Munich Center for Machine Learning

About Me

Project Page

WEBARBITER

A Principle-Guided Reasoning Process Reward Model for Web Agents

Yao Zhang[1,3], Shijie Tang[1], Zeyu Li[2], Zhen Han[1], Volker Tresp[1,3]

[1]LMU Munich [2]Technical University of Munich [3]Munich Center for Machine Learning

Why Process Reward Models for Web Agents?

The Problem with Outcome Reward Models (ORMs)

- Web interactions involve long-horizon, multi-step decisions with irreversible actions.
- ORMs provide only sparse and delayed feedback, may misclassify incorrect trajectories as successes, and cannot guide inference-time strategies such as reward-guided search.

→ This motivates **Process Reward Models (WebPRMs)** for step-level supervision

Scalar WebPRM

Collapses progress into coarse scores with little interpretability or weak grounding

Checklist WebPRM

Relies on checklists that are brittle under dynamic layouts and state-dependent action semantics

LLM-as-Judge

High cost, limited scalability, susceptible to hallucination, often rewards fluent but incorrect actions, cannot guide inference-time search

WebArbiter: Our Approach

Key Insight: Formulate process reward modeling as text generation, producing structured justifications that conclude with a preference verdict identifying the action most conducive to task completion.

1

Principle Induction

Dynamically derives principles from user intent and the current state, integrating them into the reasoning process rather than relying on fixed templates

→

2

Structured Reasoning

Produces auditable reasoning chains grounding each candidate action against the induced principles

→

3

Preference Verdict

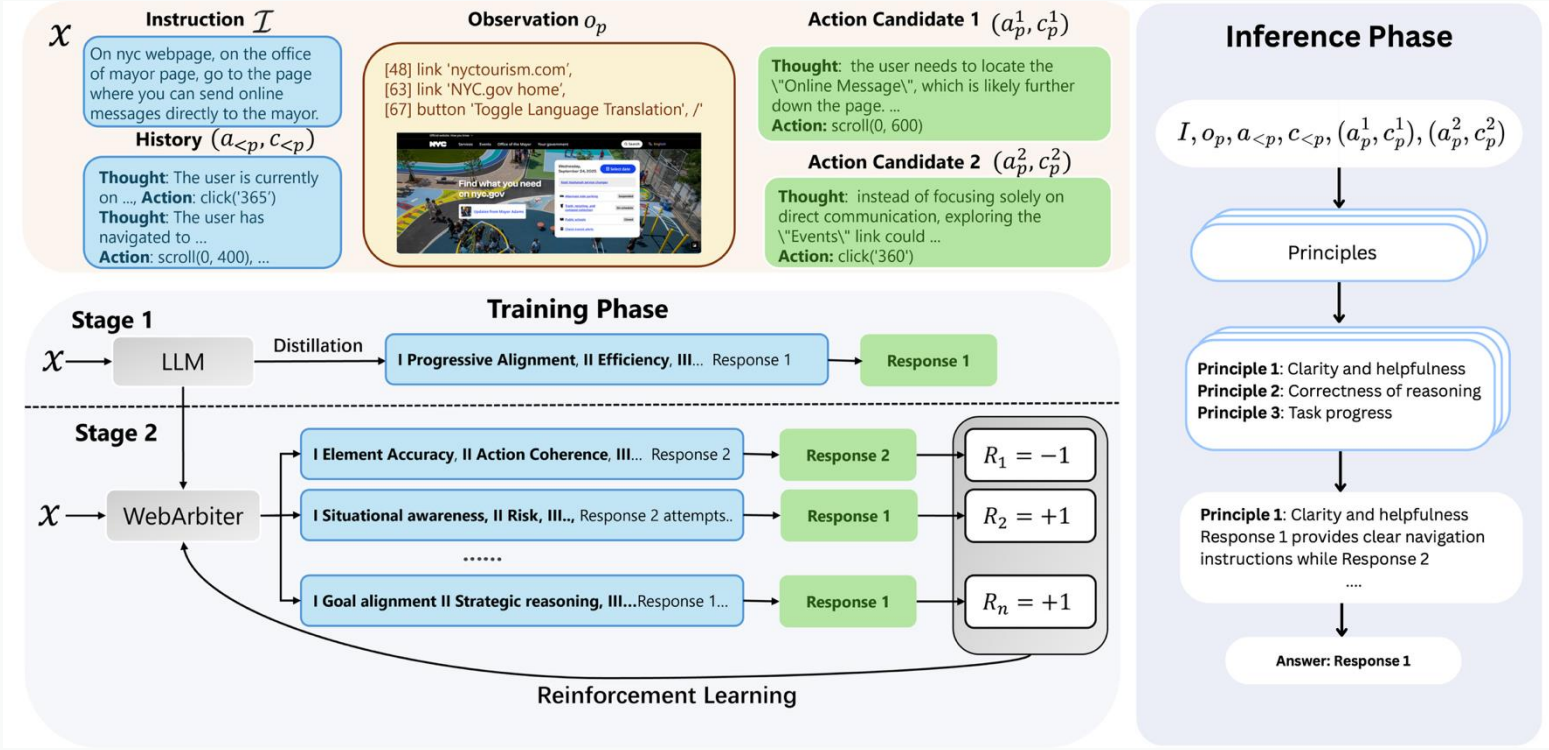
Concludes with a single verdict identifying the action most conducive to task completion under the current context

Training: Two-stage pipeline

(1) Reasoning distillation from a stronger teacher, promoting principle-grounded reasoning;

(2) Reinforcement learning (GRPO) with binary rewards $R \in \{-1, +1\}$ to align verdicts with correctness.

WebArbiter Overview



Stage 1: Reasoning Distillation
Distill principle-guided reasoning from a stronger teacher, promoting judgments grounded in explicit principles rather than surface heuristics

Stage 2: Reinforcement Learning
RL (GRPO) with binary rewards R ∈ {−1, +1} to align verdicts with correctness, enabling stronger generalization

WebPRMBench

First comprehensive evaluation benchmark for WebPRMs

1,150
Step-level preference instances

4
Diverse web environments

2
Metrics: Pairwise & Best-of-N Accuracy

Environment	Mind2Web Cross-Task	Mind2Web Cross-Website	Mind2Web Cross-Domain	WebArena	AssistantBench	WorkArena	Total
Count	142	148	417	201	30	212	1,150

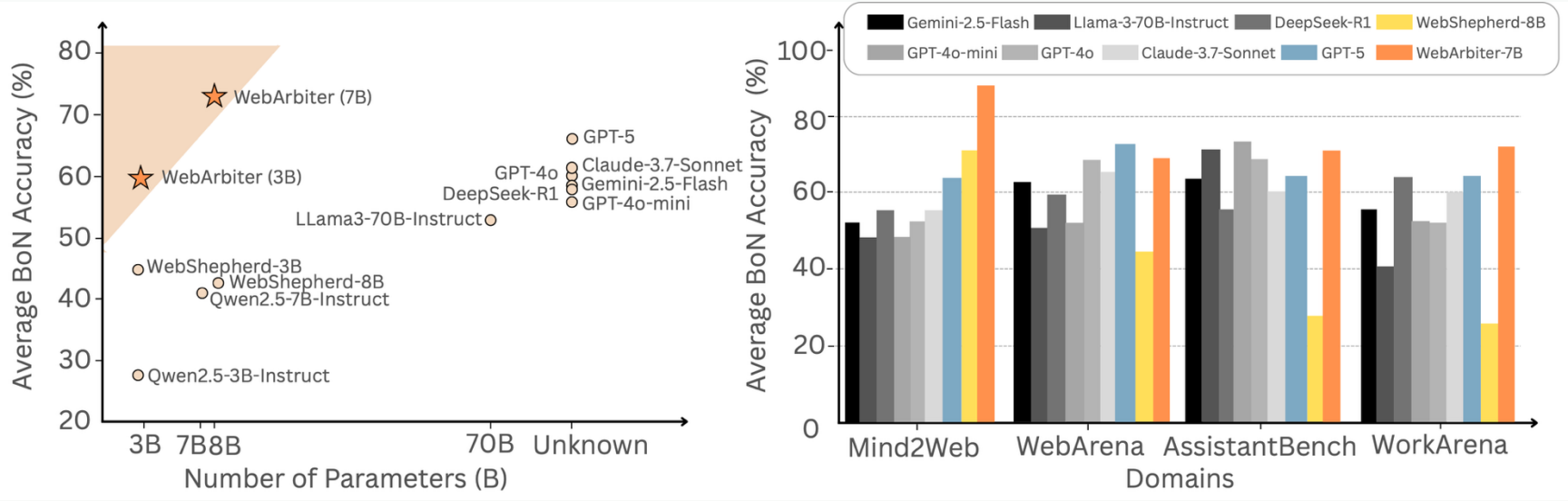
Mind2Web: cross-task generalization across heterogeneous websites

WebArena: controlled environments (Shopping, CMS, Reddit, GitLab)

AssistantBench: open-world tasks on real websites

WorkArena: enterprise workflows including IT and HR management

Main Results — Performance on WebPRMBench



Average BoN Acc versus model size (left) and across domains (right).

Key Findings

- WebArbiter-7B outperforms the strongest proprietary baseline, GPT-5, by 9.1 points in Avg. BoN Acc.
- Surpasses the prior SOTA WebPRM (WebShepherd-8B) by over 31 points in Avg. BoN Acc.
- Demonstrates robust generalization across all four diverse environments.

Main Results — Reward-Guided Trajectory Search

Policy	WebPRM	Shopping	CMS	Reddit	GitLab	MAP	Avg.	Δ
GPT-4o-mini	w/o Trajectory Search*	21.74	22.86	19.05	34.38	19.35	23.48	–
	GPT-4o-mini	24.44	22.86	26.32	33.33	15.38	24.47	+0.99
	WebShepherd-8B*	26.09	**45.71**	23.81	40.62	35.48	34.34	+10.86
	★ WebArbiter-7B	**37.78**	42.86	**36.84**	**46.67**	**38.46**	**40.52**	**+17.04**
GPT-4o	w/o Trajectory Search*	23.91	31.43	28.57	56.25	19.35	31.90	–
	GPT-4o-mini	26.67	37.14	42.11	40.00	19.23	33.03	+1.13
	WebShepherd-8B*	30.43	**42.86**	47.62	46.88	35.48	40.65	+8.75
	★ WebArbiter-7B	**44.44**	**42.86**	**52.63**	**56.67**	**38.46**	**47.01**	**+15.11**

Success Rates (%) of reward-guided trajectory search on WebArena-Lite with GPT-4o-mini / GPT-4o policies.

Key Findings

- WebArbiter-7B achieves the highest Avg. success rate gains: +17.04 (GPT-4o-mini) and +15.11 (GPT-4o) over no-search baselines.
- Surpasses the best prior WebPRM (WebShepherd) by up to 6.4 points in average success rate.
- Gains are largest in open, variable domains (Shopping, Reddit, GitLab).

Analysis of Training Design

Training Recipes

Method	*Pairwise*	*BoN*
Instruct (Original)	77.61	42.78
Instruct + Cold Start RL	78.15	48.33
Instruct + Cold Start RL + Principles	84.12	58.75
Instruct + $\text{SFT}_{\text{w/o Principles}}$ + RL	82.35	55.16
★ WebArbiter-7B	**89.19**	**74.60**

Average Pairwise and BoN Acc under different training recipes.

Key Findings

- RL alone is unstable across web environments.
- Principles enable cross-environment generalization.
- Reasoning without principles is insufficient.

Reasoning Supervision

Method	Pairwise	BoN
Train on Full Data		
Instruct + SFT	82.02	55.80
Instruct + Distilled + SFT	83.37	58.49
★ WebArbiter-7B (Instruct + Distilled + RL)	**89.19**	**74.60**
Train on 10K (Stage-1 Reasoning Distillation) Data		
Instruct + SFT	82.46	53.96
Instruct + Distilled	83.18	60.25

Average Pairwise and BoN Acc under different supervision.

Key Findings

- Reasoning distillation improves judgment stability, with RL as an amplifier.
- Reasoning supervision is especially effective under limited data.

Conclusion & Future Work

Conclusion

- **WebArbiter**: a principle-guided reasoning process reward model that formulates reward modeling as structured text generation with preference verdicts.
- **Two-stage training**: reasoning distillation + reinforcement learning converts surface correlations into robust, progress-aware reward signals.
- **WebPRMBench**: the first comprehensive benchmark for evaluating WebPRMs across 4 diverse web environments with 1,150 instances.
- **SOTA results**: surpasses all proprietary and open-source baselines on both WebPRMBench and reward-guided trajectory search on WebArena-Lite.

Future Work

- **Multimodal observations**: incorporate visual inputs beyond text-only accessibility trees.
- **Step-level agent training**: leverage PRM-based supervision for web agent post-training.