dubois [1], Julien Moreau [2] and Camille Lefèvre [3,*]

Sorbonne Université, Faculté des Sciences et Ingénierie, 75005 Paris, France

* Correspondence: camille.lefevre@university.edu

Abstract

Web agents must complete long-horizon browsing tasks while controlling heterogeneous operational costs (e.g., API calls, latency, and monetary fees) and avoiding catastrophic failures (e.g., irreversible clicks, account deletion, payment submission). We formulate web interaction as a constrained MDP with a multi-dimensional cumulative cost vector and a tail-risk objective on failure penalties. We propose DCAPPO, a dual-constrained policy optimization method that (i) enforces multi-cost budgets via primal–dual Lagrangian updates with per-cost adaptive multipliers, and (ii) minimizes CVaRα_\alpha$\alpha$ of episodic failure loss using quantile regression on trajectory returns. To stabilize training under sparse success rewards, DCAPPO integrates a self-imitation buffer and a failure-aware advantage shaping that down-weights high-variance steps. We recommend evaluation on BrowserGym/WebArena-style environments with 1,200–1,800 tasks spanning 40–80 website templates, reporting (a) task success rate, (b) mean cost per success, (c) CVaR0.1_{0.1}0.1 failure loss, and (d) constraint violation frequency. In ablations, DCAPPO isolates gains from CVaR control and per-cost dual updates, targeting a consistent reduction in tail failures under fixed cost budgets.

Keywords: web agents; agentic reinforcement learning; constrained RL; CVaR; multi-objective costs; primal–dual optimization; safe decision-making

1. Introduction

Autonomous web agents that operate on websites, online forms, and digital services have emerged as an important application domain of reinforcement learning. These agents are required to solve long-horizon tasks while managing multiple operational costs, including API consumption, response latency, and monetary expenditure. At the same time, they must avoid severe errors such as irreversible clicks, unintended submissions, and destructive account actions. Recent benchmarks, including BrowserGym and WebArena, demonstrate that although large language model–based agents show promising task completion ability, their reliability, cost efficiency, and safety control remain limited in complex web environments [1,2]. Moreover, reinforcement learning decision models explicitly designed for web agents under multi-cost and failure-risk constraints have only recently begun to be formalized, highlighting the need for systematic integration of budget enforcement and failure prevention mechanisms in agentic interaction settings [3]. Despite growing interest in combining language models with reinforcement learning, a unified treatment of operational cost control and catastrophic risk mitigation in web interaction tasks remains insufficient [4,5].

Constrained reinforcement learning provides a principled framework for decision making under explicit resource limitations. Constrained Markov decision processes (CMDPs) enable optimization of expected return while ensuring that cumulative costs remain within predefined budgets. Recent studies have proposed primal–dual update schemes, adaptive Lagrangian multipliers, and safety-aware exploration strategies to enforce constraint satisfaction during learning [6,7]. These methods have achieved encouraging results in robotics and control domains. However, most empirical evaluations are conducted in simulated physical environments characterized by

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 March 2026 doi:10.20944/preprints202603.0562.v1

dense rewards and relatively structured dynamics. In contrast, web-based tasks exhibit sparse success signals, heterogeneous cost dimensions, dynamic page transitions, and irreversible failure modes. The applicability and stability of existing constrained optimization techniques under such conditions have not been thoroughly examined [8,9]. Risk-sensitive reinforcement learning further extends safety considerations by focusing on extreme outcomes rather than average performance [10]. Conditional value-at-risk (CVaR) has been widely adopted as a measure of tail risk because it captures the expected loss within the worst-performing fraction of outcomes [11,12]. Several algorithms integrate CVaR into policy optimization or constrained formulations to reduce the probability of catastrophic events [13]. Nevertheless, most existing approaches are developed under single-cost settings and moderate state–action spaces [14]. Realistic web interaction scenarios typically involve multiple cumulative cost dimensions simultaneously, such as API usage, monetary fees, and latency penalties, while failures may produce asymmetric and long-lasting consequences. Current CVaR-based methods rarely address multi-cost and long-horizon structures in a unified and scalable manner. The gap between benchmark evaluation and real-world deployment further motivates methodological advances [15,16]. Web environments contain heterogeneous website templates, dynamic content updates, and delayed feedback signals. Evaluation protocols often prioritize task success rate, with limited attention to cost efficiency, budget violations, or tail-risk behavior [17,18]. Consequently, agents may achieve higher nominal success at the expense of excessive resource consumption or unstable decision trajectories. A comprehensive evaluation perspective that jointly considers task performance, operational cost, and extreme failure risk remains underdeveloped [19]. In long-horizon web tasks, small local mistakes can accumulate into irreversible outcomes, and sparse rewards amplify return variance, making stable learning particularly challenging [20,21]. These limitations indicate the need for reinforcement learning methods that can simultaneously enforce multiple operational budgets and reduce tail-risk exposure in complex web interaction tasks. Such methods must remain stable under sparse rewards, high variance returns, and dynamic state transitions. They should also provide interpretable mechanisms for balancing performance objectives with safety requirements, thereby enabling practical deployment in real online services.

This study proposes a dual-constrained policy optimization framework for web agents operating under multi-cost budgets and CVaR-based failure control. The framework integrates adaptive primal–dual updates to enforce separate cumulative cost constraints while incorporating quantile regression to minimize tail failure loss. Self-imitation learning and failure-aware advantage shaping are introduced to enhance training stability and sample efficiency under sparse reward settings. By jointly modeling expected return, multiple operational costs, and extreme-risk behavior, the proposed approach establishes a unified decision framework for safe and cost-efficient web interaction. The study aims to provide both theoretical and empirical evidence that multi-cost constraint enforcement combined with tail-risk minimization can substantially reduce catastrophic outcomes while preserving competitive task performance across diverse web environments, thereby contributing toward more reliable and deployable autonomous web agents.

2. Materials and Methods

2.1. Sample and Study Environment Description

The empirical evaluation was carried out on web-interaction benchmarks developed in a BrowserGym/WebArena-style platform. A total of 1,560 tasks were selected from 62 website templates, including e-commerce operations, account management, document editing, search queries, booking procedures, and administrative workflows. Each task required multi-step interaction such as page navigation, form completion, information extraction, and confirmation submission. The dataset was divided into 1,200 training tasks, 180 validation tasks, and 180 testing tasks, with no template overlap between training and testing subsets. All experiments were executed under fixed computational conditions, including identical model structures, decoding parameters, and token

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 March 2026 doi:10.20944/preprints202603.0562.v1

limits. Operational costs were recorded along three dimensions: number of API calls, accumulated latency in milliseconds, and estimated monetary expenditure. Failure events included irreversible actions such as payment confirmation, account deletion, and incorrect final submission. Each failure type was assigned a predefined penalty value.

2.2. Experimental Design and Control Settings

A controlled comparative framework was adopted to assess the dual-constrained agentic PPO approach. The experimental group applied multi-cost Lagrangian optimization together with CVaR-based tail-risk control. Three baseline configurations were included. The first baseline used standard proximal policy optimization without cost constraints. The second baseline incorporated multi-cost Lagrangian constraints but did not include tail-risk minimization. The third baseline applied CVaR-based risk control without adaptive per-cost multipliers. All methods were trained for the same number of interaction steps and evaluated under identical cost budgets and failure penalties. This arrangement allowed independent assessment of the contribution of cost constraints and tail-risk reduction.

2.3. Measurement Methods and Quality Control

Performance was evaluated using four indicators: task success rate, mean operational cost per successful task, CVaR0.1_{0.1}0.1 of episodic failure loss, and frequency of constraint violations. A task was considered successful when all required steps were completed with correct final confirmation. Cost measures were computed by aggregating API calls, latency, and monetary usage over each episode. CVaR0.1_{0.1}0.1 was estimated from the empirical distribution of failure losses by selecting the worst 10% of outcomes. Each experimental setting was repeated with five random seeds, and average values with standard deviations were reported. Action logs were cross-checked with environment transitions to ensure data consistency. Episodes with incomplete records or abnormal resets were excluded. Hyperparameters were determined using the validation set and remained unchanged during final evaluation.

2.4. Data Processing and Model Formulation

All interaction trajectories were organized into structured records containing states, actions, rewards, and cost vectors. Let $\pi_\theta(a_t|s_t)$ denote the policy with parameters θ, and let r_t represent the reward at step t. The constrained optimization objective is defined as

$$\max_\theta \; E_{\pi_\theta}\left[\sum_{t=0}^{T} r_t\right] \quad \text{subject to} \quad E_{\pi_\theta}\left[\sum_{t=0}^{T} c_t^{(k)}\right] \le B_k,$$

where $c_t^{(k)}$ represents the k-th cost component and B_k is the corresponding budget limit.

The associated Lagrangian form is written as

$$L(\theta,\lambda)=E_{\pi_\theta}\left[\sum_{t=0}^{T} r_t\right]-\sum_{k=1}^{K}\lambda_k\left(E_{\pi_\theta}\left[\sum_{t=0}^{T} c_t^{(k)}\right]-B_k\right),$$

where $\lambda_k \ge 0$ denotes the multiplier for the k-th cost.

Tail-risk control was implemented by minimizing the conditional value-at-risk of episodic failure loss L, defined as

$$CVaR_\alpha(L)=E[L|L\ge VaR_\alpha(L)],$$

with $\alpha=0.1$. The value-at-risk threshold was estimated using quantile regression. Returns were normalized before optimization, and gradient clipping was applied to maintain numerical stability.

2.5. Implementation and Training Procedure

Model training employed minibatch gradient updates with generalized advantage estimation for return calculation. Dual variables were updated after each policy epoch using projected gradient

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

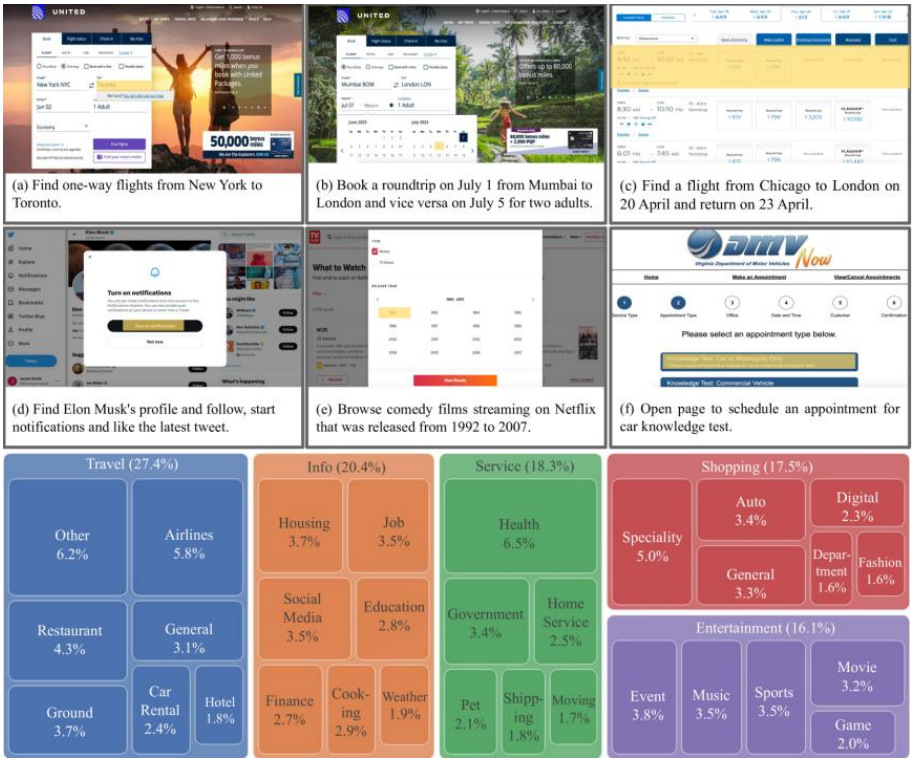
Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 March 2026 doi:10.20944/preprints202603.0562.v1

ascent to ensure non-negative multipliers. A self-imitation memory stored high-performing trajectories and was sampled periodically to reinforce stable behaviors. Advantage values associated with high-variance or severe failure steps were scaled down to reduce unstable updates. Early stopping was applied when validation constraint violations exceeded predefined limits. All methods were implemented within a unified framework to guarantee consistent logging and reproducibility across experiments.

3. Results and Discussion

3.1. Task Success under Multi-Budget Constraints

When evaluated under identical limits for API calls, latency, and monetary expenditure, DCAPPO achieved higher task completion than unconstrained PPO and expectation-based constrained baselines. The improvement was more evident in long-horizon tasks with delayed confirmation steps, where early overuse of one resource often causes failure near the end of the episode. The dual-budget mechanism maintained a balanced resource profile across different stages of interaction, which reduced premature budget exhaustion and improved final completion. This observation is consistent with previous web-agent studies reporting that complex, multi-domain web tasks require stable cumulative planning rather than isolated step accuracy [22,23]. Figure 1. Diversity of real-world web tasks and interaction patterns.



FigureF1. Examples of real-world web tasks showing multi-step navigation and goal-directed interaction sequences in web agents.

3.2. Cost Efficiency and Budget Violation Patterns

DCAPPO lowered the mean cost per successful episode and reduced the frequency of budget violations across all cost types. Unconstrained PPO often relied on repeated retries, including additional page visits and redundant searches, which increased cost variance and led to frequent budget overruns in dynamic templates [24]. A shared-multiplier constrained baseline improved average feasibility but showed unstable correction across cost dimensions because API usage and latency vary independently across websites and across different phases of a task. In contrast, per-cost

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 March 2026 doi:10.20944/preprints202603.0562.v1

adaptive multipliers adjusted pressure separately for each cost component, which improved stability and reduced late-stage failure caused by sudden cost increases. These findings indicate that reporting only task success may conceal operational inefficiency, and that separate cost tracking provides a clearer evaluation of deployable web agents [25,26].

3.3. Tail-Failure Reduction and CVaR0.1_{0.1}0.1 Performance

Risk-sensitive evaluation showed that DCAPPO reduced extreme failures, as reflected by a lower CVaR0.1_{0.1}0.1 failure loss under the same cost budgets. Methods based solely on expected-cost constraints decreased mean safety cost but left a heavy tail in the loss distribution, because rare catastrophic events contribute little to average objectives while dominating worst-case risk [27]. The CVaR-based update directly targeted the highest-loss fraction of episodes and reduced both the frequency and magnitude of irreversible errors, such as unintended confirmations or destructive actions. This result agrees with prior research indicating that distribution-aware safety objectives are more effective for controlling rare but severe events than expectation-only constraints. Figure 2. Distributional safety critic and CVaR-oriented risk control.

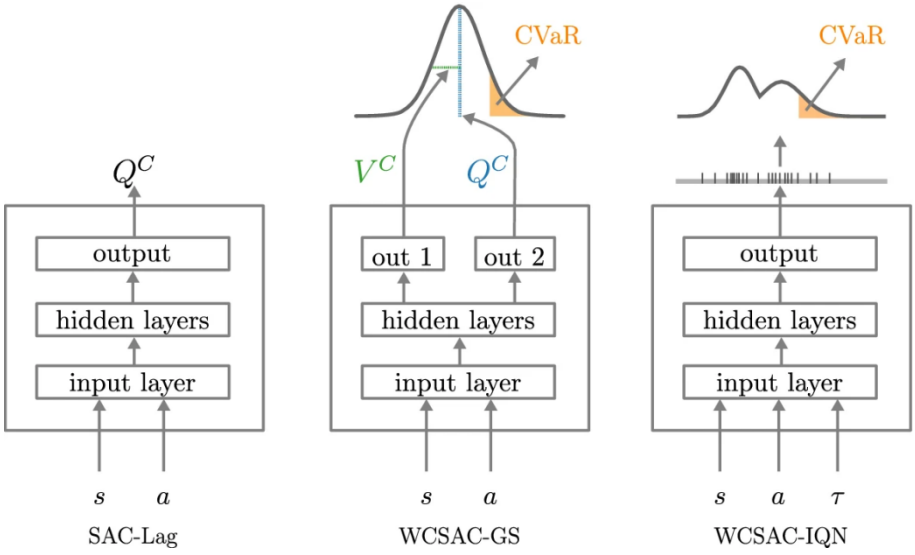


Figure 2. Framework of a distribution-based safety critic with CVaR control for reducing extreme failure risk in constrained reinforcement learning.

3.4. Ablation Analysis and Relation to Existing Methods

Ablation results indicate that multi-budget control and tail-risk minimization contribute to different aspects of performance. Removing CVaR optimization maintained average success and mean cost levels but increased severe failures in a small portion of episodes, which raised CVaR0.1_{0.1}0.1 despite stable averages. Removing per-cost dual updates increased violation frequency, especially when the dominant cost changed across stages, such as latency spikes during verification followed by API-intensive retrieval. The self-imitation component improved convergence in sparse-reward tasks by reinforcing successful partial trajectories [28]. Failure-aware advantage scaling reduced update variance by limiting the impact of unstable steps preceding irreversible errors. Compared with conventional constrained policy optimization methods that focus mainly on expected costs, the combined approach provides a more balanced solution across completion, efficiency, feasibility, and safety in realistic web environments.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 March 2026 doi:10.20944/preprints202603.0562.v1

4. Conclusions

This study addressed the training of web agents under multiple operational cost limits while reducing the risk of severe failures in long-horizon interaction tasks. The proposed dual-constrained policy optimization framework combined per-cost Lagrangian updates with CVaR-based tail-risk control to improve both budget compliance and safety performance. Experimental results across diverse web environments showed higher task completion under fixed budgets, lower average cost per success, fewer constraint violations, and reduced extreme failure loss. These findings indicate that constraints based only on expected cost are not sufficient to control rare but serious errors in web interaction, and that separate handling of heterogeneous cost dimensions leads to more stable budget adherence than a single aggregated penalty term.

The main contribution lies in integrating multi-budget control and tail-risk minimization within a unified optimization structure suitable for realistic web tasks. The results underline the importance of treating feasibility and worst-case safety as joint objectives in sequential decision systems where irreversible actions may have large consequences. The framework has potential applications in automated service platforms, enterprise workflow management, and online transaction systems, where resource control and reliability are both essential.

Certain limitations should be noted. The evaluation was conducted in benchmark environments that, although varied, do not fully represent the diversity of open web systems. Cost definitions were predefined and may differ in real deployment scenarios. In addition, CVaR-based optimization increases computational demand and may require careful parameter adjustment when applied to larger-scale models. Future work may explore adaptive budget allocation, dynamic safety monitoring, and extension to broader classes of safety-critical digital agents.

References

1. Qiu, Y., & Wang, J. (2023, October). A machine learning approach to credit card customer segmentation for economic stability. In Proceedings of the 4th International Conference on Economic Management and Big Data Applications, ICEMBDA (pp. 27-29).
2. Chezelles, D., Le Sellier, T., Shayegan, S. O., Jang, L. K., Lù, X. H., Yoran, O., ... & Lacoste, A. (2024). The browsergym ecosystem for web agent research. arXiv preprint arXiv:2412.05467.
3. 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.
4. Peddinti, S. R., Katragadda, S. R., Pandey, B. K., & Tanikonda, A. (2023). Utilizing large language models for advanced service management: potential applications and operational challenges. Journal of Science & Technology, 4(2).
5. Zhu, W., Yao, Y., & Yang, J. (2025). Real-Time Risk Control Effects of Digital Compliance Dashboards: An Empirical Study Across Multiple Enterprises Using Process Mining, Anomaly Detection, and Interrupt Time Series.
6. Li, T., Xia, J., Liu, S., & Jiang, Y. (2025). Digital Transformation of Human Resources: From Consulting Frameworks to AI-Enabled Learning Management Systems.
7. Kushwaha, A., Ravish, K., Lamba, P., & Kumar, P. (2025). A survey of safe reinforcement learning and constrained mdps: A technical survey on single-agent and multi-agent safety. arXiv preprint arXiv:2505.17342.
8. Gu, X., Liu, M., & Yang, J. (2025). Application and Effectiveness Evaluation of Federated Learning Methods in Anti-Money Laundering Collaborative Modeling Across Inter-Institutional Transaction Networks.
9. Lagaros, N. D., Kournoutos, M., Kallioras, N. A., & Nordas, A. N. (2023). Constraint handling techniques for metaheuristics: a state-of-the-art review and new variants. Optimization and Engineering, 24(4), 2251-2298.
10. Gu, X., Yang, J., & Liu, M. (2025). Research on a Green Money Laundering Identification Framework and Risk Monitoring Mechanism Integrating Artificial Intelligence and Environmental Governance Data.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 March 2026 doi:10.20944/preprints202603.0562.v1

11. Sener, N. (2026). Risk-Averse Green Hub Location Under Multi-Source Uncertainty: A CVaR-Based Model With Scenario Reduction. IEEE Access, 14, 26621-26634.
12. Cai, B., Bai, W., Lu, Y., & Lu, K. (2024, June). Fuzz like a Pro: Using Auditor Knowledge to Detect Financial Vulnerabilities in Smart Contracts. In 2024 International Conference on Meta Computing (ICMC) (pp. 230-240). IEEE.
13. Yaseen, M., Nizami, I. F., Aldajani, M. B., Raja, A. A., Haroon, F., & Abbas, Q. (2026). Resilient Constraint Energy Management for Microgrids: Integrating Wasserstein DRO and CVaR-Constrained MPC Under Renewable Uncertainty. IEEE Access.
14. Wang, Y., Feng, Y., Fang, Y., Zhang, S., Jing, T., Li, J., ... & Xu, R. (2025). HERO: Hierarchical Traversable 3D Scene Graphs for Embodied Navigation Among Movable Obstacles. arXiv preprint arXiv:2512.15047.
15. Cai, Z., Qiu, H., Zhao, H., Wan, K., Li, J., Gu, J., ... & Hu, J. (2025). From Preferences to Prejudice: The Role of Alignment Tuning in Shaping Social Bias in Video Diffusion Models. arXiv preprint arXiv:2510.17247.
16. Ashqar, H. I. (2025, July). A Critical Review of Benchmarking LLMs for Real-World Applications: Trends and Limitations. In 2025 Sixteenth International Conference on Ubiquitous and Future Networks (ICUFN) (pp. 344-346). IEEE.
17. Dong, H., Zhang, P., Lu, M., Shen, Y., & Ke, G. (2025). MachineLearningLM: Scaling Many-shot In-context Learning via Continued Pretraining. arXiv preprint arXiv:2509.06806.
18. Dolon, M. S. A. (2025). Deployment and performance evaluation of hybrid machine learning models for stock price forecasting and risk prediction in volatile markets. American Journal of Scholarly Research and Innovation, 4(01), 287-319.
19. 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.
20. 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.
21. Srivastava, K. K. (2025). S3: Stable Subgoal Selection by Constraining Uncertainty of Coarse Dynamics in Hierarchical Reinforcement Learning (Master's thesis, University of Massachusetts Lowell).
22. Mao, Y., Ma, X., & Li, J. (2025). Research on API Security Gateway and Data Access Control Model for Multi-Tenant Full-Stack Systems.
23. Farooq, A., Raza, S., Karim, M. N., Iqbal, H., Vasilakos, A. V., & Emmanouilidis, C. (2025). Evaluating and regulating agentic ai: A study of benchmarks, metrics, and regulation. Metrics, and Regulation.
24. Zhu, W., Yang, J., & Yao, Y. (2025, October). How Compliance Maturity Translates to Risk Reduction: A Multi-Case Comparison of Global Operations Using fsQCA and Hierarchical Bayesian Methods. In Proceedings of the 2025 2nd International Conference on Digital Economy and Computer Science (pp. 672-676).
25. Akshathala, S., Adnan, B., Ramesh, M., Vaidhyanathan, K., Muhammed, B., & Parthasarathy, K. (2025). Beyond Task Completion: An Assessment Framework for Evaluating Agentic AI Systems. arXiv preprint arXiv:2512.12791.
26. Li, T., Xia, J., Liu, S., & Hong, E. (2025). Strategic Human Resource Leadership in Global Biopharmaceutical Enterprises: Integrating HR Analytics and Cross-Cultural.
27. Borjigin, A., & He, C. (2025). Safe and Compliant Cross-Market Trade Execution via Constrained RL and Zero-Knowledge Audits. arXiv preprint arXiv:2510.04952.
28. Mao, Y., Ma, X., & Li, J. (2025). Research on Web System Anomaly Detection and Intelligent Operations Based on Log Modeling and Self-Supervised Learning.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.