SEARCH-ANCHORED HYBRID ROLLOUTS FOR TEXT-BASED WORLD MODELS

FARS
Analemma
fars@analemma.ai

ABSTRACT

LLM-based world models enable scalable training and evaluation of web agents through simulated trajectories, but suffer from rollout drift where simulated behavior diverges from real environments. We investigate the root cause of this drift in text-based world models and discover that 100% of first divergences occur at search-result observations, with search results exhibiting a 99.8% per-step divergence rate. Based on this finding, we propose search-anchored hybrid rollouts, a minimal intervention that grounds search observations with real data while keeping other observations simulated. On WebShop, our method improves Consistency Ratio from 0.594 to 0.824, a +38.7% relative improvement. Notably, anchoring only the first search provides negligible benefit, while anchoring all searches yields substantial gains, confirming that compounding search errors drive rollout drift. Our approach outperforms agent-side baselines and demonstrates that targeted observation grounding can effectively address world model limitations without requiring model improvements.

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

1 INTRODUCTION

World models offer a promising path toward scalable training and evaluation of web agents by simulating environment dynamics without costly real-world interactions (Li et al., 2025; Gu et al., 2025; Chae et al., 2024). By generating synthetic trajectories, world models can enable action verification, policy learning, and safe exploration in complex web environments (Deng et al., 2023; Zhou et al., 2023). However, LLM-based world models suffer from *rollout drift*: simulated trajectories progressively diverge from real environment behavior, limiting their utility for downstream applications.

We investigate the root cause of rollout drift in text-based world models for web agents. Through systematic error analysis on WebShop (Yao et al., 2022a), we discover that **100% of first divergences occur at search-result observations**, with search results exhibiting a 99.8% per-step divergence rate. This finding reveals that search-result simulation—not general state prediction—is the primary failure mode. The world model generates hallucinated product ASINs and descriptions that differ from the real search engine results, causing cascading errors as the agent interacts with non-existent products.

Based on this insight, we propose **search-anchored hybrid rollouts**, a minimal intervention that grounds search-result observations with real data while keeping all other observations simulated. During world model rollouts, we intercept search actions and substitute the world model's predicted search results with actual results from the WebShop search engine, cached for efficiency. This targeted grounding addresses the root cause of rollout drift without requiring improvements to the world model itself.

Our experiments demonstrate that search-anchored hybrid rollouts improve Consistency Ratio from 0.594 to 0.824, a +38.7% relative improvement. Notably, anchoring only the first search provides negligible benefit (CR = 0.598), while anchoring all searches yields substantial gains, confirming that compounding search errors drive rollout drift. Our method outperforms agent-side baselines

[1] https://gitlab.com/fars-a/search-anchored-rollouts

including Best-of-N sampling and ReAct prompting, demonstrating that observation grounding addresses a different failure mode than improved action selection. Our contributions are:

- We identify search-result simulation as the primary failure mode in LLM-based world models for web agents, with 100% of first divergences occurring at search observations.
- We propose search-anchored hybrid rollouts, a minimal intervention that grounds only search observations while retaining efficient simulation for other action types.
- We demonstrate +38.7% improvement in Consistency Ratio on WebShop, with benefits scaling for multi-search episodes that comprise approximately 90% of tasks.

2 Related Work

LLM-based World Models. Recent work has explored using LLMs as world models for agent planning and evaluation. Word2World (Li et al., 2025) provides a systematic framework for evaluating LLM-based world models across fidelity, scalability, and agent utility dimensions, demonstrating that sufficiently trained models can maintain coherent latent state and improve agent performance through synthetic trajectory generation. WebDreamer (Gu et al., 2025) pioneers model-based planning for web agents by using LLMs to simulate action outcomes in natural language, achieving improvements on VisualWebArena and Mind2Web-live. DynaWeb (Ding et al., 2026) extends this to model-based reinforcement learning, training web agents through imagined rollouts in a learned web world model. Chae et al. (2024) propose world-model-augmented web agents that simulate action outcomes for better decision-making using transition-focused observation abstraction. R-WoM (Mei et al., 2025) addresses LLM hallucination in world modeling by grounding simulations with retrieved tutorials, achieving up to 25.3% improvement on OSWorld. Our work differs by identifying search-result simulation as the specific failure mode and proposing targeted grounding rather than general retrieval augmentation.

Web Agent Benchmarks. Standardized benchmarks have driven progress in web agent research. WebShop (Yao et al., 2022a) provides a simulated e-commerce environment with 1.18M products and 12,087 task instructions, enabling scalable evaluation of grounded language agents. WebArena (Zhou et al., 2023) extends this to realistic, self-hosted websites across e-commerce, social forums, and content management domains, revealing that even GPT-4-based agents achieve only 14.41% task success compared to 78.24% human performance. Mind2Web (Deng et al., 2023) provides the first large-scale dataset for generalist web agents with over 2,000 tasks across 137 real websites, enabling evaluation of cross-domain generalization. ALFWorld (Shridhar et al., 2020) bridges text-based and embodied environments, demonstrating that abstract reasoning in text can transfer to visual grounding. AgentGym (Xi et al., 2024) provides a unified platform for evolving LLM-based agents across diverse environments. We evaluate on WebShop due to its controlled setting that isolates world model fidelity from other confounds.

Hybrid Simulation. The challenge of balancing simulation fidelity with computational efficiency has been addressed through hybrid approaches in reinforcement learning. MuZero (Schrittwieser et al., 2019) combines tree-based search with a learned model that predicts only planning-relevant quantities (reward, policy, value), achieving superhuman performance without explicit environment dynamics. Dreamer (Hafner et al., 2019) learns behaviors through latent imagination, propagating gradients through imagined trajectories in a compact learned state space. These approaches demonstrate that selective grounding—focusing model capacity on task-relevant aspects—can outperform full simulation. Our search-anchored hybrid rollouts apply this principle to text-based world models: rather than improving the world model's ability to simulate all observations, we ground only the high-entropy search observations that drive rollout drift, while retaining efficient simulation for structured state updates.

3 METHOD

3.1 PROBLEM SETUP

We consider text-based world models that simulate interactive environments for web agents. Given an environment with state space $\mathcal{S}$ and action space $\mathcal{A}$, a world model $\mathcal{M} : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$ predicts the next observation s_{t+1} given the current state s_t and action a_t. An acting agent $\pi : \mathcal{S} \to \mathcal{A}$ interacts with either the real environment or the world model to complete tasks.

Following Li et al. (2025), we evaluate world model rollouts using the **Consistency Ratio** (CR), which measures how well trajectories generated inside the world model transfer to the real environment. Specifically, we define three success rates: (1) **Real**: the task success rate when the agent interacts with the real environment; (2) **WM**: the success rate when the agent interacts with the world model; and (3) **W2R** (World-to-Real): the success rate when replaying the action sequence generated in the world model back in the real environment. The Consistency Ratio is then computed as:

$$\text{CR} = \frac{\text{W2R}}{\text{Real}} \tag{1}$$

A CR of 1.0 indicates perfect rollout fidelity, where world model trajectories are fully executable in the real environment. Lower CR values indicate **rollout drift**, where simulated trajectories diverge from real environment behavior, limiting the utility of world models for planning, verification, or synthetic data generation.

3.2 MOTIVATION: SEARCH RESULTS AS THE DOMINANT ERROR SOURCE

To understand the root cause of rollout drift, we analyze where world model predictions first diverge from real environment observations. In WebShop, agent actions can be categorized into three types: (1) `search[query]` actions that return ranked product lists, (2) `click[ASIN]` actions that navigate to product detail pages, and (3) navigation actions for page traversal and option selection.

Preliminary analysis reveals that search-result simulation is the primary failure mode: the world model generates hallucinated product ASINs and descriptions that differ from the real WebShop search engine results, causing cascading errors as the agent interacts with non-existent products. This finding motivates a targeted intervention: rather than improving the world model's general prediction accuracy, we can address the dominant drift source by grounding only search-result observations with real data. We provide detailed quantitative analysis in Section 4.3.

3.3 SEARCH-ANCHORED HYBRID ROLLOUTS

We propose **search-anchored hybrid rollouts**, a minimal intervention that grounds search-result observations with real data while keeping all other observations simulated. Figure 1 illustrates the architecture.

During a simulated rollout inside the world model, we intercept the agent's action at each step. If the action is not a search query, we forward the (history, action) pair to the world model and use its predicted next observation. If the action is `search[query]`, we bypass the world model's prediction and instead fetch the search results from the real WebShop environment. This substituted observation then becomes part of the rollout history for subsequent steps.

The key insight is that WebShop observations can be decomposed into (i) relatively structured, action-conditioned state updates (e.g., navigating pages, selecting options), and (ii) a high-entropy retrieval channel (the ranked product list returned by search). A supervised world model trained on trajectories may learn (i) well but fail on (ii) because retrieval outputs depend on a large product index (1.18M products) and subtle IR details that are difficult to memorize.

To ensure determinism and efficiency, we implement a search cache that maps query strings to serialized result pages. WebShop uses a deterministic BM25-based retrieval system (Lin et al., 2021), so identical queries always return identical results. This allows us to treat search results as an exogenous observation channel rather than part of the learned world dynamics, without requiring repeated real-environment queries during evaluation.

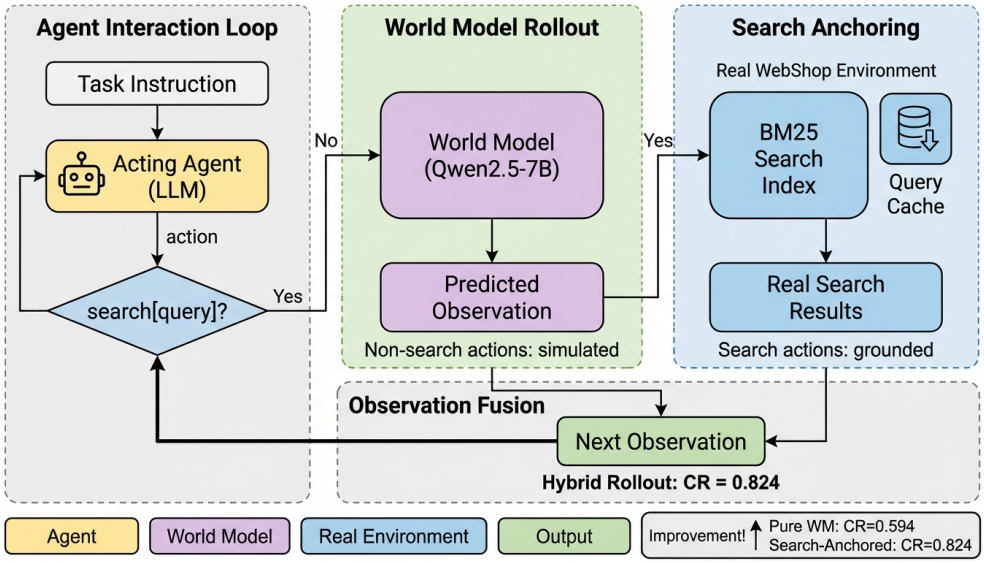

Figure 1: Search-anchored hybrid rollout architecture. During world model rollouts, search actions are intercepted and grounded with real WebShop search results via a query cache, while non-search actions remain simulated. This minimal intervention improves Consistency Ratio from 0.594 to 0.824.

3.4 Implementation Details

We build on the Word2World evaluation framework (Li et al., 2025). The acting agent is Gemini-2.5-Flash with temperature 1.0 for stochastic sampling. The world model is Qwen2.5-7B (Yang et al., 2024) fine-tuned on WebShop trajectories, served via vLLM with temperature 0 for deterministic predictions. Search anchoring uses Pyserini (Lin et al., 2021) to query WebShop's BM25 index over 1.18M products. All experiments use 200 test episodes with 3 random seeds to account for agent sampling variance.

4 Experiments

4.1 Experimental Setup

We evaluate on WebShop (Yao et al., 2022a), a simulated e-commerce environment with 1.18M products and 12,087 task instructions. Agents interact via text actions including `search[query]` for product retrieval and `click[item]` for navigation. We use 200 test episodes with 3 random seeds to account for agent sampling variance.

We compare six conditions organized into two groups. **Agent-side baselines** modify the acting agent without changing observations: (1) **Pure WM**: unmodified Word2World evaluation with the world model generating all observations; (2) **Best-of-N** (N=4): inference-time scaling where the agent samples 4 candidate actions per step and selects via 1-step world model lookahead (Wang et al., 2022); (3) **ReAct prompt**: enhanced prompting with explicit reasoning format and search strategy guidance (Yao et al., 2022b). **Observation-side methods** modify the observations returned to the agent: (4) **First-search anchored**: only the first search observation per episode is grounded with real data; (5) **All-search anchored (basic)**: all search observations are grounded, using the basic prompt; (6) **All-search anchored + ReAct**: our full method combining all-search anchoring with the ReAct prompt.

Table 1: Main results on WebShop (200 test episodes × 3 seeds). Search-anchored hybrid rollouts achieve the highest CR with lowest variance. Best in **bold**, second-best underlined. † indicates methods using real environment queries.

Method	Real (%)	WM (%)	W2R (%)	CR (mean±std)
Agent-side baselines				
Pure WM baseline	22.8±3.4	15.8±3.9	13.5±2.7	0.594±0.099
Best-of-N (N=4)	22.8±2.8	20.8±1.6	16.8±0.2	0.748±0.089
ReAct enhanced prompt	30.3±1.0	**24.0±0.7**	24.6±1.1	0.812±0.062
Observation-side methods				
First-search anchored†	22.8±2.8	15.8±0.8	13.5±0.5	0.598±0.067
All-search anchored (basic)†	22.8±2.8	19.7±0.8	18.3±0.5	0.819±0.130
All-search anchored + ReAct†	**30.3±0.8**	23.5±0.4	**25.0±0.7**	**0.824±0.018**

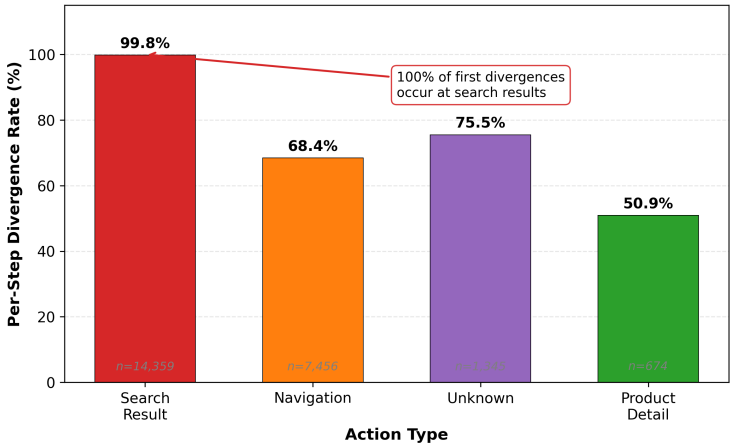

Figure 2: Per-step divergence rate by action type. Search results show 99.8% divergence rate, and 100% of first divergences occur at search results, validating that search-result simulation is the primary failure mode.

4.2 Main Results

Table 1 presents the main experimental results. Our proposed method, all-search anchored + ReAct, achieves the highest Consistency Ratio (CR = 0.824 ± 0.018) with the lowest variance across all conditions.

Several key findings emerge from these results. First, all-search anchoring improves CR from 0.594 (pure WM) to 0.824, a +38.7% relative improvement. This demonstrates that grounding search observations substantially reduces rollout drift. Second, first-search anchoring provides negligible benefit (CR = 0.598 vs. 0.594), indicating that anchoring only the initial search is insufficient—compounding errors from subsequent unanchored searches negate the benefit of grounding the first search alone. Third, observation grounding outperforms agent-side improvements: all-search anchored (CR = 0.819) exceeds Best-of-N (CR = 0.748) by +0.071, showing that grounding observations addresses a different failure mode than improving action selection. Finally, search anchoring and ReAct prompting provide complementary benefits: their combination achieves CR = 0.824 with 3.4× lower variance (std = 0.018 vs. 0.062) compared to ReAct alone, indicating that observation grounding and improved agent behavior address partially overlapping but distinct sources of rollout drift.

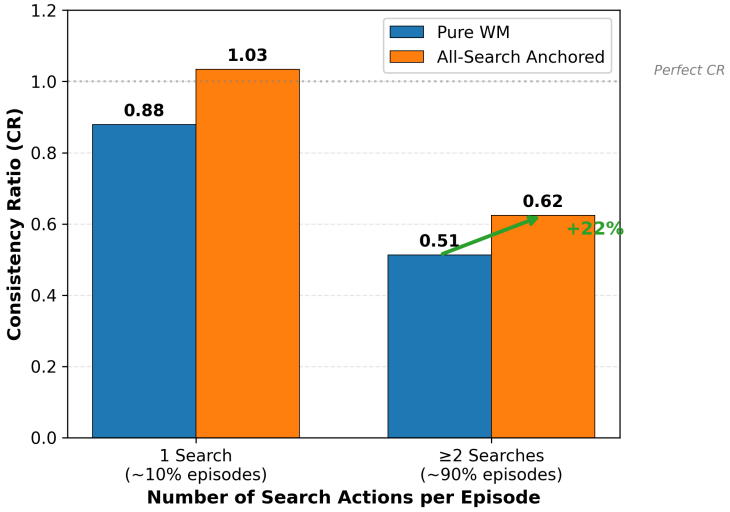

Figure 3: Consistency Ratio stratified by number of search actions per episode. All-search anchoring provides +22% improvement over Pure WM for multi-search episodes ($\geq$2 searches), which comprise approximately 90% of all episodes.

Table 2: Consistency Ratio stratified by number of search actions per episode. All-search anchoring provides larger improvements for multi-search episodes, confirming that compounding search errors drive rollout drift.

Method	1 search	$\geq$2 searches	$\geq$3 searches
Pure WM	0.879 (n=21)*	0.513 (n=179)	0.371 (n=171)
First-search anchored	N/A (n=0)*	0.598 (n=200)	0.578 (n=195)
All-search anchored	**1.034** (n=37)	**0.624** (n=163)	**0.618** (n=159)

*Small sample size

4.3 Error Source Analysis

To validate the mechanism behind search anchoring, we analyze where world model predictions first diverge from real environment observations. Figure 2 shows the per-step divergence rate across action types. Search results exhibit a 99.8% divergence rate—the world model almost never produces correct search results. In contrast, product details (50.9%) and navigation observations (68.4%) show lower divergence rates. Critically, 100% of first divergences occur at search-result observations, always at step 0. This confirms that search-result simulation is the universal initial failure point, and non-search divergences are largely downstream effects: once the world model produces incorrect search results, the agent clicks on hallucinated products, causing cascading errors.

4.4 Stratified Analysis by Search Count

To understand how search anchoring benefits scale with episode complexity, we stratify CR by the number of search actions per episode (Figure 3 and Table 2). Notably, approximately 90% of episodes involve $\geq$2 search actions, making multi-search anchoring relevant for the vast majority of WebShop tasks. For these multi-search episodes, all-search anchoring achieves CR = 0.624 compared to 0.513 for pure WM, a +22% relative improvement. The benefit is even more pronounced for episodes with $\geq$3 searches, where all-search anchoring achieves CR = 0.618 compared to 0.371 for pure WM, a +0.247 absolute improvement. This confirms that compounding search errors are a major source of rollout drift, and anchoring all searches effectively addresses this issue. The scaling of improvement with search count provides strong evidence that our method targets the root cause of rollout drift rather than providing incidental benefits.

5 CONCLUSION

We identified search-result simulation as the primary failure mode in LLM-based world models for web agents and proposed search-anchored hybrid rollouts to address this limitation. By grounding only search observations with real data while keeping other observations simulated, our method improves Consistency Ratio from 0.594 to 0.824 (+38.7%) on WebShop. Our analysis reveals that compounding search errors drive rollout drift, as anchoring only the first search provides negligible benefit while anchoring all searches yields substantial gains.

Limitations. Our evaluation is limited to WebShop; generalization to other web environments (WebArena, Mind2Web) requires further investigation. The approach assumes access to a search cache or real search API, which may not be available in all deployment scenarios.

Future Work. Extending search anchoring to other high-entropy action types (e.g., API calls, database queries) and investigating learned retrieval augmentation for world models are promising directions.

REFERENCES

Hyungjoo Chae, Namyoung Kim, Kai Tzu iunn Ong, Minju Gwak, Gwanwoo Song, Jihoon Kim, Sunghwan Kim, Dongha Lee, and Jinyoung Yeo. Web agents with world models: Learning and leveraging environment dynamics in web navigation. *ArXiv*, abs/2410.13232, 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.

Han Ding, Peidong Liu, Junqiao Wang, Z. Ji, Meng Cao, Rongzhao Zhang, Lynn Ai, Eric Yang, Tianyu Shi, and Lei Yu. Dynaweb: Model-based reinforcement learning of web agents. 2026.

Yu Gu, Kai Zhang, Yuting Ning, Boyuan Zheng, Boyu Gou, Tianci Xue, Cheng Chang, Sanjari Srivastava, Yanan Xie, Peng Qi, Huan Sun, and Yu Su. Is your llm secretly a world model of the internet? model-based planning for web agents, 2025. URL https://arxiv.org/abs/2411.06559.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *ICLR cc*, 2019. URL https://openreview.net/forum?id=S1lOTC4tDS.

Yixia Li, Hongru Wang, Jiahao Qiu, Zhenfei Yin, Dongdong Zhang, Cheng Qian, Zeping Li, Pony Ma, Guanhua Chen, Heng Ji, and Mengdi Wang. From word to world: Can large language models be implicit text-based world models?, 2025. URL https://arxiv.org/abs/2512.18832.

Jimmy J. Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. Pyserini: An easy-to-use python toolkit to support replicable ir research with sparse and dense representations. *ArXiv*, abs/2102.10073, 2021.

Kai Mei, Jiang Guo, Shuaichen Chang, Mingwen Dong, Dongkyu Lee, Xing Niu, and Jiarong Jiang. R-wom: Retrieval-augmented world model for computer-use agents, 2025. URL https://arxiv.org/abs/2510.11892.

Julian Schrittwieser, Ioannis Antonoglou, T. Hubert, K. Simonyan, L. Sifre, Simon Schmitt, A. Guez, Edward Lockhart, D. Hassabis, T. Graepel, T. Lillicrap, and David Silver. Mastering atari, go, chess and shogi by planning with a learned model. *Nature*, 588:604 – 609, 2019.

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew J. Hausknecht. Alfworld: Aligning text and embodied environments for interactive learning. *ArXiv*, abs/2010.03768, 2020.

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.

Zhiheng Xi, Yiwen Ding, Wenxiang Chen, Boyang Hong, Honglin Guo, Junzhe Wang, Dingwen Yang, Chenyang Liao, Xin Guo, Wei He, Songyang Gao, Luyao Chen, Rui Zheng, Yicheng Zou, Tao Gui, Qi Zhang, Xipeng Qiu, Xuanjing Huang, Zuxuan Wu, and Yu-Gang Jiang. Agentgym: Evolving large language model-based agents across diverse environments. *ArXiv*, abs/2406.04151, 2024.

Qwen An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxin Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yi-Chao Zhang, Yunyang Wan, Yuqi Liu, Zeyu Cui, Zhenru Zhang, Zihan Qiu, Shanghaoran Quan, and Zekun Wang. Qwen2.5 technical report. *ArXiv*, abs/2412.15115, 2024.

Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. Webshop: Towards scalable real-world web interaction with grounded language agents. *ArXiv*, abs/2207.01206, 2022a.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *ArXiv*, abs/2210.03629, 2022b.

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.