e-ISSN: 3079-6377 p-ISSN: 3079-6369

Pareto-Front Agentic RL with Dynamic Preference Conditioning for Cost–Risk–Success Trade-offs in Web Tasks

Marco Rossi[1], Giulia Bianchi[2], Alessandro Conti[3*]

Department of Information Engineering, University of Padua, 35131 Padua, Italy

Corresponding author: a.conti@unipd.it

Abstract

Web agent deployment requires navigating trade-offs among success rate, monetary/API costs, latency, and failure risk, which vary by user and scenario. We propose a preference-conditioned multi-objective RL framestudy that learns a Pareto set of policies for web tasks. The method trains a single agent conditioned on a preference vector w over (success, cost1_11, cost2_22, ..., risk), using (i) multi-gradient conflict resolution to stabilize updates across objectives, and (ii) Pareto replay that balances samples from distinct regions of the frontier. To ensure tail-risk control, the risk objective is defined as quantile-based failure loss (e.g., 90th/95th percentile). We recommend benchmarking on 1,200+ tasks across 40–70 site templates, sweeping preference vectors to obtain a Pareto curve and reporting hypervolume improvement, frontier coverage, and policy switching stability. This approach enables "one model, many operating points," supporting practical deployment where budgets and risk tolerance change dynamically.

Keywords

Multi-objective reinforcement learning; Pareto optimality; preference conditioning; web agents; risk quantiles; cost–latency trade-off; policy calibration

1.Introduction

Web agents are increasingly deployed to accomplish multi-step tasks across heterogeneous websites, including information retrieval, form submission, reservation booking, and online transactions. Advances in large language models and tool-integrated agents have substantially improved task completion by enabling structured reasoning, API invocation, and adaptive planning in dynamic web environments [1, 2]. In particular, recent reinforcement learning–based decision frameworks for web agents under multi-cost and failure risk constraints demonstrate that explicit modeling of operational trade-offs can significantly enhance stability and deployment flexibility [3]. Benchmark environments such as Web Shop and Mind2Web further provide standardized settings for evaluating generalization across diverse site templates and task distributions [4, 5]. Despite these advances, practical deployment still faces inherent trade-offs among success rate, API or monetary cost, response latency, and failure risk. Empirical evidence indicates that higher success often relies on increased sampling, longer reasoning chains, or more frequent tool calls, which inevitably raise computational cost and delay [6, 7]. Because user requirements and task complexity vary

e-ISSN: 3079-6377 p-ISSN: 3079-6369

substantially, a single scalar evaluation metric is insufficient to reflect real-world operational constraints. Systems optimized solely for average success may incur unacceptable latency or risk levels in safety-critical or cost-sensitive scenarios. Multi-objective reinforcement learning (MORL) provides a principled framework for modeling competing objectives [8, 9]. Conventional scalarization methods combine objectives through weighted sums, yet they may fail to recover non-convex regions of the Pareto frontier and can suffer from gradient interference during optimization [10, 11]. More recent approaches introduce gradient adjustment mechanisms and preference-conditioned policies to approximate a continuous Pareto set within a single model, thereby improving sample efficiency and adaptability across varying objective trade-offs [12]. In parallel, distributional and quantile-based reinforcement learning techniques enable explicit control of tail risk by optimizing high-percentile outcomes rather than mean returns, offering a more realistic characterization of system reliability [13, 14]. However, most MORL studies are validated in simulated control domains with low-dimensional state spaces. Applications to language-driven web interaction remain limited. Web environments are characterized by sparse rewards, delayed feedback, dynamic page structures, and stochastic transitions, all of which increase optimization difficulty and reduce training stability. Within existing web-agent research, cost and latency control are frequently implemented through heuristic budget limits or simple reward shaping rather than explicit Pareto modeling [15, 16]. Failure risk is typically measured as average failure rate, without modeling distributional variability or high-percentile failure outcomes that directly influence user trust and service-level guarantees [17, 18]. Furthermore, many empirical evaluations rely on a limited number of site templates or several hundred tasks, which constrains statistical robustness and obscures performance variation under different preference configurations [19]. These limitations highlight the necessity of scalable and stable learning frameworks that can flexibly adapt to heterogeneous cost–risk–performance requirements. This study develops a preference-conditioned multi-objective reinforcement learning framework for web task optimization. The framework jointly optimizes task success, multiple cost components, latency, and quantile-based failure risk within a unified training process. A single policy is conditioned on a continuous preference vector, enabling dynamic behavioral adaptation under diverse operational constraints without retraining separate models. To mitigate instability arising from conflicting optimization directions, a multi-gradient adjustment strategy is incorporated to harmonize objective updates. A Pareto-aware replay mechanism is further designed to balance sampling across different regions of the frontier, enhancing coverage and reducing bias toward dominant objective combinations. Failure risk

e-ISSN: 3079-6377 p-ISSN: 3079-6369

is modeled through quantile-based loss, allowing explicit control over high-percentile adverse outcomes and aligning optimization objectives with reliability requirements in practical deployments. Extensive large-scale benchmarking over more than one thousand tasks and multiple heterogeneous site templates is conducted to approximate the Pareto frontier under systematically varied preference vectors. Performance is evaluated using hypervolume improvement, frontier coverage, and policy switching stability, thereby assessing both efficiency and robustness under dynamic preference changes. By integrating multi-objective optimization, preference conditioning, gradient coordination, and quantile-based risk control into a unified framework, this work establishes a scalable and deployment-oriented solution for web agents operating under complex cost–risk–success trade-offs, contributing both methodological advances and practical guidance for reliable large-scale web automation.

2. Materials and Methods

2.1 Sample and Study Setting

The dataset contained 1,248 web task cases derived from 56 website templates, including online retail, travel booking, information services, and account management pages. Each task required several interaction steps such as page navigation, structured input, or information retrieval. The web environments reflected practical conditions, including different layouts, partial visibility of states, response delay, and occasional interface changes. All experiments were carried out in a controlled cloud setting with fixed computing resources. Model structure, decoding parameters, and maximum interaction steps were kept the same for all experiments. Each episode was limited to 30 actions. For every episode, task outcome, API or monetary cost, execution time, and failure-related loss were recorded. The dataset was divided into training (70%), validation (15%), and testing (15%) subsets based on website templates to avoid overlap in structural patterns across splits.

2.2 Experimental Design and Control Groups

The proposed approach used a preference-conditioned multi-objective reinforcement learning model. A single policy was trained with a continuous preference vector that assigned weights to success, cost, latency, and risk. During optimization, gradient adjustment was applied to reduce conflicts among objective signals. A replay strategy was introduced to ensure balanced sampling across different trade-off regions. Two baseline models were included. The first baseline applied a fixed weighted-sum reward without preference conditioning, representing a standard multi-objective method. The second baseline optimized only task success and applied simple limits on cost and latency. All models were trained under

e-ISSN: 3079-6377 p-ISSN: 3079-6369

the same environment settings and interaction constraints. This design enabled comparison of performance stability, trade-off diversity, and risk control across methods.

2.3 Measurement Methods and Quality Control

Task success was defined as full completion of predefined goals verified by rule-based checks. Monetary or API cost was calculated as the total number of tool calls multiplied by a fixed unit price. Latency was measured as total wall-clock time per episode under identical hardware conditions. Failure risk was measured using the empirical upper quantile of episode loss. Each policy was evaluated using five independent random seeds. Mean values and standard deviations were reported. Episodes affected by system-level errors not caused by the agent were removed according to predefined criteria. All interaction steps, actions, and reward components were recorded for verification. Statistical comparison across models used paired bootstrap resampling with 1,000 repetitions to control variance.

2.4 Data Processing and Model Formulation

Let $w=(w_s,w_c,w_l,w_r)$represent a normalized preference vector for success, cost, latency, and risk. The overall return under a given preference was defined as

$$J(\pi,w)=E_{\tau\sim\pi}\left[w_sR_s(\tau)-w_cC(\tau)-w_lL(\tau)-w_rQ_\alpha(\tau)\right],$$

Where $R_s(\tau)$ denotes success reward, $C(\tau)$ represents total cost, $L(\tau)$denotes latency, and $Q_\alpha(\tau)$ is the empirical upper α-quantile of failure loss.

Pareto performance was assessed using hypervolume (HV) relative to a reference point z^{ref}:

$$HV=\int_{z\in P}\prod_{i=1}^{m}(z_i^{ref}-z_i)\,dz,$$

Where P is the set of non-dominated solutions and m is the number of objectives. All objective values were normalized before calculation. Data preprocessing included scaling rewards to unit variance and removing extreme values beyond the 1st and 99th percentiles. Convergence trends were presented using moving average smoothing.

2.5 Evaluation Protocol and Stability Assessment

Evaluation was conducted using 40 preference vectors sampled from a Dirichlet distribution over the objective simplex. For each preference configuration, the trained policy was tested on the independent test set. Frontier coverage was defined as the proportion of non-dominated solutions identified relative to all candidate solutions. Policy switching stability was measured by performance changes between adjacent preference settings. Convergence was examined using rolling averages over fixed training intervals. This evaluation procedure enabled

e-ISSN: 3079-6377 p-ISSN: 3079-6369

systematic assessment of trade-off flexibility, risk management, and operational stability under varying deployment preferences.

3.Results and Discussion

3.1 Pareto Frontier Shape and Operating-Point Stability

Preference conditioning produced a wider and more evenly distributed Pareto frontier across success, cost, latency, and risk compared with fixed-weight baselines. Scalarized training tended to cluster solutions around several dominant weight settings, while the conditioned policy generated non-dominated solutions across a broader range of trade-off combinations [20,21]. The largest improvements appeared in balanced regimes, where moderate success requirements were combined with strict cost and latency limits. In these cases, the agent reduced unnecessary tool calls and shortened interaction paths without a sharp decline in completion rate. This observation is consistent with recent reviews of tool-using agents, which describe deployment as a process of selecting operating points rather than optimizing a single score (Fig.1).

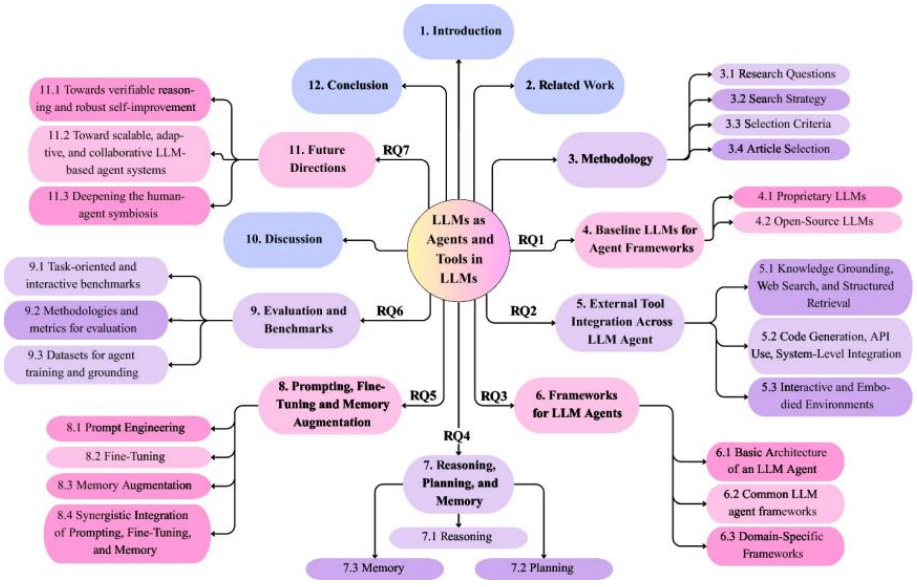

Figure 1 Framework illustrating preference-conditioned policy adjustment across different cost–risk–success trade-off settings.

3.2 Effect of Conflict-Reduced Updates and Balanced Sampling

When gradient conflict handling was removed, training became less stable, especially in regions where higher success required increased cost or longer execution time. Under fixed-weight optimization, the agent frequently shifted between aggressive exploration and conservative early stopping, which resulted in irregular frontier shapes and unstable behavior when preference weights changed slightly. The inclusion of conflict-reduced updates improved convergence consistency and reduced oscillation across objectives. In addition,

e-ISSN: 3079-6377 p-ISSN: 3079-6369

frontier-balanced replay increased exposure to trajectories from less frequent preference settings. This mechanism improved hypervolume values and increased the density of non-dominated solutions [22,23]. Compared with common practice in web-agent systems, where budgets are manually adjusted for each setting, the proposed approach maintained smoother transitions between adjacent preferences, indicating improved adaptability across deployment scenarios.

3.3 Tail-Risk Reduction with Quantile-Based Objectives

The quantile-based risk objective reduced high-percentile failure loss more effectively than mean-based penalties. On templates with unstable page behavior or delayed responses, average failure rates did not reflect rare but costly breakdowns. The quantile formulation limited these extreme outcomes by discouraging action sequences that occasionally triggered cascading retries or stalled navigation. This behavior follows the principle of risk-sensitive reinforcement learning, which emphasizes control of upper-tail losses rather than average performance (Fig.2). Policies trained with quantile risk showed lower variability in worst-case outcomes while maintaining competitive average success rates [24].

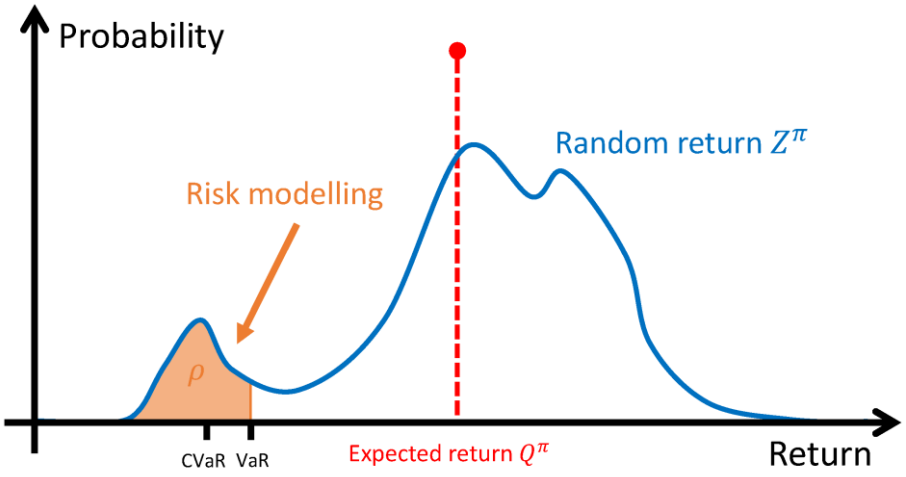

Figure 2 Diagram showing Value-at-Risk (VaR) and Conditional Value-at-Risk (CVaR) for controlling upper-tail losses in reinforcement learning.

3.4 Comparison with Existing Approaches and Practical Implications

Compared with single-objective web-agent methods that rely on fixed tool-call budgets or manual tuning, preference-conditioned multi-objective learning supports flexible adjustment of cost and risk tolerance without retraining separate models. In contrast to many multi-objective reinforcement learning studies conducted in simplified simulation environments, the present web-based evaluation demonstrates that large task diversity and template variation are essential for reliable frontier estimation [25]. Two practical constraints remain. First, performance depends on accurate preference specification, which may require

e-ISSN: 3079-6377 p-ISSN: 3079-6369

additional calibration in real systems. Second, stronger tail-risk control may increase conservatism under strict latency limits. Despite these limitations, the results indicate that a single conditioned policy can operate at multiple trade-off points, reducing the need for repeated model optimization and improving deployment efficiency under changing operational requirements.

4.Conclusion

This study investigated a preference-conditioned multi-objective reinforcement learning framework for web agents that must balance task success, cost, latency, and failure risk. The results indicate that conditioning a single policy on a continuous preference vector allows stable adjustment across different trade-off settings without training separate models. The use of gradient conflict reduction improved optimization stability, and frontier-balanced sampling increased the range and density of non-dominated solutions. In addition, quantile-based risk objectives reduced upper-tail failure outcomes and improved reliability in dynamic web environments. These findings demonstrate that web-agent deployment should be treated as a structured trade-off problem rather than a single-metric optimization task. The proposed framework supports flexible control of budget and risk tolerance in practical applications and is suitable for scenarios where operating requirements change over time. However, performance relies on appropriate selection of preference weights, and stronger tail-risk control may lead to more conservative behavior under strict latency limits. Further research may focus on automatic preference calibration and improved robustness to changes in website structure and task distribution.

References

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

[2] Anderson, A. S., & Adams, M. K. Designing Resource-Efficient Agentic AI: Architectural Patterns for Deploying Small Language Models in Autonomous Decision Systems.

[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] Peeters, R., Steiner, A., Schwarz, L., Caspary, J. Y., & Bizer, C. (2025). WebMall--A Multi-Shop Benchmark for Evaluating Web Agents [Technical Report]. arXiv preprint arXiv:2508.13024.

[5] 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.

e-ISSN: 3079-6377 p-ISSN: 3079-6369

[6] Stanley-Marbell, P., Alaghi, A., Carbin, M., Darulova, E., Dolecek, L., Gerstlauer, A., ... & Zufferey, D. (2020). Exploiting errors for efficiency: A survey from circuits to applications. ACM Computing Surveys (CSUR), 53(3), 1-39.

[7] 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).

[8] Nguyen, T. T., Nguyen, N. D., Vamplew, P., Nahavandi, S., Dazeley, R., & Lim, C. P. (2020). A multi-objective deep reinforcement learning framework. Engineering Applications of Artificial Intelligence, 96, 103915.

[9] 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.

[10] Ghane-Kanafi, A., & Khorram, E. (2015). A new scalarization method for finding the efficient frontier in non-convex multi-objective problems. Applied Mathematical Modelling, 39(23-24), 7483-7498.

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

[12] Janmohamed, H., Faldor, M., Pierrot, T., & Cully, A. (2024). Preference-Conditioned Gradient Variations for Multi-Objective Quality-Diversity. ACM Transactions on Evolutionary Learning.

[13] 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).

[14] Khatun, Z. (2025). Hybrid Digital Twin and Monte Carlo Simulation For Reliability Of Electrified Manufacturing Lines With High Power Electronics. International Journal of Scientific Interdisciplinary Research, 6(2), 143-194.

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

[16] 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.

[17] 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.

[18] Hasan, M. M., & Islam, M. M. (2023). Reinforcement Learning Approaches to Optimize IT Service Management Under Data Security Constraints. American Journal of Scholarly Research and Innovation, 2(02), 373-414.

e-ISSN: 3079-6377 p-ISSN: 3079-6369

[19] 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.

[20] Van Moffaert, K., Drugan, M. M., & Nowé, A. (2013, April). Scalarized multi-objective reinforcement learning: Novel design techniques. In 2013 IEEE symposium on adaptive dynamic programming and reinforcement learning (ADPRL) (pp. 191-199). IEEE.

[21] 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.

[22] Sato, H., Aguirre, H. E., & Tanaka, K. (2007, March). Controlling dominance area of solutions and its impact on the performance of MOEAs. In International conference on evolutionary multi-criterion optimization (pp. 5-20). Berlin, Heidelberg: Springer Berlin Heidelberg.

[23] 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.

[24] Bodnar, C., Li, A., Hausman, K., Pastor, P., & Kalakrishnan, M. (2019). Quantile qt-opt for risk-aware vision-based robotic grasping. arXiv preprint arXiv:1910.02787.

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