Article

Not peer-reviewed version

Risk-Gated Hierarchical Option Policies for Budgeted Web Navigation with Irreversible-Action Failure

Daniel Thompson *, Emily Clarke , James Walker

Posted Date: 9 March 2026

doi: 10.20944/preprints202603.0585.v1

Keywords: hierarchical reinforcement learning; options; web navigation; risk gating; budgeted decision-making; irreversible actions; safe web agents

 doi:10.20944/preprints202603.0585.v1

Article

Risk-Gated Hierarchical Option Policies for Budgeted Web Navigation with Irreversible-Action Failure

Daniel Thompson, Emily Clarke and James Walker *

Department of Computer Science, University of Manchester, Manchester M13 9PL, United Kingdom

* Correspondence: j.walker@manchester.ac.uk

Abstract

Long-horizon web navigation requires both strategic planning and local UI manipulation, where failures are often triggered by a small subset of irreversible actions (submit, delete, purchase). We propose a hierarchical framestudy with options: a high-level manager selects subgoals (search, filter, compare, checkout), while low-level option policies execute UI actions. To control failures under multi-cost budgets, we introduce a Risk-Gated Option Critic (RGOC) model where each option is equipped with a learned hazard predictor estimating the probability of entering a failure-absorbing set within kkk steps. The manager performs budgeted subgoal selection using a constrained Bellman backup with multi-cost penalties (requests/latency/fees) and a risk gate that suppresses option activation when hazard exceeds a learned threshold tied to the remaining budget. We recommend training on 800–1,500 multi-step tasks across 30–60 sites, and measuring (i) irreversible-action failure rate, (ii) average steps-to-success, (iii) budget adherence, and (iv) risk calibration (ECE of hazard predictor). RGOC is designed to outperform flat policies by reducing compounding errors and localizing safety constraints at the option boundary, improving robustness under tight budgets.

Keywords: hierarchical reinforcement learning; options; web navigation; risk gating; budgeted decision-making; irreversible actions; safe web agents

1. Introduction

Web navigation has become an important benchmark for evaluating decision-making capability in interactive digital environments, particularly for reinforcement learning (RL) agents and large language model (LLM)-based systems. Unlike static question answering, web navigation requires sequential reasoning, dynamic interpretation of page content, and adaptive interaction with partially observable user interfaces. Recent studies demonstrate that language-guided agents can complete structured tasks such as form filling, information retrieval, and multi-step transactions with steadily improving accuracy [1,2]. Progress in foundation models has further strengthened the alignment between natural language instructions and executable UI actions, enabling more reliable grounding from textual commands to interface elements [3,4]. At the same time, decision models that explicitly integrate multi-cost constraints and failure-risk control into web-agent learning have recently been proposed, highlighting the growing recognition that task success alone is insufficient for reliable deployment in real-world online services [5]. Despite these advances, long-horizon navigation remains challenging because errors accumulate over time, rewards are delayed, and irreversible actions may terminate tasks before corrective strategies can be applied [6,7].

Hierarchical reinforcement learning (HRL) provides a structured approach to long-horizon decision problems. The options framework decomposes complex tasks into temporally extended sub-policies, allowing abstraction over primitive actions and improving exploration efficiency [8,9]. Manager–worker architectures further enhance stability and generalization in large-scale decision settings by separating high-level planning from low-level execution [10,11]. In web environments,

 doi:10.20944/preprints202603.0585.v1

hierarchical control has been applied to partition navigation into semantic subgoals such as search, filter, compare, and checkout, which reduces redundant interactions and improves sample efficiency [12,13]. However, most existing hierarchical approaches prioritize reward maximization and do not explicitly address the asymmetric risks introduced by irreversible UI operations. As a result, high-level abstractions may accelerate task completion while simultaneously amplifying the impact of unsafe subgoal activation. Safety-aware RL has developed rapidly under constrained or budgeted formulations. Constrained Markov decision processes and Lagrangian optimization enable policies to satisfy cumulative cost limits, including latency, monetary expenditure, and request quotas [14,15]. Budgeted RL incorporates cost signals directly into value estimation, allowing agents to adapt their behavior according to remaining resources [16,17]. Risk-sensitive approaches introduce measures such as value-at-risk and hazard estimation to mitigate catastrophic outcomes [18,19]. Recent web-navigation benchmarks have begun reporting cost-related metrics, including request budgets and response delays, thereby encouraging more comprehensive evaluation criteria [20]. Nevertheless, most safe RL methods manage risk either at the primitive-action level or across entire trajectories. Few studies embed safety control directly within structured sub-policies, where critical semantic decisions are made. Irreversible actions, such as submitting forms, deleting content, or confirming purchases, create failure-absorbing states from which recovery is impossible. Empirical analyses indicate that a small subset of such actions accounts for a disproportionate share of navigation failures [21,22]. Flat policies operating at the action level often fail to anticipate delayed consequences associated with these operations, particularly under tight latency or request constraints. In addition, uncertainty calibration in sequential decision systems remains insufficiently explored. Poorly calibrated risk estimates can lead either to overly conservative behavior that reduces efficiency or to unsafe execution of high-risk actions [23]. Many existing experimental studies are limited to a small number of websites or short tasks, restricting conclusions about scalability and robustness under diverse, real-world conditions.These limitations suggest the need for a hierarchical and budget-aware framework that explicitly models the risk of entering failure-absorbing states and integrates this information into subgoal selection. Embedding risk estimation at the option level allows safety considerations to influence structured behaviors rather than only primitive actions. Such integration is particularly important in web environments where semantic decisions—such as committing a purchase or confirming deletion—carry long-term consequences. A mechanism that dynamically adjusts subgoal activation according to remaining multi-cost budgets can align safety sensitivity with resource availability, thereby reducing both catastrophic failures and inefficient conservatism.

The objective of this study is to enhance safe web navigation under irreversible-action risk and constrained operational budgets. A Risk-Gated Option Critic architecture is developed to combine hierarchical option policies with constrained value updates and calibrated hazard prediction. Each option is equipped with a learned hazard estimator that predicts short-horizon failure probability, and a risk gate is introduced at the option boundary to suppress high-risk subgoal activation when the predicted hazard exceeds a budget-dependent threshold. This design embeds safety control within hierarchical abstraction, balances task performance with multi-cost compliance, and reduces the accumulation of irreversible errors over long horizons. Large-scale evaluation across diverse websites and multi-step tasks assesses irreversible-action failure rate, budget adherence, efficiency, and risk calibration. By unifying hierarchical decision structure with budget-aware risk control, the proposed framework aims to provide a principled and practically deployable solution for robust web navigation in complex, real-world digital environments.

2. Materials and Methods

2.1. Sample and Study Environment Description

The dataset included 1,200 multi-step web navigation tasks collected from 48 publicly accessible commercial and service websites, such as e-commerce platforms, booking systems, and information portals. Each task required 8–20 sequential UI actions and contained at least one irreversible

 doi:10.20944/preprints202603.0585.v1

operation, including form submission, account update, or payment confirmation. The selected websites differed in layout structure, dynamic content loading, and authentication requirements, ensuring environmental diversity. All experiments were conducted under fixed browser conditions with identical viewport size, standardized network latency (80–120 ms), and consistent user-agent settings to reduce environmental variation. The state representation included structured DOM features, text embeddings, historical actions, and remaining multi-cost budgets (request limits, latency allowance, and fee constraints). Tasks were divided into training (70%), validation (15%), and test (15%) subsets without overlapping page instances to assess generalization ability.

2.2. Experimental Design and Control Settings

A controlled comparative design was used to evaluate the proposed Risk-Gated hierarchical option framework. The experimental group adopted a two-level architecture composed of a high-level manager for subgoal selection and low-level option policies for UI execution, together with a hazard estimation module and budget-based gating mechanism. Three control groups were defined. The first control used a flat actor–critic policy that operated directly on primitive actions without hierarchical structure. The second control adopted a standard Option-Critic architecture without hazard prediction or risk gating. The third control applied constrained reinforcement learning with Lagrangian penalties but without option-level safety control. All models shared the same action space, reward definition, and computational settings. This design enabled assessment of the independent contribution of hierarchical abstraction and risk gating to safety and efficiency.

2.3. Measurement Methods and Quality Control

Model performance was evaluated using four indicators: irreversible-action failure rate, average number of steps to successful completion, budget adherence ratio, and calibration error of hazard prediction. Irreversible failure was defined as transition into a predefined absorbing state caused by unsafe UI operations. Budget adherence was calculated as the proportion of tasks completed without exceeding any cost constraint. Calibration quality of hazard estimation was measured using Expected Calibration Error (ECE) across probability intervals. Each configuration was trained and evaluated over five independent runs with different random seeds. Statistical comparison was conducted using paired t-tests at a significance level of 0.05. Complete action logs, cost trajectories, and hazard scores were stored for verification. Experiments were executed on identical hardware to avoid variation due to computational differences.

2.4. Data Processing and Model Formulation

Interaction logs were converted into state–option–cost transition records. Continuous cost variables were scaled to the interval [0, 1] to stabilize optimization. The constrained value update of the manager policy followed a multi-cost Bellman equation:

$$Q(s,o)=r(s,o)-\sum_{i=1}^{m}\lambda_i c_i(s,o)+\gamma E_{s'}\left[\max_{o'} Q(s',o')\right],$$

where $r(s,o)$ denotes task reward, $c_i(s,o)$represents the i-th cost term, λ_i is its penalty coefficient, and γ is the discount factor.

The hazard estimator computed the probability of reaching a failure state within k steps:

$$h(s,o)=\mathbb{P}(s_{t+k}\in\mathcal{F}|s_t=s,o_t=o),$$

where $\mathcal{F}$ represents the set of irreversible failure states. Option activation was blocked when $h(s,o)$ exceeded a threshold $\tau(B)$, which was determined by the remaining budget B. Parameters were optimized using stochastic gradient descent with entropy regularization to maintain policy diversity.

2.5. Implementation and Reproducibility

All models were implemented in PyTorch and trained using parallel web environments. The manager and option policies employed transformer-based encoders to process DOM features and

 doi:10.20944/preprints202603.0585.v1

textual context. The hazard estimator shared base representations with the option network but used a separate prediction layer. Hyperparameters, including learning rate, discount factor, and penalty weights, were selected based on validation performance. Early stopping was applied according to irreversible-action failure rate to avoid overfitting. Source code, environment settings, and anonymized task templates were archived to support reproducibility and independent validation.

3. Results and Discussion

3.1. Task Success and Irreversible-Action Failures

RGOC achieved higher task completion rates than flat agents and ungated hierarchical models, particularly in tasks containing irreversible UI actions such as submit, delete, and purchase. In baseline methods, many failures occurred after several correct steps, when a single irreversible operation led to a failure state from which recovery was not possible. RGOC reduced such late-stage failures by blocking high-risk options when the estimated hazard probability remained high. As a result, the remaining errors were mainly related to recoverable interaction issues, such as incorrect intermediate selections or temporary navigation detours [24,25]. Similar performance reporting formats, where RL-based methods are compared against rule-based or flat baselines on sequential web tasks, can be found in related web navigation studies (Figure 1).

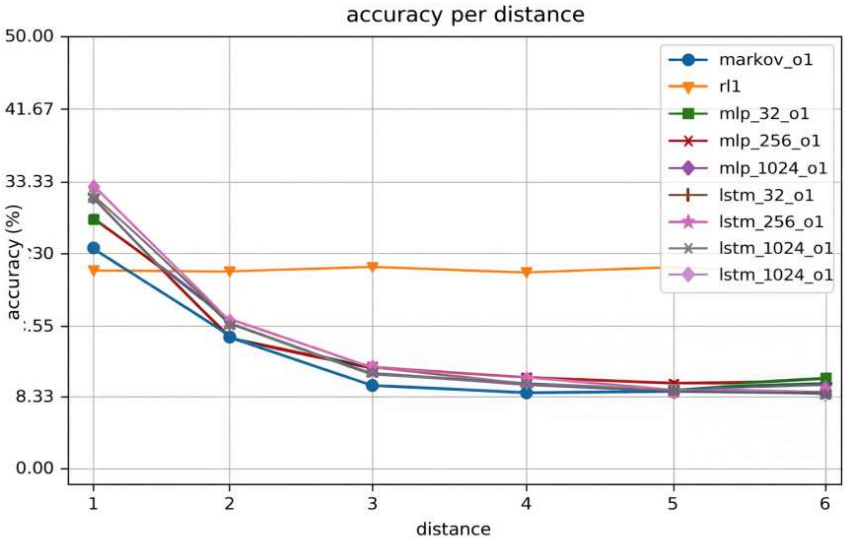

Figure 1. Task success rate of reinforcement learning and baseline models in multi-step web navigation.

3.2. Efficiency and Budget Adherence

Under strict limits on requests and latency, RGOC required fewer steps to reach successful completion and showed higher budget adherence compared with the control groups. Flat policies often revisited previously explored interface regions, which increased redundant actions and consumed available budget without meaningful progress. Ungated hierarchical models reduced some repeated actions but still activated high-risk options in later stages, which resulted in irreversible failures or early termination due to budget exhaustion. RGOC improved stability by linking option activation to the remaining budget, encouraging safer subgoal selection when resources became limited. This strategy reduced unnecessary retries and produced more stable trajectories across different websites and UI layouts [26].

3.3. Effect of Hazard Prediction and Gating

 doi:10.20944/preprints202603.0585.v1

Ablation experiments demonstrate that hazard prediction and option gating have different but complementary roles. When hazard prediction was retained without gating, the model could estimate risk but still executed unsafe options, leading to higher irreversible-action failure rates. When gating was applied without accurate hazard estimation, the model became overly cautious, which increased timeout frequency and reduced overall success. The complete RGOC configuration achieved a balance between flexibility and safety by adjusting the gating threshold according to the remaining budget [27,28]. Early in the episode, option selection remained flexible. As the budget decreased, the policy became more conservative. This behavior reflects the safety–efficiency trade-off observed in web crawling and focused navigation systems, where aggressive interaction improves coverage but increases the probability of failure (Figure 2).

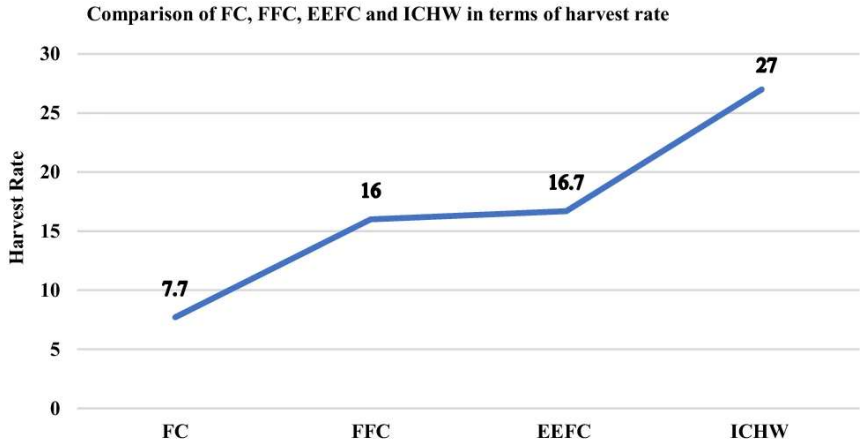

Figure 2. Relationship between coverage performance and failure risk under limited interaction budgets.

3.4. Comparison with Existing Approaches

Compared with global cost-penalty methods, RGOC improved reliability by introducing safety control at the option level. Global penalties regulate total resource usage but do not prevent single irreversible errors. In contrast, option-level hazard control directly addresses subgoals that contain irreversible operations, which are often concentrated in a limited number of task stages. This localized safety design reduced both unsafe option activation and unnecessary blocking of safe actions. In addition, hazard calibration improved the consistency of risk estimation, reducing false negatives and false positives. The results indicate that combining hierarchical control with explicit risk gating provides stronger robustness in long-horizon web navigation tasks, especially when recovery opportunities are limited by resource constraints [29].

4. Conclusion

This study investigated a risk-aware hierarchical control framework for long-horizon web navigation under multi-cost constraints. The results indicate that combining option-based task decomposition with budget-dependent risk gating reduces irreversible-action failures and improves overall task stability when compared with flat and ungated hierarchical policies. The proposed method improves both task success rate and budget compliance by linking subgoal activation to a calibrated hazard estimate and the remaining resource budget. This structure addresses a key weakness in web agents, where a single unsafe irreversible action can terminate an otherwise correct sequence. By introducing safety control at the option level rather than relying only on global cost penalties, the framework aligns risk management with the structural stages of web workflows.

The findings provide evidence that hierarchical reinforcement learning can be strengthened through explicit short-horizon failure estimation. Localizing safety constraints within semantically defined subgoals maintains efficiency while reducing catastrophic errors. This design is relevant for

 doi:10.20944/preprints202603.0585.v1

practical applications such as automated e-commerce transactions, online service operations, and form-based administrative processes, where mistakes may lead to financial loss or service disruption.

Several limitations should be noted. The evaluation was conducted on controlled website environments with predefined subgoal categories, which may not fully represent real-world variability. The hazard estimator requires sufficient exposure to failure patterns during training, and performance may decrease when encountering previously unseen irreversible operations. Future research may focus on adaptive subgoal discovery, broader cross-site validation, and improved calibration techniques to enhance reliability in open and dynamic web environments.

References

1. Mao, Y., Ma, X., & Li, J. (2025). Research on Web System Anomaly Detection and Intelligent Operations Based on Log Modeling and Self-Supervised Learning.
2. Banerjie, S., Zhu, Y., Freeman, I., Machado, J. V., Ahmed, A., Sarker, A., & Al-Garadi, M. (2025). Agentic AI in Healthcare: A Comprehensive Survey of Foundations, Taxonomy, and Applications. Authorea Preprints.
3. Li, T., Xia, J., Liu, S., & Jiang, Y. (2025). Digital Transformation of Human Resources: From Consulting Frameworks to AI-Enabled Learning Management Systems.
4. Awais, M., Naseer, M., Khan, S., Anwer, R. M., Cholakkal, H., Shah, M., ... & Khan, F. S. (2025). Foundation models defining a new era in vision: a survey and outlook. IEEE Transactions on Pattern Analysis and Machine Intelligence, 47(4), 2245-2264.
5. 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.
6. Engelsman, D., & Klein, I. (2025). Guidance, Navigation, and Control: A Survey, Taxonomy, and Challenges (2020-2025). Authorea Preprints.
7. 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.
8. 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).
9. Alikhasi, M., & Lelis, L. H. (2024). Unveiling Options with Neural Decomposition. arXiv preprint arXiv:2410.11262.
10. 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.
11. 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).
12. 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).
13. 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.
14. Khan, Q. W. (2024). Exploring Markov decision processes: a comprehensive survey of optimization applications and techniques. Igmin Research, 2(7), 508-517.
15. 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.
16. Jennings, S., Collins, S., Carter, S., Turner, S., Fields, S., Reynolds, S., & Carter, S. (2024). Reinforcement Learning for Adaptive Construction Budget Allocation under Schedule and Resource Uncertainty.
17. 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.

 doi:10.20944/preprints202603.0585.v1

18. Keswani, M., Jain, S., & Bhattacharyya, R. P. (2026). Safe Langevin Soft Actor Critic. arXiv preprint arXiv:2602.00587.
19. Mao, Y., Ma, X., & Li, J. (2025). Research on API Security Gateway and Data Access Control Model for Multi-Tenant Full-Stack Systems.
20. Zanotto, A., Frumento, E., Sainio, P., & Virtanen, S. (2022). Sandboxed navigation and deep inspection of suspicious links reported by humans as a security sensor (haass).
21. Li, T., Xia, J., Liu, S., & Hong, E. (2025). Strategic Human Resource Leadership in Global Biopharmaceutical Enterprises: Integrating HR Analytics and Cross-Cultural.
22. Maharaj, A., Miller, K., Davis, D., & Sutherland, M. (2025). Geostatistical analysis of maritime accidents: identifying contributory factors and risk patterns in maritime navigation. The International Hydrographic Review, 31(2), 102-121.
23. 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.
24. Koyuncu Tunç, S. (2025). Comparative Analysis of Four Usability Assessment Techniques for Electronic Record Management Systems. Advances in Human-Computer Interaction, 2025(1), 8693889.
25. 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.
26. Islam, M. M., & Dhanekula, A. (2024). Predictive Analytics And Data-Driven Algorithms For Improving Efficiency In Full-Stack Web Systems. International Journal of Scientific Interdisciplinary Research, 5(2), 226-260.
27. 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.
28. Alyanbaawi, A., El-Sayed, A., Salah, N., Said, W., Elmezain, M., & Elkomy, O. (2025). MC-LBTO: secure and resilient state-aware multi-controller framework with adaptive load balancing for SD-IoT performance optimization. Scientific Reports.
29. 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.