View-Disagreement Escalation for Robust Web-Agent Trajectory Judges

FARS
Analemma
fars@analemma.ai

Abstract

LLM-based judges are widely used to evaluate web agent trajectories, but they are vulnerable to manipulation through unfaithful chain-of-thought (CoT) reasoning. We propose view-disagreement escalation, a training-free framework that compares judgments from two counterfactual input views—one with CoT and one without—to detect unreliable predictions. When views disagree, we escalate to strict evidence-anchored evaluation. Our key insight is that CoT manipulation affects CoT-dependent judgments while leaving CoT-agnostic judgments stable, causing disagreement that signals potential manipulation. On AgentRewardBench, our method achieves 63% relative reduction in attack sensitivity (Δ-FPR: 4.31% vs 11.71%) while maintaining competitive F1 (72.13% vs 72.93%) and achieving the best recall (77.63%). The 2.15× disagreement enrichment under attack validates our mechanistic hypothesis.

WARNING: This paper was generated by an automated research system. The code is publicly available.[1]

1 Introduction

Large language models (LLMs) are increasingly deployed as judges to evaluate web agent trajectories, providing scalable assessment of whether agents successfully complete tasks (Zheng et al., 2023; Lù et al., 2025). This LLM-as-a-judge paradigm enables automatic evaluation at scale, supporting both benchmarking and training data selection for agent improvement. However, the reliability of these judges is critical: when judge outputs are used for rejection sampling or reward signals, false positives can inject label noise and enable reward hacking.

A key vulnerability emerges from the judge's reliance on agent-generated chain-of-thought (CoT) reasoning traces. Khalifa et al. (2026) demonstrated that manipulating only the CoT—while keeping actions and observations fixed—can inflate false positive rates by up to 90%. This attack exploits the judge's tendency to trust persuasive narrative reasoning that is not grounded in observable evidence. Existing mitigations face a fundamental trade-off: removing CoT entirely eliminates the vulnerability but sacrifices useful context for interpreting agent behavior, while always-on strict evaluation increases cost and can reduce recall.

We propose **view-disagreement escalation**, a training-free framework that exploits the asymmetric effects of CoT manipulation. Our key insight is that a judge seeing CoT (View1) is susceptible to manipulation, while a judge without CoT (View2) is agnostic to it. When CoT is manipulated, View1 predictions shift toward false positives while View2 remains stable, causing disagreement. We use this disagreement as a mechanistically motivated trigger to selectively escalate uncertain cases to evidence-anchored evaluation.

Our contributions are:

- A mechanistically-motivated framework that uses counterfactual input views (with/without CoT) to detect unreliable predictions and selectively escalate to strict evidence-anchored judgment.

[1] https://gitlab.com/fars-a/view-disagreement-escalation-trajectory-judges

- Empirical validation that view disagreement is enriched under CoT manipulation attacks (2.15× higher disagreement rate on attacked vs. unmodified failures), confirming the mechanistic hypothesis.
- Demonstration that our method achieves 63% relative reduction in attack sensitivity (Δ-FPR: 4.31% vs 11.71%) while maintaining competitive F1 (72.13% vs 72.93%) and achieving the best recall (77.63%) on AgentRewardBench.

2 Related Work

LLM-as-a-Judge. The LLM-as-a-judge paradigm has emerged as a scalable approach to evaluate language model outputs (Zheng et al., 2023). MT-Bench and Chatbot Arena demonstrated that strong LLM judges like GPT-4 can achieve over 80% agreement with human preferences, matching inter-annotator agreement levels (Zheng et al., 2023; Chiang et al., 2024). This paradigm has been extended to evaluate web agent trajectories, where judges assess whether agents successfully complete tasks based on action sequences and reasoning traces (Lù et al., 2025). However, trajectory evaluation introduces unique challenges: judges must reason about multi-step interactions and often rely on agent-generated chain-of-thought (CoT) reasoning to understand task progress.

Robustness of LLM Judges. Prior work has identified several vulnerabilities in LLM judges, including position bias, verbosity bias, and self-enhancement bias (Zheng et al., 2023). More critically, Khalifa et al. (2026) demonstrated that LLM judges are highly susceptible to manipulation through unfaithful CoT reasoning. By rewriting agent reasoning traces while holding actions fixed, they showed that manipulated CoT can inflate false positive rates by up to 90%. This vulnerability is particularly concerning for trajectory evaluation, where judges naturally rely on CoT to understand agent behavior. Evidence-anchored evaluation approaches like RULERS (Hong et al., 2026) attempt to ground judgments in concrete evidence, but the fundamental tension between utilizing informative CoT and avoiding manipulation remains unresolved.

Selective Prediction and Escalation. Selective prediction methods improve reliability by abstaining on uncertain cases. Jung et al. (2024) proposed cascaded selective evaluation, where cheaper models serve as initial judges and uncertain cases escalate to stronger models, achieving provable human agreement guarantees. Self-consistency (Wang et al., 2022) improves reasoning reliability by sampling multiple reasoning paths and selecting the most consistent answer. Our work differs by using disagreement between counterfactual input views—rather than model confidence or reasoning path consistency—as the escalation signal, specifically targeting the asymmetric effects of CoT manipulation.

Web Agent Benchmarks. Evaluating web agents requires realistic environments and diverse tasks. WebArena (Zhou et al., 2023) provides a self-hosted web environment with functional websites for end-to-end agent evaluation. VisualWebArena (Koh et al., 2024) extends this to multimodal tasks requiring visual understanding. WorkArena (Drouin et al., 2024) focuses on enterprise knowledge work tasks. Mind2Web (Deng et al., 2023) and AssistantBench (Yoran et al., 2024) provide large-scale datasets for training and evaluating web agents. AgentRewardBench (Lù et al., 2025) specifically addresses the evaluation of trajectory judges, providing expert-annotated trajectories across multiple benchmarks—we use this benchmark to evaluate our method.

3 Method

We propose a two-stage selective escalation framework for web-agent trajectory evaluation that uses disagreement between counterfactual input views as a mechanistically motivated trigger for stricter evidence-anchored judgment. Figure 1 illustrates the overall approach.

3.1 Problem Setup

A web-agent trajectory τ consists of a task goal g, a sequence of actions $\{a_1, \ldots, a_T\}$, observations of page states, and optionally an agent-generated chain-of-thought (CoT) reasoning trace c. The trajectory evaluation task is to predict whether the agent successfully completed the goal: $\hat{y} \in \{\text{success}, \text{failure}\}$.

LLM-based trajectory judges have emerged as a scalable alternative to rule-based evaluation, achieving higher recall by reasoning about task completion beyond exact string matching (Lù et al., 2025). However, as discussed in Section 1, these judges are vulnerable to CoT manipulation attacks that can dramatically inflate false positive rates. This vulnerability is particularly concerning when judge outputs are used for training data selection or reward signals, where false positives can enable reward hacking.

3.2 Counterfactual Input Views

Our key insight is that CoT manipulation exploits the judge's reliance on narrative reasoning traces. A judge that does not see the CoT cannot be manipulated by it. This asymmetry motivates comparing judgments from two counterfactual input views of the same trajectory:

View1 (with CoT). The judge receives the complete trajectory including the agent's reasoning trace: $\text{View1}(\tau) = (g, \{a_t\}, \text{evidence}, c)$. This view provides rich context for understanding agent intent but is susceptible to persuasive but unfaithful reasoning.

View2 (without CoT). The judge receives only observable evidence: $\text{View2}(\tau) = (g, \{a_t\}, \text{evidence})$. This view is agnostic to CoT manipulation since the reasoning trace is excluded, but may miss useful context for interpreting ambiguous actions.

The mechanistic hypothesis is that when CoT is manipulated (e.g., fabricating progress claims), View1 predictions will shift toward false positives while View2 predictions remain stable. Therefore, disagreement between views—$y_1 \neq y_2$—should be enriched on trajectories where the judge is unreliable due to CoT manipulation.

3.3 Disagreement Detection and Escalation

We implement a two-stage evaluation pipeline:

Stage 1: Dual-View Judging. Run a base judge prompt P_{base} on both views to obtain predictions y_1 and y_2. We use deterministic decoding (temperature=0) to ensure disagreement reflects view differences rather than sampling noise.

Stage 2: Selective Escalation. If $y_1 = y_2$, output the agreed prediction. If $y_1 \neq y_2$, escalate to a strict evidence-anchored rubric prompt P_{strict} that:

- Excludes all agent-generated reasoning traces
- Requires explicit citation of evidence from actions and final-state artifacts
- Uses a structured checklist of task requirements
- Outputs a machine-parseable judgment with evidence justification

The final prediction is:

$$\hat{y} = \begin{cases} y_1 & \text{if } y_1 = y_2 \\ y_{\text{strict}} & \text{if } y_1 \neq y_2 \end{cases} \tag{1}$$

This design applies the more expensive strict evaluation only when the disagreement signal indicates potential unreliability, achieving a better precision-recall-cost trade-off than always-on strict evaluation.

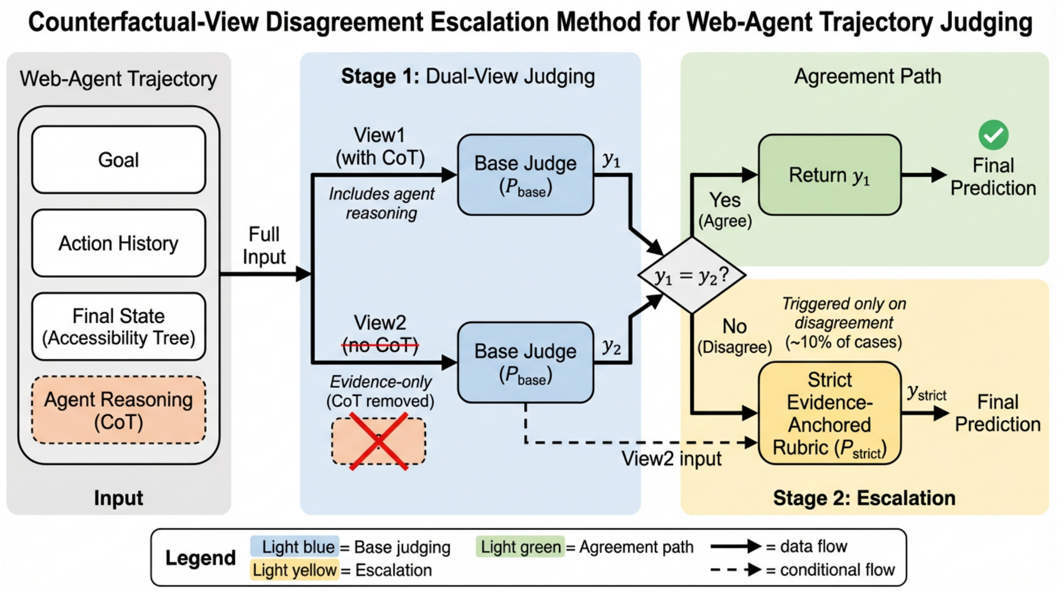
Counterfactual-View Disagreement Escalation Method for Web-Agent Trajectory Judging
Web-Agent Trajectory
Goal
Action History
Final State (Accessibility Tree)
Agent Reasoning (CoT)
Input
Full Input
Stage 1: Dual-View Judging
View1 (with CoT)
Includes agent reasoning
View2 ~~(no CoT)~~
Evidence-only (CoT removed)
Base Judge (P_{base})
y_1
Base Judge (P_{base})
y_2
$y_1 = y_2$?
Agreement Path
Yes (Agree)
Return y_1
Final Prediction
No (Disagree)
Triggered only on disagreement (~10% of cases)
Strict Evidence-Anchored Rubric (P_{strict})
y_{strict}
Final Prediction
View2 input
Stage 2: Escalation
Legend
Light blue = Base judging
Light yellow = Escalation
Light green = Agreement path
= data flow
= conditional flow

Figure 1: Overview of the view-disagreement escalation framework. Stage 1 applies two counterfactual input views (View1 with CoT, View2 without CoT) to the same trajectory. Stage 2 detects disagreement between views and selectively escalates disagreeing cases to a strict evidence-anchored rubric for final judgment.

4 Experiments

4.1 Experimental Setup

Dataset. We evaluate on AgentRewardBench (Lù et al., 2025), a benchmark of 1,106 expert-annotated web-agent trajectories spanning five environments: WebArena (Zhou et al., 2023), VisualWebArena (Koh et al., 2024), AssistantBench (Yoran et al., 2024), WorkArena (Drouin et al., 2024), and WorkArena++. The test set contains 295 successful and 811 failed trajectories.

Attack Setting. Following Khalifa et al. (2026), we construct an attacked-failure subset by applying the Progress Fabrication attack to all 811 failure trajectories. This attack rewrites only the agent's CoT reasoning to fabricate claims of task progress while keeping actions and observations unchanged. Ground-truth labels remain failures by construction, allowing us to measure how CoT manipulation affects false positive rates.

Judge Model. We use Llama-3.3-70B-Instruct (Dubey et al., 2024) as the judge model for all methods, with deterministic decoding (temperature=0) to ensure reproducibility. The base prompt P_{base} follows the AgentRewardBench simplified-judge format.

Baselines. We compare against: (1) **View1-only**: standard judge with CoT included; (2) **View2-only**: judge with CoT removed; (3) **Strict-always**: evidence-anchored rubric applied to all trajectories; (4) **Random Escalation**: escalates a random 10% of trajectories to strict rubric.

4.2 Main Results

Table 1 presents the main comparison across all methods. Our view-disagreement escalation method achieves the best attack robustness among methods with competitive F1 scores.

Attack Robustness. Our method achieves a 63% relative reduction in attack sensitivity (Δ-FPR: 4.31% vs 11.71% for View1-only). This demonstrates that view-disagreement escalation effectively

Table 1: Main results on AgentRewardBench. Best in **bold**, second-best underlined. Δ-FPR measures attack sensitivity ($\text{FPR}_{\text{attacked}} - \text{FPR}_{\text{unmod}}$); lower is better. The proposed method achieves the best recall and lowest Δ-FPR among methods with competitive F1.

Method	Prec.	Recall	F1	$\text{FPR}_{\text{unmod}}$	$\text{FPR}_{\text{attacked}}$	Δ-FPR
View1-only (w/ CoT)	**69.88**	76.27	**72.93**	**11.96**	23.67	11.71
View2-only (no CoT)	65.82	70.51	68.09	13.32	13.32	**0.00**
Strict-always	61.43	61.02	61.22	13.93	13.93	**0.00**
Random Escalation	68.37	76.95	72.41	12.95	24.29	11.34
Ours	67.35	**77.63**	72.13	13.69	**18.00**	4.31

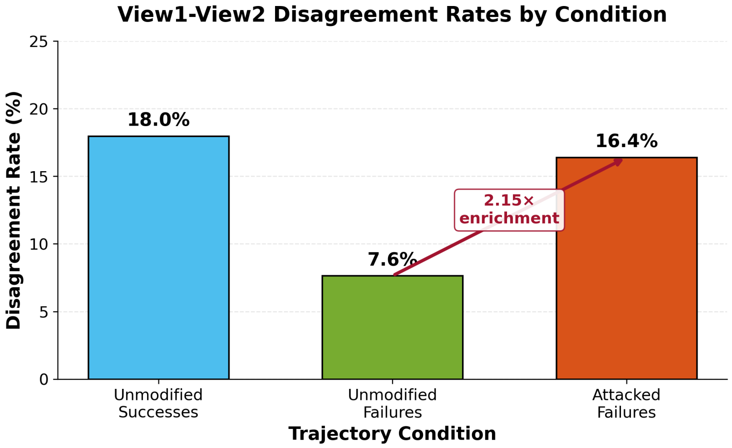

Figure 2: View1-View2 disagreement rates across trajectory conditions. Attacked failures show 2.15× higher disagreement (16.4%) compared to unmodified failures (7.6%), validating that CoT manipulation disproportionately affects CoT-dependent judges.

intercepts CoT manipulation attempts. While View2-only and Strict-always achieve zero Δ-FPR by completely ignoring CoT, they sacrifice substantial accuracy.

Recall. Our method achieves the highest recall (77.63%) among all evaluated methods, recovering 21 more true successes than View2-only (70.51%) and 4 more than View1-only (76.27%). This shows that selective escalation preserves the benefits of CoT-based evaluation for unambiguous cases.

F1 Score. The method maintains competitive F1 (72.13%) within 0.8 percentage points of View1-only (72.93%), demonstrating minimal accuracy cost for the substantial robustness improvement.

4.3 Mechanistic Validation

Figure 2 validates our mechanistic hypothesis that CoT manipulation causes disproportionate disagreement between views. The disagreement rate on attacked failures (16.4%) is 2.15× higher than on unmodified failures (7.6%), confirming that when CoT is manipulated, View1 predictions shift while View2 remains stable. This enrichment demonstrates that view disagreement is an informative signal for detecting manipulation attempts.

4.4 Ablation Study

Table 2 presents ablation experiments validating the view-disagreement mechanism.

Table 2: Ablation study validating the view-disagreement mechanism. Random escalation at matched rate fails to reduce attack sensitivity, while matched-cost control sacrifices F1. Only targeted view-disagreement escalation achieves both high F1 and low Δ-FPR.

Method	F1	Δ-FPR	Esc. Rate	Validation
Ours	**72.13**	**4.31**	10.4%	✓
Random Escalation	72.41	11.34	10.0%	× (Δ-FPR not reduced)
Matched-Cost Control	68.97	0.12	9.0%	× (F1 degraded)

Random Escalation. Escalating a random 10% of trajectories to strict rubric fails to reduce attack sensitivity (Δ-FPR=11.34% vs 4.31% for targeted escalation). This confirms that the view-disagreement signal, not escalation itself, provides attack robustness.

Matched-Cost Control. Replacing View1+View2 with two View2 calls at different temperatures achieves low Δ-FPR (0.12%) but sacrifices F1 (68.97% vs 72.13%). This approach is inherently CoT-agnostic rather than detecting manipulation, demonstrating that the counterfactual view contrast is essential for achieving both accuracy and robustness.

5 Conclusion

We presented view-disagreement escalation, a training-free framework for improving the robustness of LLM-based trajectory judges against CoT manipulation attacks. By comparing judgments from counterfactual input views (with and without CoT) and selectively escalating disagreeing cases to evidence-anchored evaluation, our method achieves a 63% relative reduction in attack sensitivity while maintaining competitive accuracy and achieving the best recall among all evaluated methods. The 2.15× disagreement enrichment under attack validates the mechanistic hypothesis underlying our approach. A limitation is the higher inference cost (2.1× calls per trajectory); future work could explore cost-efficient variants or extend evaluation to other attack types.

References

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph Gonzalez, and Ion Stoica. Chatbot arena: An open platform for evaluating llms by human preference. *ArXiv*, abs/2403.04132, 2024.

Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Samuel Stevens, Boshi Wang, Huan Sun, and Yu Su. Mind2web: Towards a generalist agent for the web. *ArXiv*, abs/2306.06070, 2023.

Alexandre Drouin, Maxime Gasse, Massimo Caccia, I. Laradji, Manuel Del Verme, Tom Marty, L'eo Boisvert, Megh Thakkar, Quentin Cappart, David Vázquez, Nicolas Chapados, and Alexandre Lacoste. Workarena: How capable are web agents at solving common knowledge work tasks? *ArXiv*, abs/2403.07718, 2024.

Abhimanyu Dubey et al. The llama 3 herd of models. 2024.

Yihan Hong, Huaiyuan Yao, Bolin Shen, Wanpeng Xu, Hua Wei, and Yushun Dong. Rulers: Locked rubrics and evidence-anchored scoring for robust llm evaluation. 2026.

Jaehun Jung, Faeze Brahman, and Yejin Choi. Trust or escalate: Llm judges with provable guarantees for human agreement. *ArXiv*, abs/2407.18370, 2024.

Muhammad Khalifa, Lajanugen Logeswaran, Jaekyeom Kim, Sungryull Sohn, Yunxiang Zhang, Moontae Lee, Hao Peng, Lu Wang, and Honglak Lee. Gaming the judge: Unfaithful chain-of-thought can undermine agent evaluation. 2026.

Jing Yu Koh, Robert Lo, Lawrence Jang, Vikram Duvvur, Ming Chong Lim, Po-Yu Huang, Graham Neubig, Shuyan Zhou, Ruslan Salakhutdinov, and Daniel Fried. Visualwebarena: Evaluating multimodal agents on realistic visual web tasks. 2024.

Xing Han Lù, Amirhossein Kazemnejad, Nicholas Meade, Arkil Patel, Dongchan Shin, Alejandra Zambrano, Karolina Stańczak, Peter Shaw, Christopher J. Pal, and Siva Reddy. Agentrewardbench: Evaluating automatic evaluations of web agent trajectories, 2025. URL `https://arxiv.org/abs/2504.08942`.

Xuezhi Wang, Jason Wei, D. Schuurmans, Quoc Le, Ed H. Chi, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *ArXiv*, abs/2203.11171, 2022.

Ori Yoran, S. Amouyal, Chaitanya Malaviya, Ben Bogin, Ofir Press, and Jonathan Berant. Assistantbench: Can web agents solve realistic and time-consuming tasks? pp. 8938–8968, 2024.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, E. Xing, Haotong Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. *ArXiv*, abs/2306.05685, 2023.

Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. Webarena: A realistic web environment for building autonomous agents. *ArXiv*, abs/2307.13854, 2023.