INTERNATIONAL JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGY TRENDS

Vol. 06 No. 01 (2026)

e. ISSN: 2409-5222

https://ijictt.ictt.info

Survival-Aware Agentic RL: Hazard Modeling of Failure Events under Budgeted Web Interaction

Hiroshi Tanaka1, Yuki Nakamura2, Kenji Sato3*

Department of Information Science and Technology, The University of Tokyo, Tokyo 113-8656, Japan

Corresponding author: kenji.sato@university.ac.jp

Abstract. *Failures in web automation often occur as discrete terminal events (e.g., irreversible submission, account lockout) whose risk accumulates over steps. We propose Survival-Aware Agentic RL (SAARL), which models failure as a time-to-event process and learns a hazard function conditioned on state–action pairs. The policy is trained to maximize success while satisfying multi-cost budgets and minimizing an integrated failure hazard. SAARL integrates (i) a Cox-style hazard head on top of the agent's state encoder, (ii) a risk-adjusted return that subtracts a weighted cumulative hazard, and (iii) budget-aware censoring where truncated episodes contribute censored likelihood terms to stabilize hazard estimation. Recommended evaluation uses 10k–30k episodes (or 1,200–2,000 tasks with 8–25 steps) to provide enough event/censoring statistics, reporting concordance index for hazard prediction, failure rate at fixed success, and cost per successful completion. This formulation yields interpretable risk curves ("where risk spikes") and improves safety calibration in long-horizon web tasks.*

Keywords: *Survival analysis; hazard modeling; risk-sensitive RL; constrained decision-making; web automation; failure prediction; calibration*

1. INTRODUCTION

Web automation has evolved from brittle rule-based scripts to large language model (LLM)–driven agents that support multi-step reasoning, tool use, and adaptive decision-making in dynamic online environments. Recent benchmarks and agent frameworks indicate that reinforcement learning (RL) and instruction-tuned LLMs can complete complex tasks—such as form submission, account setup, and information retrieval—under partial observability and heterogeneous website structures [1,2]. Beyond maximizing task success or cumulative reward, emerging agentic RL lines also formalize multi-cost budgets and failure-risk constraints as first-class objectives for deployable web agents [3]. In practice, web interactions are governed not only by reward signals but also by operational budgets (tokens, latency, API fees) and safety requirements (preventing irreversible mistakes), motivating learning objectives that explicitly manage cost and risk rather than treating them as secondary penalties [4,5]. Despite steady improvements, LLM-based agents remain unstable in long-horizon web tasks, especially when trajectories approach irreversible decision points. Irreversible actions—such as submitting incorrect information, triggering account lockout, or confirming unintended transactions—terminate episodes abruptly and can invalidate all prior progress [6,7]. These outcomes are discrete terminal events whose probability often increases as interaction proceeds: early steps may involve low-stakes navigation and exploration, whereas later stages concentrate high-risk actions (authentication, payment, submission) and tighter constraints (timeouts, rate limits). This temporal structure creates a mismatch with common training

signals. Agents can appear strong under average success metrics while remaining poorly calibrated near terminal hazards, producing rare but costly failures that dominate deployment risk [8]. Risk-sensitive and constrained RL methods have been developed to reduce unsafe behavior in sequential decision-making. Distributional RL, conditional value-at-risk (CVaR) optimization, and Lagrangian constrained learning have shown effectiveness in robotics and autonomous control for trading performance against risk or constraint violations [9,10]. In the context of LLM agents, budget-aware strategies regulate tool invocation and inference expenditure, often through explicit cost terms or constrained optimization during training [11,12]. However, most approaches still treat risk as an immediate penalty, or as a constraint on expected cumulative cost. Such formulations do not explicitly represent how the probability of catastrophic events evolves over time. In web automation, failures frequently follow a time-dependent pattern: risk can remain low during early exploration but increase sharply near submission or authentication steps. Reward- and expectation-based constraints alone are often insufficient to capture this escalation, leading to limited safety calibration in long-horizon interaction even when average success improves. Survival analysis provides a statistical framework for modeling time-to-event outcomes via hazard functions and censored observations. Cox proportional hazards models and neural survival approaches have been widely used in healthcare prognosis, reliability modeling, and user-behavior analytics to estimate event risk as a function of covariates and time [13,14]. More recent work connects survival modeling with sequential prediction and decision processes, including hazard-aware bandits and time-to-event forecasting under partial observability [15]. In RL, event-time objectives have been introduced for safety control and maintenance scheduling, where preventing or delaying failure events is central to policy quality [16,17]. Web environments, however, differ in ways that complicate direct transfer. Observations are high-dimensional and language-heavy, interaction dynamics are partially observed and interface-dependent, and catastrophic events are sparse relative to the number of steps. Episodes are also commonly truncated by operational budgets, which creates pervasive censoring that is informative about both cost accumulation and risk exposure. These characteristics motivate a survival-aware design that can learn stable hazard estimates from limited event counts while remaining compatible with modern neural encoders used by LLM agents. Current evaluation practice further limits understanding of failure dynamics. Many web-agent benchmarks contain only a few hundred tasks or short trajectories, yielding few failure events and limited censoring diversity, which reduces the statistical stability of risk estimation [18,19]. Reporting typically emphasizes aggregate success rate or average reward, while calibration properties—such as failure probability at a fixed success level, or the concentration of failures near terminal actions—are rarely examined. Moreover, budget-terminated episodes are often treated as neutral failures, discarding their censoring information and weakening the ability to model how risk develops across time. A more informative evaluation should therefore quantify both decision quality and risk calibration under realistic truncation and cost constraints. To address these gaps, Survival-Aware Agentic Reinforcement Learning (SAARL) models catastrophic failure as a time-to-event outcome conditioned on state–action pairs in web interaction. A Cox-style hazard head is integrated into the agent's state encoder to estimate stepwise failure risk, enabling explicit modeling of risk escalation as the trajectory progresses. Training optimizes a risk-adjusted return that subtracts a weighted cumulative hazard while enforcing multi-cost budget constraints, aligning policy improvement with both completion objectives and safety requirements. Episodes truncated by budget limits are treated as censored observations within a partial-likelihood formulation, allowing the learning signal to exploit informative censoring rather than ignoring it. This design supports more stable hazard estimation when event counts are limited, remains compatible with neural encoders and standard constrained RL procedures, and provides a principled mechanism for calibrating risk near irreversible actions. Evaluation is conducted on large episodic datasets designed to yield sufficient failure and censoring signals for reliable hazard learning. Metrics include concordance index for hazard prediction, failure rate at fixed success, constraint-violation frequency under budget limits, and cost per successful

completion. Together, these measures expose whether an agent improves safety calibration rather than merely shifting failures into rarer but more severe tail events. By combining survival modeling with budget-aware reinforcement learning, SAARL captures the temporal pattern of failure risk in long-horizon web automation and offers a structured path toward safer, more deployable LLM-driven agents in dynamic online environments.

2.Materials and Methods

2.1 Sample and Study Setting

The study used a web interaction benchmark containing 1,800 task instances. The tasks covered e-commerce transactions, account operations, document submission, and structured information retrieval. Each task required 8 to 25 sequential steps and included at least one irreversible action, such as final submission or credential confirmation. A total of 24,000 episodes were generated under fixed cost budgets. All episodes were executed in a controlled browser environment with stable network latency and fixed rendering settings. For each episode, state–action trajectories, reward signals, cumulative cost values, and terminal outcomes were recorded. Terminal outcomes were categorized as success, failure, or truncation due to budget limits. Failure events were defined as discrete terminal states caused by incorrect irreversible actions or system lockouts. Episodes stopped by budget limits without failure were treated as censored observations.

2.2 Experimental Design and Control Settings

A comparative design was used to evaluate the Survival-Aware Agentic Reinforcement Learning (SAARL) framework against baseline methods. The experimental group included agents trained with a hazard-based objective that accounted for cumulative failure risk and budget constraints. The control group included three settings: standard reward-based RL, constrained RL using Lagrangian relaxation without hazard estimation, and a risk-penalty RL model with fixed stepwise failure costs. All models used the same state encoder, policy structure, and action space. Token limits and API cost budgets were identical across groups. Each configuration was trained using the same random seed strategy and evaluated on a separate test set of 400 tasks. This setup ensured that performance differences were related to the modeling of failure risk rather than model architecture or resource allocation.

2.3 Measurement Methods and Quality Control

Evaluation metrics included task success rate, failure rate under matched success levels, cost per successful task, and concordance index (C-index) for hazard prediction. The C-index measured the consistency between predicted hazard scores and observed time-to-failure ordering. Hazard values were estimated at each interaction step using the model output. All experiments were repeated five times with independent training runs. Mean values and standard deviations were reported. Episodes affected by browser errors or incomplete logs were excluded based on predefined screening rules. Model stability was checked by monitoring training loss and validation performance. Calibration quality was assessed using reliability curves that compared predicted cumulative hazard with observed failure frequency across quantile groups.

2.4 Data Processing and Model Formulation

State–action sequences were encoded into fixed-length vectors using a transformer-based encoder. For each episode i, let T_i denote the step index at failure or censoring, and let $\delta_i \in \{0,1\}$ indicate whether failure occurred. The hazard function conditioned on state–action pair $(s_t,a_t)$was defined as

$$h(t|s_t,a_t)=h_0(t)\exp\left(\beta^{\top}\phi(s_t,a_t)\right),$$

where $h_0(t)$ is the baseline hazard and $\phi(s_t,a_t)$ is the encoded feature vector. Parameter estimation followed the partial likelihood approach based on observed failure events.

The policy objective incorporated a risk-adjusted return defined as

$$J(\pi)=E\left[\sum_{t=0}^{T} r_t - \lambda \sum_{t=0}^{T} \hat{h}_t - \sum_{k} \mu_k c_{k,t}\right],$$

Where r_t represents task reward, $\hat{h}_t$ is the predicted hazard at step t, $c_{k,t}$ denotes the k-th cost component, and μ_k are weighting parameters. Episodes terminated due to budget limits contributed censored terms to the likelihood function but were not treated as failure events. Continuous variables were standardized before training. Hyperparameters were selected using validation-based tuning to ensure stable convergence and comparable model performance.

3.Results and Discussion

3.1 Completion stability under fixed budgets

Under identical token and API-call limits, SAARL achieved more stable task completion than reward-based RL and standard constrained RL. The advantage was most evident in tasks with more than 18 interaction steps. In these settings, baseline agents often reached the final submission stage with limited remaining budget and showed a higher probability of irreversible errors. SAARL reduced this problem by selecting lower-risk actions in earlier steps, which limited the accumulation of risky transitions before critical boundaries. As shown in Fig.1, the risk–efficiency frontier shifted toward lower failure probability at comparable success levels, without a marked increase in cost per successful completion. This pattern indicates that the hazard-based objective improves safety while preserving efficiency [20,21].

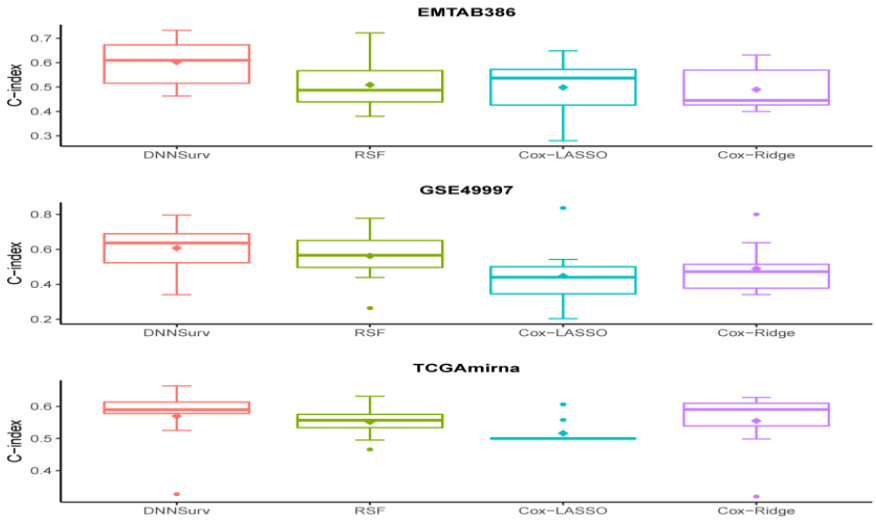


Figure 1.Risk–efficiency frontier showing failure rate versus task success rate under identical interaction budgets for different training methods.

3.2 Step-level risk distribution across interaction phases

The learned hazard values provided a clear description of how failure risk changed over time. During early information collection and reversible navigation, hazard estimates remained low and stable. Near authentication checks and final confirmation steps, hazard values increased sharply. In contrast, baseline models with fixed penalties showed flatter risk patterns and weaker separation between reversible preview states and true submission states. This distinction is important in web environments because visually similar pages may differ in whether actions are reversible [22,23]. The time-indexed survival curves presented in Fig.2 illustrate how cumulative survival probability decreases as the interaction approaches irreversible stages, allowing direct inspection of risk concentration across steps.

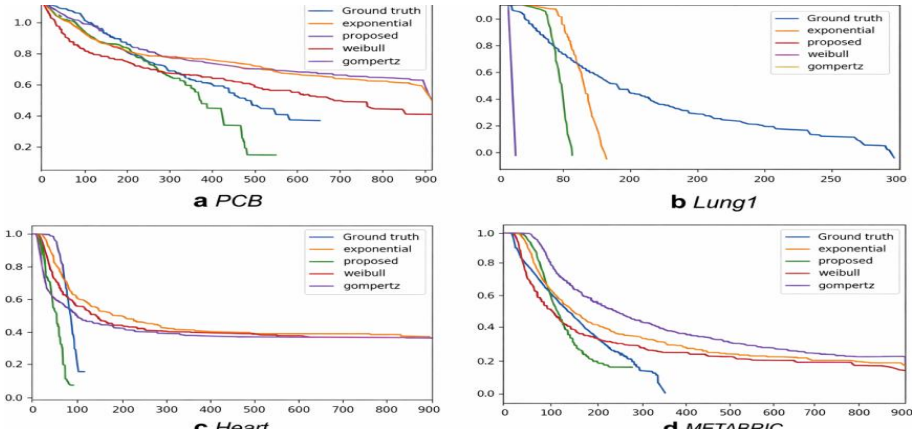


Figure 2:Kaplan–Meier curves showing the change in failure probability over interaction steps for the survival-aware agent and baseline models.

3.3 Hazard calibration and ranking performance

Hazard prediction accuracy was evaluated using concordance index and calibration analysis. SAARL produced higher concordance values, indicating better agreement between predicted risk ordering and observed time-to-failure. Reliability analysis showed that predicted cumulative hazard aligned more closely with empirical failure frequency across quantile groups. Baseline models that relied on fixed step penalties or expected cost constraints tended to underestimate late-stage risk. Incorporating censored episodes from budget truncation improved estimation stability because these trajectories still provide information about survival up to the stopping point [24,25]. This approach reduced variance in hazard estimation under limited event counts and supported consistent risk ranking across evaluation runs.

3.4 Relation to existing web-agent research and remaining challenges

Most recent web-agent studies focus on stronger reasoning ability, improved page representation, or enhanced exploration strategies. Safety is commonly reported as overall failure rate or rule violation frequency. The present results show that modeling time-to-event risk adds a complementary perspective by identifying where failure probability increases within a trajectory. The hazard signal supports clearer interpretation and can guide risk-aware stopping near high-risk boundaries. However, several challenges remain. Reliable hazard learning requires adequate numbers of failure and censored episodes; rare failure types may require additional sampling. Risk estimates may shift under interface changes or domain drift, which can affect calibration. Future studies should therefore combine success metrics with calibrated risk evaluation under realistic budget conditions to obtain a balanced assessment of performance and safety.

4.Conclusion

This study presented a survival-aware reinforcement learning framework for web automation, where failure was treated as a time-to-event outcome rather than as a single penalty signal. By combining a Cox-type hazard module with a risk-adjusted objective and censoring treatment under budget limits, the method improved task completion stability while maintaining fixed interaction costs. The results showed lower late-stage irreversible errors, stronger agreement between predicted risk ranking and observed failure order, and stable efficiency across different budget levels. The step-level hazard estimates provided a clear description of how failure probability changed during interaction, which supports safety monitoring and risk-based control in long-horizon web tasks. The framework can be extended to other sequential decision problems that involve discrete terminal events, such as automated workflow execution and interactive service systems. Limitations remain. Reliable hazard learning requires sufficient failure and censored samples, and risk calibration may weaken under interface changes or domain shifts. Further work should focus on rare-event coverage, model recalibration across domains, and larger-scale validation to improve robustness and generalizability.

References

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).

Rahman, M., Bhuiyan, A., Islam, M. S., Laskar, M. T. R., Mahbub, R., Masry, A., ... & Hoque, E. (2025). Llm-based data science agents: A survey of capabilities, challenges, and future directions. arXiv preprint arXiv:2510.04023.

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.

Chambers, C. R. (2024). Nonmonetary reward systems, counterproductive behavior, and responses to sanctions in open collaboration environments. Organization Science, 35(3), 928-947.

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.

Thiyagarajan, G., Bist, V., & Nayak, P. (2025). Securing Credit Inquiries: The Role of Real-Time User Approval in Preventing SSN Identity Theft. arXiv preprint arXiv:2505.18861.

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.

Baduwal, M., & Paudel, P. (2026). Evaluating Agentic Artificial Intelligence: A Comprehensive Survey of Metrics, Benchmarks, and Methodologies. Authorea Preprints.

Mao, Y., Ma, X., & Li, J. (2025). Research on API Security Gateway and Data Access Control Model for Multi-Tenant Full-Stack Systems.

Enkhsaikhan, B., & Jo, O. (2025). Risk-constrained reinforcement learning with augmented lagrangian multiplier for portfolio optimization. IEEE Transactions on Big Data, 11(5), 2489-2502.

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.

Alomrani, M. A., Zhang, Y., Li, D., Sun, Q., Pal, S., Zhang, Z., ... & Coates, M. (2025). Reasoning on a budget: A survey of adaptive and controllable test-time compute in llms. arXiv preprint arXiv:2507.02076.

Li, T., Xia, J., Liu, S., & Jiang, Y. (2025). Digital Transformation of Human Resources: From Consulting Frameworks to AI-Enabled Learning Management Systems.

Grau-Escolano, J., Bassolas, A., & Vicens, J. (2024). Cycling into the workshop: e-bike and m-bike mobility patterns for predictive maintenance in Barcelona's bike-sharing system. EPJ Data Science, 13(1), 48.

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.

Yuan, T., Assilevi, K. R., Adjallah, K. H., Ajavon, A. S. A., & Wang, H. (2024). Microgrid resilience enhancement with sensor network-based monitoring and risk assessment involving uncertain data. Energies, 17(23), 6141.

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).

Krupp, L. A. R. S., Geißler, D. A. N. I. E. L., Woźniak, P. W., Lukowicz, P., & Karolus, J. (2025). Quantifying Web Agents-A Survey on Web Agent Performance and Efficiency.

Mao, Y., Ma, X., & Li, J. (2025). Research on Web System Anomaly Detection and Intelligent Operations Based on Log Modeling and Self-Supervised Learning.

Rahman, S. T., & Dhanekula, A. (2024). Market-Driven Management Strategies Using Artificial Intelligence To Strengthen Food Safety And Advance One Health Initiatives. International Journal of Scientific Interdisciplinary Research, 5(2), 377-414.

Li, T., Xia, J., Liu, S., & Hong, E. (2025). Strategic Human Resource Leadership in Global Biopharmaceutical Enterprises: Integrating HR Analytics and Cross-Cultural.

Chae, H., Kim, N., Ong, K. T. I., Gwak, M., Song, G., Kim, J., ... & Yeo, J. (2024). Web agents with world models: Learning and leveraging environment dynamics in web navigation. arXiv preprint arXiv:2410.13232.

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.

Morenz, E. (2023). Novel Statistical Methods for Causal Inference Based on Truncated and Censored Time-to-Event Data. University of Washington.

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.