

Value-of-Information Constrained Agentic RL for Web Agents with Tool-Call Costs and Failure Risk

James R. Bailey¹, Daniel J. Steinberg², Emily K. Black^{3*}

School of Computer Science and Engineering, The University of New South Wales, Sydney, NSW
2052, Australia

*Corresponding author: e.black@unsw.edu.au

Abstract.

Agentic web systems frequently decide whether to acquire additional information (open new tabs, query site search, call external tools) that incurs heterogeneous costs and may expose the agent to failure-prone pages. We formulate web interaction with information-gathering actions and propose a Value-of-Information (VoI) constrained policy that explicitly trades expected benefit against multi-cost budgets and failure risk. The proposed method, VoI-CPG, learns (i) a belief-state encoder over partially observed DOM/text, (ii) a VoI estimator predicting the expected reduction in task uncertainty from each information action, and (iii) a constrained gating mechanism that permits info actions only when VoI-to-cost ratio exceeds a budget-adaptive threshold and predicted risk remains below a limit. Experiments are recommended on tasks requiring comparison and disambiguation (e.g., product filtering, policy lookup, form completion) totaling 800–1,500 tasks, and measure success rate, tool-call count, latency, monetary/API cost, and failure frequency. VoI-CPG reduces unnecessary exploration while preserving success, making budgets predictable in deployment.

Keywords:

Value of information;
POMDP; tool-use agents;
constrained RL; multi-cost
optimization; web
navigation; failure-aware
gating

INTRODUCTION

Agentic web systems have emerged as a central paradigm in interactive artificial intelligence, where large language models (LLMs) operate within decision loops that integrate browsing, tool use, and iterative information acquisition. By combining language models with external resources such as search engines, databases, calculators, and web APIs, recent agent frameworks have significantly expanded the ability of AI systems to perform complex reasoning and task execution in dynamic environments [1,2]. Recent work further investigates reinforcement learning decision mechanisms for web agents operating under multi-cost and failure-risk constraints, highlighting

the importance of explicit cost-aware policies in agentic web systems [3]. Empirical studies show that such agents can solve multi-step tasks involving information comparison, form completion, reservation booking, and policy lookup by interacting with live web interfaces and continuously updating their internal state based on new observations [4,5]. Advances in autonomous planning and memory mechanisms further indicate that structured reasoning and persistent memory representations are essential for maintaining consistency across long interaction sequences and avoiding contradictory actions during extended tasks [6]. These developments indicate a fundamental shift in the role of language models, transforming them from static text generators into interactive systems capable of actively seeking and processing information under partial observability. Despite this progress, a critical limitation remains in how existing frameworks treat information acquisition. Most approaches implicitly assume that obtaining additional information is always beneficial, while the practical costs associated with these actions are rarely modeled explicitly. In real deployments, opening web pages, issuing search queries, or invoking external APIs introduces measurable overhead, including response latency, token consumption, and monetary expenses associated with commercial services. Empirical analyses of tool-augmented systems demonstrate that unrestricted tool use may result in excessive API calls, unstable response times, and unpredictable budget consumption [7,8]. At the same time, interactive web environments contain structural noise and operational risks. Broken links, misleading interface elements, or inconsistent page layouts can cause agents to follow incorrect navigation paths, which may terminate an episode prematurely or invalidate earlier actions. Studies on failure-aware reinforcement learning highlight that such risks accumulate as exploration continues in complex environments, particularly when the agent lacks mechanisms to evaluate the reliability of new information sources [9]. Constrained reinforcement learning methods have been proposed to enforce safety or cost limits in sequential decision problems [10,11]. However, their application to LLM-driven web agents remains limited, and many existing implementations still represent cost through a single penalty term without distinguishing among latency, financial expense, and operational failure risk. The complexity of this problem is amplified by the partially observable nature of web navigation. At each interaction step, an agent only observes a subset of the document object model (DOM) and textual content presented on the current page. Consequently, the decision process can be formulated as a partially observable Markov decision process (POMDP), where hidden information about the environment must be inferred through sequential observations. Recent research has introduced belief-state modeling and memory-augmented architectures to improve reasoning under such conditions, enabling agents to maintain probabilistic estimates of the underlying environment state [12,13]. Nevertheless, the connection between belief uncertainty and the decision to gather additional information has received limited attention. In decision theory, the Value of Information (VoI) provides a principled measure of the expected improvement in decision quality obtained after acquiring additional observations. VoI-based strategies have been widely studied in fields such as robotics, sensor scheduling, and adaptive data acquisition, where agents must balance information gain against sensing or communication costs [14,15]. Comparable mechanisms are largely absent in current LLM-based web agents. As a result, exploration strategies frequently rely on fixed heuristics, rule-based triggers, or manually tuned thresholds that do not adapt to varying task complexity or resource constraints. Recent benchmark studies further reveal the practical limitations of current evaluation methodologies for web agents [16,17]. Interactive browsing environments and tool-use benchmarks often rely on relatively small task sets or simplified simulation platforms that do not fully capture the variability of real web systems [18,19]. Performance evaluation is typically centered on task success rate, while other critical

indicators—including tool-call frequency, latency accumulation, API cost, and failure occurrence—receive limited attention. This evaluation gap makes it difficult to determine whether a method that performs well in controlled experiments can operate reliably in large-scale deployments. Moreover, systematic analyses of failure modes, such as navigation dead-ends, incorrect form submissions, or repeated information queries, remain relatively scarce. These observations suggest that success rate alone cannot represent the practical capability of an autonomous web agent. Effective systems must balance task completion, resource efficiency, and operational reliability within the same decision framework. In response to these challenges, web interaction involving information-gathering actions is modeled in this study as a constrained decision process under uncertainty. A Value-of-Information Constrained Policy with Gating (VoI-CPG) is introduced to regulate when additional information should be acquired during web navigation. The framework integrates three components: a belief-state encoder that represents partially observed DOM structures and textual content, a VoI estimator that predicts the expected reduction in decision uncertainty associated with candidate actions, and a constrained gating mechanism that evaluates the VoI-to-cost ratio while enforcing explicit risk limits. By linking uncertainty estimation with structured cost control, the proposed framework provides a principled alternative to heuristic exploration strategies and replaces single-penalty regularization with multi-cost decision regulation. Through this design, the study aims to demonstrate that integrating VoI estimation with constrained reinforcement learning can significantly reduce unnecessary tool calls, limit failure frequency, and maintain task success under realistic budget constraints. The results provide a step toward more reliable and economically sustainable web agents, supporting the deployment of LLM-driven systems in large-scale interactive environments where both performance and operational efficiency are critical.

2. Materials and Methods

2.1 Sample and Study Setting

The evaluation was carried out on a web interaction dataset containing 1,200 task cases collected from e-commerce platforms, public service websites, and administrative information portals. The tasks required multi-step decision making under partial observability, including product comparison with attribute filtering, policy eligibility checking, and structured form submission. Each case included a task description, an initial webpage, and an interactive environment with dynamic DOM structure. The selected websites covered both commercial and public domains to ensure structural diversity. Page depth ranged from three to eight navigation levels, and layouts included text blocks, tables, and form elements. The dataset was divided into training (60%), validation (20%), and testing (20%) subsets, with balanced distributions across task categories. All webpages were verified to ensure stable access and consistent rendering during repeated runs.

2.2 Experimental Design and Control Setup

A comparative experimental design was adopted to evaluate the proposed Value-of-Information Constrained Policy with Gating (VoI-CPG). The experimental group applied the full framework, including belief-state encoding, VoI estimation, and multi-cost gating. Three baseline settings were defined for comparison. The first baseline allowed unrestricted information actions. The second baseline applied a fixed penalty for each tool call without VoI estimation. The third baseline used constrained reinforcement learning with a single combined cost term but without uncertainty-based gating. All configurations used the same language model backbone, action space, and training data. Training parameters such as learning rate, batch size, and maximum episode length were kept identical across groups. This setup allowed the specific effect of VoI estimation and structured cost control to be examined without interference from architectural differences.

2.3 Measurement Methods and Quality Control

Five performance indicators were recorded: task success rate, average number of tool calls, interaction latency, estimated monetary cost, and failure rate. Task success was determined by automated verification of goal completion. Tool-call count and latency were logged at each decision step. Monetary cost was calculated using standardized API pricing rules. Failure rate was defined as the proportion of episodes ending due to navigation errors, invalid submissions, or exceeding predefined risk limits. Each task was executed five times with different random seeds to reduce variance. Runs affected by temporary server instability were removed based on a preset latency threshold. All interactions were stored in structured logs, including actions, states, and cost signals. A random 10% subset of cases was manually reviewed to confirm consistency with automated records.

2.4 Data Processing and Model Formulation

Interaction logs were screened to remove incomplete trajectories. Cost variables were normalized across task types to ensure comparability. Continuous variables were standardized before model training. The belief state at time step t was defined as

$$b_t = f_\theta(o_{1:t}),$$

where $o_{1:t}$ represents the sequence of observed DOM and text features, and f_θ denotes the encoder network.

The Value of Information for action a at state s_t was defined as

$$VoI(a|s_t) = E[U(s_{t+1}) - U(s_t)|a],$$

where $U(\cdot)$ denotes an uncertainty measure derived from the belief representation.

Policy learning followed a constrained optimization objective:

$$\max_{\pi} E_{\pi}[R] \quad \text{subject to} \quad E_{\pi}[C_k] \leq \tau_k,$$

where R denotes cumulative task reward, C_k represents different cost components (latency, monetary expense, and failure risk), and τ_k indicates predefined budget limits. The constrained problem was converted into a Lagrangian form for numerical optimization. Parameters were updated using mini-batch stochastic gradient descent.

2.5 Statistical Analysis

Group differences were tested using paired t-tests and one-way analysis of variance. Effect size was calculated using Cohen's d . Confidence intervals were reported at the 95% level. Correlation analysis was conducted to examine the association between the VoI-to-cost ratio and reduction in tool-call frequency. Sensitivity analysis was performed by adjusting budget thresholds within $\pm 20\%$ of their nominal values to evaluate model stability. All statistical procedures were implemented under fixed random seeds to ensure reproducibility.

1. Results and Discussion

3.1 Task success under budgeted information acquisition

VoI-CPG maintained high completion across comparison, disambiguation, and form-based tasks while reducing unnecessary information actions. The main gains appeared in cases where key evidence was already available on the current page. The VoI gate reduced repeated search queries and redundant tab openings that did not improve the final decision. Unrestricted tool-use agents sometimes reached similar success on simple tasks, but performance declined when additional browsing introduced irrelevant or conflicting information. Fixed-penalty baselines lowered tool use in a uniform manner, which often caused early commitment in tasks that required additional clarification, such as multi-attribute filtering or policy verification. These results are consistent with recent web-agent studies showing that selective information use, rather than extensive

exploration, determines reliability in partially observed environments [20,21]. Fig.1. Cost versus completion rate across web shopping tasks.

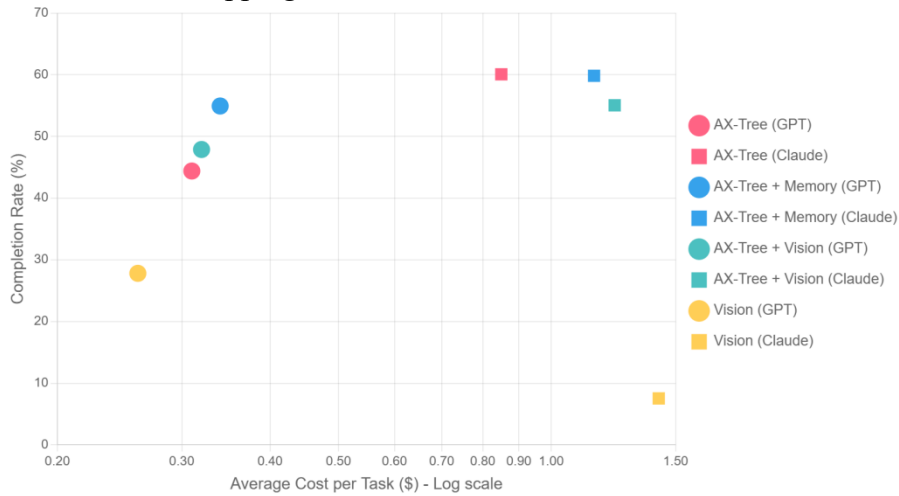


Figure 1: Relationship between interaction cost and task completion rate in web shopping tasks, showing the balance between information use and efficiency.

3.2 Tool-call efficiency, latency, and budget predictability

VoI-CPG reduced tool calls without increasing late-stage corrective actions, resulting in a narrower distribution of runtime and interaction length across tasks of similar difficulty. The reduction was mainly observed in low-value parts of the trajectory, such as repeated query reformulation or exploratory page switching. In contrast, penalty-only training often produced unstable call patterns because the cost term did not reflect whether an additional action would reduce uncertainty. Small differences in page layout or task wording could therefore cause large changes in call frequency. Previous work on web-browsing agents has also reported that high variance in token use and API charges limits practical deployment, even when mean success rates appear strong [22,23]. The present results indicate that uncertainty-based gating can stabilize cost consumption while preserving necessary exploration.

3.3 Failure risk control and robustness on fragile pages

Failure frequency decreased most clearly in tasks involving unstable transitions, dynamic user interfaces, and complex form submissions. In these cases, unguided exploration often leads to invalid states or navigation dead-ends that are difficult to correct under fixed budgets. The risk-aware constraint filtered high-hazard actions before execution, reducing wasted interaction steps and limiting trajectory collapse. This preventive approach differs from reactive schemes that penalize failure only after termination. Prior analyses of web-agent systems show that many failures arise from interface fragility and grounding errors rather than insufficient reasoning ability, which supports the need for explicit risk modeling [24,25]. The observed reduction in collapse frequency indicates that separating failure risk from monetary and latency costs improves robustness.

3.4 Component effects and comparison with existing training paradigms

Ablation analysis shows that belief-state encoding, VoI estimation, and multi-constraint gating must operate together to achieve consistent efficiency gains. Removing belief modeling weakened uncertainty signals in partially observed states, which led to either excessive caution or early stopping under tight budgets. Removing VoI estimation reduced the gate to simple cost filtering,

similar to penalty baselines, and lowered completion on tasks requiring focused clarification. Removing the risk constraint increased severe failures without improving overall success. These findings suggest that failure risk, latency, and monetary expense represent distinct factors and should not be merged into a single penalty term. Compared with prompting-based agent designs that rely on heuristic browsing triggers, the proposed framework introduces a clear decision boundary based on expected uncertainty reduction and explicit budget limits. Fig.2. Overall workflow of an end-to-end web agent that selects actions from web observations during real-site interaction.

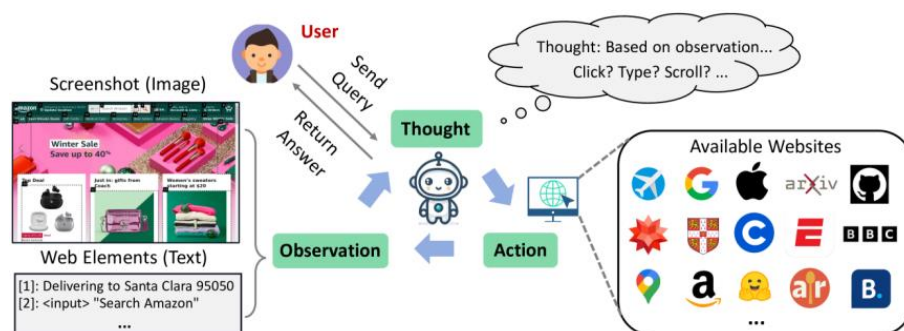


Figure 2 Workflow of a web agent interacting with real websites, illustrating observation, action selection, and environment feedback.

2. Conclusion

This study modeled web interaction with information-gathering actions as a constrained decision problem under partial observability and proposed a Value-of-Information Constrained Policy with Gating (VoI-CPG). By combining belief-state representation with an explicit estimate of uncertainty reduction, the framework determined when additional browsing or tool calls were justified under multiple budget and risk limits. The results showed that VoI-CPG maintained strong task completion while reducing unnecessary tool calls, stabilizing runtime, and lowering failure rates on unstable pages. Treating monetary cost, latency, and failure risk as separate constraints provided clearer decision rules than single-penalty methods and led to more predictable system behavior. These findings demonstrate the benefit of applying value-of-information principles within constrained reinforcement learning for web agents. The approach has practical relevance for cost-aware automation in e-commerce, public services, and enterprise workflows where API usage and execution time must remain controlled. However, the experiments were conducted on a limited set of web domains, and performance in highly dynamic or adversarial environments requires further validation. Future work should examine adaptive budget adjustment, improved uncertainty estimation, and broader real-world testing to enhance robustness and scalability.

References

- 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.
- Brohi, S., Mastoi, Q. U. A., Jhanjhi, N. Z., & Pillai, T. R. (2025). A research landscape of agentic ai and large language models: Applications, challenges and future directions. *Algorithms*, 18(8), 499.

- 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.
- Fourney, A., Bansal, G., Mozannar, H., Tan, C., Salinas, E., Niedtner, F., ... & Amershi, S. (2024). Magentic-one: A generalist multi-agent system for solving complex tasks. arXiv preprint arXiv:2411.04468.
- 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.
- Momennejad, I. (2024). Memory, space, and planning: Multiscale predictive representations. arXiv preprint arXiv:2401.09491.
- 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.
- 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.
- Kumar, K., Ashraf, T., Thawakar, O., Anwer, R. M., Cholakkal, H., Shah, M., ... & Khan, S. (2025). Llm post-training: A deep dive into reasoning large language models. arXiv preprint arXiv:2502.21321.
- Wachi, A., Shen, X., & Sui, Y. (2024). A survey of constraint formulations in safe reinforcement learning. arXiv preprint arXiv:2402.02025.
- Mao, Y., Ma, X., & Li, J. (2025). Research on API Security Gateway and Data Access Control Model for Multi-Tenant Full-Stack Systems.
- 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).
- Bedi, A., Crestani, L., El-Rayes, Y., & Marcelli, D. A Compressed Suffix Memory Approach to Improving RL Performance in Partially Observable Domains.
- 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.
- Li, T., Xia, J., Liu, S., & Jiang, Y. (2025). Digital Transformation of Human Resources: From Consulting Frameworks to AI-Enabled Learning Management Systems.
- Garg, D., VanWeelden, S., Caples, D., Draguns, A., Ravi, N., Putta, P., ... & Motwani, S. (2025). Real: Benchmarking autonomous agents on deterministic simulations of real websites. arXiv preprint arXiv:2504.11543.
- Mohammadi, M., Li, Y., Lo, J., & Yip, W. (2025, August). Evaluation and benchmarking of llm agents: A survey. In Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2 (pp. 6129-6139).
- 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.
- Srivastava, S., Li, G., Chang, C., Garg, R., Kaur, M., Lee, C. Y., ... & Qi, P. (2025). WARC-Bench: Web Archive Based Benchmark for GUI Subtask Executions. arXiv preprint arXiv:2510.09872.

- Li, T., Xia, J., Liu, S., & Hong, E. (2025). Strategic Human Resource Leadership in Global Biopharmaceutical Enterprises: Integrating HR Analytics and Cross-Cultural.
- 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.
- Kuntz, T., Duzan, A., Zhao, H., Croce, F., Kolter, Z., Flammarion, N., & Andriushchenko, M. (2025). Os-harm: A benchmark for measuring safety of computer use agents. arXiv preprint arXiv:2506.14866.
- 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).
- Jones, D., Severi, G., Pouliot, M., Lopez, G., de Gruyter, J., Zanella-Beguelin, S., ... & Minnich, A. (2025). A systematization of security vulnerabilities in computer use agents. arXiv preprint arXiv:2507.05445.