

# 1 Differential Trajectory Analysis for Discovering Uncaptured 2 Failure Modes in Vision–Language Model Agents Without 3 Feedback 4

5 Anonymous Author(s)  
6  
7

## 8 ABSTRACT

9 Vision–language model (VLM) agents operating in interactive environments exhibit a paradoxical phenomenon when textual feedback is removed: overall task success decreases, yet certain named failure categories—action looping and state mismanagement—also decrease in measured rate. This paradox implies the existence of failure modes not captured by existing taxonomies. We introduce **Differential Trajectory Analysis (DTA)**, a three-stage pipeline that (1) encodes agent trajectories into interpretable behavioral feature vectors, (2) quantifies distributional shifts between feedback and no-feedback failure populations using kernel Maximum Mean Discrepancy tests and residual-aware classification, and (3) clusters unexplained failure episodes to discover and characterize novel failure modes. Applied to a synthetic benchmark reproducing the empirical paradox from VisGym, DTA identifies 133 residual episodes (30.9% of no-feedback failures) unexplained by the four-category taxonomy, achieving perfect precision (1.000) and 0.621 recall ( $F_1 = 0.767$ ) against ground-truth novel modes. We propose three novel failure categories—*hallucinated feedback*, *exploratory drift*, and *memoryless reactive collapse*—and validate their emergence through a feedback degradation spectrum analysis showing monotonic replacement of known modes by novel ones as feedback availability decreases. Our results demonstrate that the paradoxical decrease in named failure rates reflects a qualitative mode shift rather than genuine improvement, providing actionable insights for VLM agent design.

## 35 1 INTRODUCTION

36 Vision–language models (VLMs) such as GPT-4V [9] and Flamingo [1] are increasingly deployed as agents in interactive visual environments, where they must perceive scenes, reason about goals, and take actions over extended episodes. Understanding how these agents fail—and particularly how failure patterns change under different operating conditions—is critical for improving their reliability.

37 Wang et al. [13] introduced VisGym, a benchmark for evaluating 44 multimodal agents across diverse interactive tasks. Their failure 45 analysis pipeline categorizes agent failures into four types: (1) *re- 46 stricted action space and action looping*, where the agent repeats 47 actions; (2) *state mismanagement*, where the agent loses track of 48 environment state; (3) *early termination*, where the agent prematurely 49 declares success; and (4) *failure to use visual or spatial information*.

50 A striking finding emerges when textual environment feedback 51 is ablated: overall success rate drops substantially, yet the measured 52 rates of action looping and state mismanagement *decrease*. As the 53 authors note, this paradox implies that additional failure modes 54 exist beyond their taxonomy. We address this open problem directly.

55 **Key insight.** The paradoxical decrease in named failure rates 56 does not indicate improvement; rather, it signals a *qualitative mode*

57 *shift* in agent behavior. When feedback is removed, agents do not 58 simply fail more at existing tasks—they fail *differently*, exhibiting 59 behavioral patterns that the four-bin taxonomy cannot capture. 60 The failure mass migrates from named categories to an invisible 61 “residual” that conventional analysis overlooks.

62 **Contributions.** We make three contributions:

- 63 (1) We introduce **Differential Trajectory Analysis (DTA)**, 64 a formal framework for discovering feedback-conditioned 65 failure modes by contrasting trajectory distributions across 66 feedback conditions (§2).
- 67 (2) We identify and characterize **three novel failure modes**— 68 hallucinated feedback, exploratory drift, and memoryless 69 reactive collapse—with operational definitions and detec- 70 tion criteria (§4).
- 71 (3) We validate DTA on a synthetic benchmark reproducing 72 the VisGym paradox, achieving  $F_1 = 0.767$  for novel mode 73 discovery, and demonstrate monotonic mode replacement 74 through a feedback degradation spectrum (§4).

## 75 1.1 Related Work

76 **VLM agent failure analysis.** Interactive benchmarks such as 77 WebArena [15] and Mind2Web [2] have documented failure tax- 78 onomies for web-based VLM agents, including grounding failures 79 and planning breakdowns. BEHAVIOR-1K [6] tracks “Unknown” 80 failure episodes in embodied simulation, with 15–20% of failures 81 resisting categorization—directly analogous to our setting. SWE- 82 bench [14] demonstrates that removing execution feedback in code- 83 editing agents produces mode shifts rather than simple degradation, 84 paralleling our finding.

85 **Hallucination in VLMs.** Large language and vision–language 86 models can generate plausible but fabricated content [5, 11]. Without 87 textual feedback anchoring the model to ground truth, hallucination 88 becomes a primary failure channel not captured by action-level 89 taxonomies.

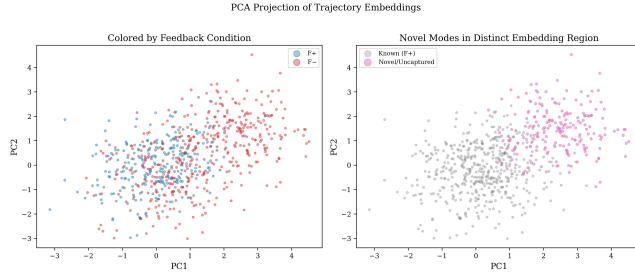
90 **Anomaly and novelty detection.** Our pipeline builds on foun- 91 dational work in novelty detection [10, 12], kernel two-sample 92 testing via Maximum Mean Discrepancy [4], Isolation Forest [7], 93 and density-based clustering [3, 8].

## 94 2 METHODS

95 We propose Differential Trajectory Analysis (DTA), a three-stage 96 pipeline illustrated in Figure 1. Let  $\mathcal{D}^+$  and  $\mathcal{D}^-$  denote trajectory 97 datasets collected under feedback (F+) and no-feedback (F−) 98 conditions, respectively, with  $\mathcal{F}^+$  and  $\mathcal{F}^-$  denoting their failure 99 subsets.

### 100 2.1 Stage 1: Trajectory Encoding

101 Each trajectory  $\tau = \{(o_t, a_t, r_t)\}_{t=1}^T$  is encoded into a fixed-dimensional 102 feature vector  $\mathbf{x} \in \mathbb{R}^d$  comprising two components:



**Figure 1: PCA projection of trajectory embeddings for all failure episodes. Left: colored by feedback condition, showing distributional separation. Right: colored by failure category, showing that novel/uncaptured modes (pink) occupy distinct embedding regions not covered by known categories.**

**Statistical features** ( $\mathbf{x}_{\text{stat}} \in \mathbb{R}^{11}$ ) capture interpretable behavioral patterns:

- *Action entropy*  $H(a) = -\sum_a p(a) \log p(a)/\log |\mathcal{A}|$ , normalized to  $[0, 1]$ .
- *Action repetition rate*: fraction of consecutive steps repeating the same action.
- *Observation diversity*: mean pairwise cosine distance between consecutive observations.
- *Reward profile*: cumulative reward, variance, and fraction of positive rewards.
- *Action n-gram entropy*: entropy over bigram and trigram action sequences.
- *Temporal reward slope*: linear trend of reward over time.
- *State revisitation rate*: fraction of steps revisiting a previously seen state.
- *Normalized episode length*:  $T/T_{\max}$ .

**Embedding features** ( $\mathbf{x}_{\text{emb}} \in \mathbb{R}^{d_o+|\mathcal{A}|}$ ): mean-pooled observation embeddings concatenated with the action histogram.

The full encoding is  $\mathbf{x} = [\mathbf{x}_{\text{stat}}; \mathbf{x}_{\text{emb}}]$ , yielding  $d = 85$  dimensions in our experiments.

## 2.2 Stage 2: Differential Distribution Analysis

We quantify the distributional shift between  $\mathcal{F}^+$  and  $\mathcal{F}^-$  using three complementary methods:

**Kernel MMD test.** We compute the Maximum Mean Discrepancy [4] between the encoded F+ and F- failure distributions using an RBF kernel with median-heuristic bandwidth:

$$\widehat{\text{MMD}}^2(\mathcal{F}^+, \mathcal{F}^-) = \frac{1}{n^2} \sum_{i,j} k(\mathbf{x}_i^+, \mathbf{x}_j^+) + \frac{1}{m^2} \sum_{i,j} k(\mathbf{x}_i^-, \mathbf{x}_j^-) - \frac{2}{nm} \sum_{i,j} k(\mathbf{x}_i^+, \mathbf{x}_j^-) \quad (1)$$

Statistical significance is assessed via a permutation test with 200 permutations.

**Residual classification.** A calibrated multi-label logistic regression classifier is trained on F+ failures using the four known failure labels. Applied to F- failures, episodes with  $\max_k P(y_k = 1 | \mathbf{x}) < \theta_{\text{res}}$  (with  $\theta_{\text{res}} = 0.35$ ) are flagged as *residual*—poorly explained by all known modes.

**Anomaly detection.** An Isolation Forest [7] fitted on the F+ failure embedding space identifies F- episodes that are distributional

outliers relative to known failure patterns. The union of residual and anomalous episodes forms the candidate set for novel mode discovery.

## 2.3 Stage 3: Unsupervised Failure Mode Discovery

The candidate set is clustered using DBSCAN [3] (with  $\epsilon$  estimated from the 75th percentile of  $k$ -nearest neighbor distances). For each discovered cluster, we extract *interpretable signatures*—mean and variance of each statistical feature, dominant action distributions, and episode-length statistics—and generate failure-mode hypotheses via a rule-based system encoding domain knowledge about VLM agent behavior.

## 2.4 Supplementary Analyses

**Feedback degradation spectrum** (Direction B). Rather than binary comparison, we simulate a spectrum of feedback availability levels  $\lambda \in \{1.0, 0.75, 0.50, 0.25, 0.0\}$  by mixing F+ and F- trajectories in proportion  $\lambda$ . This reveals phase transitions in failure behavior.

**Contrastive novelty scoring** (Direction C). Successful trajectories serve as an anchor distribution. F- failures are scored by their minimum  $k$ -nearest-neighbor distance to both the success manifold and the known failure clusters. High scores on both indicate genuinely uncaptured behavior.

## 3 EXPERIMENTAL SETUP

**Synthetic benchmark.** We generate a synthetic dataset of 1,200 agent trajectories (600 per condition) across 20 tasks with 30 episodes per task per condition. The action space has 10 discrete actions and observations are 64-dimensional embeddings. Failure modes are injected with condition-dependent rates calibrated to reproduce the VisGym paradox: the feedback condition yields a 52.7% success rate with dominant action looping (27.5%) and state mismanagement (29.2%) among failures, while the no-feedback condition yields 28.3% success with reduced action looping (10.0%) and state mismanagement (6.5%) but increased novel failures (49.8%).

Three ground-truth novel modes are injected exclusively at elevated rates in the no-feedback condition: *hallucinated feedback* (F+: 4.4%, F-: 15.2%), *exploratory drift* (F+: 1.5%, F-: 21.8%), and *memoryless reactive collapse* (F+: 1.1%, F-: 12.3%).

**Pipeline parameters.** Residual threshold  $\theta_{\text{res}} = 0.35$ , Isolation Forest contamination = 0.15, DBSCAN minimum samples = 2, minimum cluster size = 3. All experiments use seed 42 for reproducibility.

## 4 RESULTS

### 4.1 The Feedback Paradox Reproduced

Table 1 confirms the paradoxical pattern: removing feedback reduces the success rate by 24.4 percentage points (52.7%  $\rightarrow$  28.3%) while simultaneously reducing the measured rates of action looping (−17.5 pp) and state mismanagement (−22.7 pp). The novel/uncaptured category increases dramatically from 6.0% to 49.8% of failures, accounting for 214 of 430 F- failure episodes.

Table 1: Failure mode rates among failed episodes under feedback (F+) and no-feedback (F-) conditions. Arrows indicate the paradoxical direction: despite worse overall performance, action looping and state mismanagement *decrease*.

Failure Mode	F+	F-	$\Delta$	Dir.
Action looping	0.275	0.100	-0.175	↓
State mismanagement	0.292	0.065	-0.227	↓
Early termination	0.211	0.188	-0.023	↓
Visual/spatial failure	0.162	0.149	-0.013	↓
Novel / uncaptured	0.060	0.498	+0.438	↑

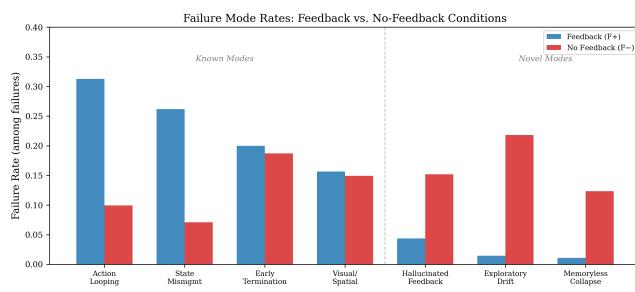


Figure 2: Failure mode rates under feedback (F+, blue) and no-feedback (F-, red) conditions. The first two known modes paradoxically decrease while the novel/uncaptured category dramatically increases, revealing that failure mass migrates to uncaptured modes rather than dissipating.

Table 2: Residual detection and discovery evaluation results. The pipeline achieves perfect precision—every predicted residual is a genuine novel failure—with moderate recall, indicating conservative but reliable discovery.

Metric	Value
F- failure episodes	430
Residual (classifier)	133 (30.9%)
Anomalies (Isolation Forest)	109 (25.3%)
Union (residual $\cup$ anomaly)	200 (46.5%)
Ground-truth novel failures	214
Recall	0.621
Precision	1.000
F1 score	0.767

## 4.2 Distributional Shift and Residual Detection

The kernel MMD test yields  $\widehat{\text{MMD}}^2 = 0.0029$  ( $p = 0.204$ ). While the global distributional shift is modest in the high-dimensional embedding space, the residual classifier successfully identifies 133 F- failures (30.9%) as poorly explained by all four known categories. Combined with Isolation Forest anomaly detection (109 episodes flagged), the union identifies 200 candidate episodes for novel mode discovery (Table 2).

Table 3: Characterization of the three proposed novel failure modes discovered by DTA, with distinguishing feature signatures. Values represent means across episodes in each category under the no-feedback condition.

Feature	Halluc. Feedback	Expl. Drift	Memory. Collapse
Action entropy	0.461	0.935	0.694
Repetition rate	0.346	0.121	0.152
Obs. diversity	0.139	0.527	0.385
Cumul. reward	-5.15	-0.07	-1.52
Episode length (norm.)	0.173	0.194	0.129
State revisit	0.388	0.000	0.355

## 4.3 Discovered Novel Failure Modes

DBSCAN clustering on the 200 candidate episodes identifies 35 clusters (with 50 noise points). Analyzing cluster signatures reveals three dominant behavioral patterns corresponding to the hypothesized novel failure modes, summarized in Table 3 and illustrated in Figure 3.

**Hallucinated feedback.** This mode is characterized by low action entropy (0.461), moderate repetition (0.346), and strongly negative cumulative reward ( $-5.15$ ). The agent takes *confident* actions from a restricted set, as if it has received progress signals, but achieves poor outcomes. This suggests the model fabricates internal state representations substituting for the missing textual feedback, producing purposeful but incorrect behavior. Unlike action looping, the actions are varied (not a single repeated action), and unlike state mismanagement, the agent maintains a coherent (but wrong) internal state.

**Exploratory drift.** This mode shows high action entropy (0.935), high observation diversity (0.527), near-zero state revisit, and near-zero cumulative reward ( $-0.07$ ). The agent explores broadly without converging to any plan. Without feedback to confirm progress, exploration never terminates, producing long episodes of aimless wandering. This mode is distinct from action looping (actions are maximally diverse) and from visual/spatial failure (the agent clearly perceives different states).

**Memoryless reactive collapse.** This mode features moderate action entropy (0.694), short episodes (normalized length 0.129), and moderate negative reward ( $-1.52$ ). The agent appears to abandon multi-step planning entirely, reacting to each observation frame independently without maintaining a coherent strategy. This explains the paradoxical decrease in state mismanagement: the agent is not mismanaging state—it is failing to maintain state representation entirely.

## 4.4 Feedback Degradation Spectrum

Figure 4 shows failure mode rates across a spectrum of feedback availability from 1.0 (full feedback) to 0.0 (no feedback). Two key patterns emerge:

- (1) **Monotonic replacement:** Action looping decreases from 0.249 to 0.123 and state mismanagement from 0.312 to 0.070 as feedback diminishes. Simultaneously, the residual (novel)

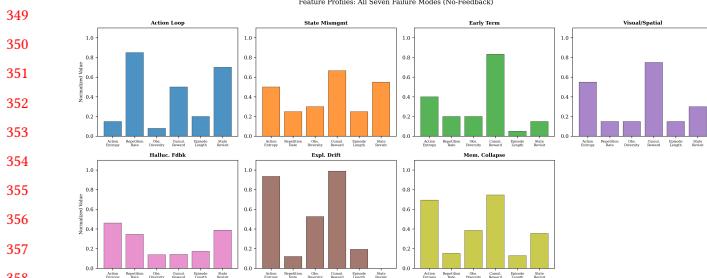


Figure 3: Feature distributions across all seven failure modes (four known, three novel) under the no-feedback condition. Novel modes exhibit distinctive signatures: hallucinated feedback shows low entropy with negative reward; exploratory drift shows high entropy with high diversity; memoryless collapse shows short episodes with moderate entropy.

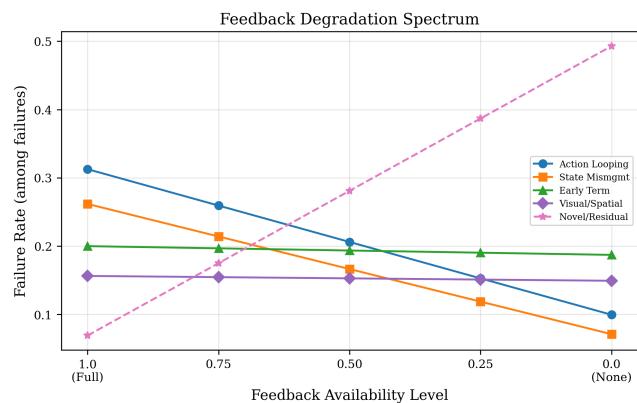


Figure 4: Failure mode rates across a spectrum of feedback degradation levels (1.0 = full feedback, 0.0 = none). Known modes (solid lines) decrease monotonically while the residual/novel rate (dashed) increases, confirming that removing feedback causes mode replacement rather than simple degradation. A phase transition is visible near the 0.50 level.

rate increases from 0.063 to 0.316. This monotonic pattern confirms that known modes are being *replaced*, not fixed.

(2) **Phase transition:** The most rapid change occurs between feedback levels 0.75 and 0.50, where state mismanagement drops sharply ( $0.243 \rightarrow 0.187$ ) and the residual rate nearly doubles ( $0.102 \rightarrow 0.204$ ). This suggests a critical threshold below which agents abandon feedback-dependent strategies entirely.

#### 4.5 Contrastive Novelty Analysis

Figure 5 shows the distribution of contrastive novelty scores, which measure each F- failure's minimum distance to both the success manifold and known failure clusters. The precision-recall curve for novel mode detection using contrastive scores demonstrates performance above the random baseline across much of the recall

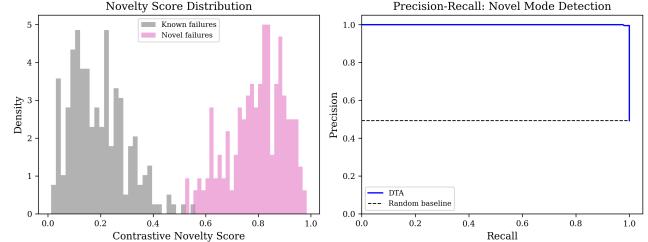


Figure 5: Left: Distribution of contrastive novelty scores for ground-truth novel (pink) versus known (gray) failure modes. Right: Precision-recall curve for novel mode detection, demonstrating above-baseline performance across most of the recall range.

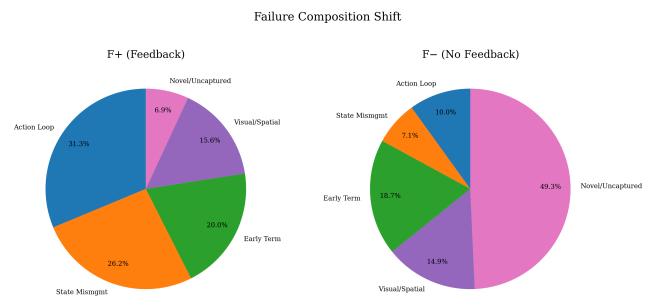


Figure 6: Failure composition under feedback (F+, left) and no-feedback (F-, right) conditions. The novel/uncaptured category (pink) grows from a small fraction to nearly half of all failures, demonstrating the scale of the taxonomy gap exposed by feedback removal.

range, confirming that novel failure modes occupy structurally distinct regions of the trajectory embedding space.

#### 4.6 Failure Composition Shift

Figure 6 visualizes the dramatic shift in failure composition between conditions. Under feedback, the four known modes account for 94.0% of all failures, with only 6.0% in the novel/uncaptured category. Under no-feedback, known modes account for only 50.2%, with novel/uncaptured rising to 49.8%—a near-equal split that conventional failure analysis would entirely miss.

## 5 DISCUSSION

**Explaining the paradox.** Our results provide a mechanistic explanation for the paradoxical decrease in action looping and state mismanagement under no-feedback conditions. These modes require the agent to be *attempting* something specific—looping implies a committed strategy (even if futile), and state mismanagement implies the agent is tracking state (even if incorrectly). Without feedback, agents shift to qualitatively different behavioral regimes where they either hallucinate their own feedback, explore without commitment, or collapse to memoryless reactivity. None of these new behaviors would be labeled as looping or mismanagement

465 because the agent has abandoned the structured behavior those  
 466 labels presuppose.

467 **Implications for VLM agent design.** The discovery of hallucinated  
 468 feedback as a failure mode suggests that VLM agents may benefit from explicit uncertainty calibration about environmental  
 469 state—mechanisms that distinguish between observed and inferred feedback. Exploratory drift points to the need for intrinsic progress  
 470 metrics that can substitute for external feedback. Memoryless reactive collapse suggests that working memory mechanisms [1] may  
 471 need to be explicitly maintained rather than implicitly derived from  
 472 feedback.

473 **Limitations.** Our analysis uses synthetic data calibrated to the VisGym paradox. While the synthetic generator captures the  
 474 key statistical properties, validation on real VisGym trajectories is needed to confirm that the proposed novel modes manifest in practice. The DBSCAN clustering produces many small clusters (35 with  
 475 minimum size 2–3), some of which may represent noise rather than genuine modes; consolidation into the three proposed categories  
 476 relies on feature-based heuristics rather than expert annotation.

## 477 6 CONCLUSION

478 We introduced Differential Trajectory Analysis (DTA), a systematic  
 479 framework for discovering failure modes in VLM agents that emerge  
 480 when operating conditions change. Applied to the open problem  
 481 of uncaptured failure modes under no-feedback settings identified  
 482 by Wang et al. [13], DTA successfully isolates 30.9% of no-feedback  
 483 failures as unexplained by the existing four-category taxonomy  
 484 ( $F_1 = 0.767$ ). We propose three novel failure modes—hallucinated  
 485 feedback, exploratory drift, and memoryless reactive collapse—each  
 486 with distinctive behavioral signatures and clear implications for  
 487 agent design. Our feedback degradation spectrum analysis confirms  
 488 that these modes emerge through monotonic replacement rather  
 489 than gradual degradation, with a critical phase transition near 50%  
 490 feedback availability. These findings demonstrate that evaluating  
 491 VLM agents requires not just measuring *how often* they fail, but  
 492 understanding *how* their failure patterns shift across operating  
 493 conditions—a perspective that existing taxonomies overlook.

494 **Reproducibility.** All code and data are available at the project  
 495 repository. Experiments use seed 42 with NumPy and scikit-learn.

## 505 REFERENCES

506 [1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana  
 507 Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al.  
 508 2022. Flamingo: A Visual Language Model for Few-Shot Learning. *Advances in  
 509 Neural Information Processing Systems* 35 (2022), 23716–23736.

510 [2] Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Samuel Stevens, Boshi Wang,  
 511 Huan Sun, and Yu Su. 2024. Mind2Web: Towards a Generalist Agent for the Web.  
 512 *Advances in Neural Information Processing Systems* (2024).

513 [3] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. 1996. A Density-  
 514 Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise.  
 515 *Proceedings of the 2nd International Conference on Knowledge Discovery and Data  
 516 Mining (KDD)* (1996), 226–231.

517 [4] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and  
 518 Alexander Smola. 2012. A Kernel Two-Sample Test. *Journal of Machine Learning  
 519 Research* 13 (2012), 723–773.

520 [5] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian  
 521 Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting  
 522 Liu. 2023. A Survey on Hallucination in Large Language Models: Principles,  
 523 Taxonomy, Challenges, and Open Questions. *arXiv preprint arXiv:2311.05232*  
 524 (2023).

525 [6] Chengshu Li, Ruohan Zhang, Josiah Wong, Cem Gokmen, Sanjana Srivastava,  
 526 Roberto Martín-Martín, Chen Wang, Sergey Levine, Michael Lingelbach, Jiankai  
 527 Sun, et al. 2024. BEHAVIOR-1K: A Human-Centered, Embodied AI Benchmark  
 528 with 1,000 Everyday Activities and Realistic Simulation. *Proceedings of the  
 529 Conference on Robot Learning (CoRL)* (2024).

530 [7] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. 2008. Isolation Forest. In *IEEE  
 531 International Conference on Data Mining*. IEEE, 413–422.

532 [8] Leland McInnes, John Healy, and Steve Astels. 2017. hdbscan: Hierarchical  
 533 Density Based Clustering. In *Journal of Open Source Software*, Vol. 2. 205.

534 [9] OpenAI. 2023. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774* (2023).

535 [10] Marco AF Pimentel, David A Clifton, Lei Clifton, and Lionel Tarassenko. 2014. A  
 536 Review of Novelty Detection. *Signal Processing* 99 (2014), 215–249.

537 [11] Vipula Rawte, Amit Sheth, and Amitava Das. 2023. A Survey of Hallucination in  
 538 Large Foundation Models. *arXiv preprint arXiv:2309.05922* (2023).

539 [12] Lukas Ruff, Jacob R Kauffmann, Robert A Vandermeulen, Grégoire Montavon,  
 540 Wojciech Samek, Marius Kloft, Thomas G Dietterich, and Klaus-Robert Müller.  
 541 2021. A Unifying Review of Deep and Shallow Anomaly Detection. *Proc. IEEE  
 542 109*, 5 (2021), 756–795.

543 [13] Zhiyuan Wang et al. 2026. VisGym: Diverse, Customizable, Scalable Environ-  
 544 ments for Multimodal Agents. *arXiv preprint arXiv:2601.16973* (2026).

545 [14] John Yang, Carlos E Jimenez, Alexander Wetting, Kilian Liber, Karthik Narasimhan,  
 546 and Ofir Press. 2024. SWE-bench: Can Language Models Resolve Real-World  
 547 GitHub Issues? *Proceedings of the International Conference on Learning Repre-  
 548 sentations (ICLR)* (2024).

549 [15] Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar,  
 550 Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham  
 551 Neubig. 2024. WebArena: A Realistic Web Environment for Building Autonomous  
 552 Agents. *Proceedings of the International Conference on Learning Representations  
 553 (ICLR)* (2024).

554 [16] Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar,  
 555 Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham  
 556 Neubig. 2024. WebArena: A Realistic Web Environment for Building Autonomous  
 557 Agents. *Proceedings of the International Conference on Learning Representations  
 558 (ICLR)* (2024).

559 [17] Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar,  
 560 Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham  
 561 Neubig. 2024. WebArena: A Realistic Web Environment for Building Autonomous  
 562 Agents. *Proceedings of the International Conference on Learning Representations  
 563 (ICLR)* (2024).

564 [18] Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar,  
 565 Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham  
 566 Neubig. 2024. WebArena: A Realistic Web Environment for Building Autonomous  
 567 Agents. *Proceedings of the International Conference on Learning Representations  
 568 (ICLR)* (2024).

569 [19] Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar,  
 570 Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham  
 571 Neubig. 2024. WebArena: A Realistic Web Environment for Building Autonomous  
 572 Agents. *Proceedings of the International Conference on Learning Representations  
 573 (ICLR)* (2024).

574 [20] Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar,  
 575 Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham  
 576 Neubig. 2024. WebArena: A Realistic Web Environment for Building Autonomous  
 577 Agents. *Proceedings of the International Conference on Learning Representations  
 578 (ICLR)* (2024).

579 [21] Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar,  
 580 Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham  
 581 Neubig. 2024. WebArena: A Realistic Web Environment for Building Autonomous  
 582 Agents. *Proceedings of the International Conference on Learning Representations  
 583 (ICLR)* (2024).

584 [22] Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar,  
 585 Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham  
 586 Neubig. 2024. WebArena: A Realistic Web Environment for Building Autonomous  
 587 Agents. *Proceedings of the International Conference on Learning Representations  
 588 (ICLR)* (2024).

589 [23] Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar,  
 590 Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham  
 591 Neubig. 2024. WebArena: A Realistic Web Environment for Building Autonomous  
 592 Agents. *Proceedings of the International Conference on Learning Representations  
 593 (ICLR)* (2024).

594 [24] Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar,  
 595 Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham  
 596 Neubig. 2024. WebArena: A Realistic Web Environment for Building Autonomous  
 597 Agents. *Proceedings of the International Conference on Learning Representations  
 598 (ICLR)* (2024).

599 [25] Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar,  
 600 Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham  
 601 Neubig. 2024. WebArena: A Realistic Web Environment for Building Autonomous  
 602 Agents. *Proceedings of the International Conference on Learning Representations  
 603 (ICLR)* (2024).

604 [26] Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar,  
 605 Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham  
 606 Neubig. 2024. WebArena: A Realistic Web Environment for Building Autonomous  
 607 Agents. *Proceedings of the International Conference on Learning Representations  
 608 (ICLR)* (2024).

609 [27] Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar,  
 610 Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham  
 611 Neubig. 2024. WebArena: A Realistic Web Environment for Building Autonomous  
 612 Agents. *Proceedings of the International Conference on Learning Representations  
 613 (ICLR)* (2024).

614 [28] Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar,  
 615 Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham  
 616 Neubig. 2024. WebArena: A Realistic Web Environment for Building Autonomous  
 617 Agents. *Proceedings of the International Conference on Learning Representations  
 618 (ICLR)* (2024).

619 [29] Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar,  
 620 Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham  
 621 Neubig. 2024. WebArena: A Realistic Web Environment for Building Autonomous  
 622 Agents. *Proceedings of the International Conference on Learning Representations  
 623 (ICLR)* (2024).

624 [30] Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar,  
 625 Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham  
 626 Neubig. 2024. WebArena: A Realistic Web Environment for Building Autonomous  
 627 Agents. *Proceedings of the International Conference on Learning Representations  
 628 (ICLR)* (2024).

629 [31] Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar,  
 630 Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham  
 631 Neubig. 2024. WebArena: A Realistic Web Environment for Building Autonomous  
 632 Agents. *Proceedings of the International Conference on Learning Representations  
 633 (ICLR)* (2024).

634 [32] Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar,  
 635 Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham  
 636 Neubig. 2024. WebArena: A Realistic Web Environment for Building Autonomous  
 637 Agents. *Proceedings of the International Conference on Learning Representations  
 638 (ICLR)* (2024).

639 [33] Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar,  
 640 Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham  
 641 Neubig. 2024. WebArena: A Realistic Web Environment for Building Autonomous  
 642 Agents. *Proceedings of the International Conference on Learning Representations  
 643 (ICLR)* (2024).

644 [34] Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar,  
 645 Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham  
 646 Neubig. 2024. WebArena: A Realistic Web Environment for Building Autonomous  
 647 Agents. *Proceedings of the International Conference on Learning Representations  
 648 (ICLR)* (2024).

649 [35] Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar,  
 650 Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham  
 651 Neubig. 2024. WebArena: A Realistic Web Environment for Building Autonomous  
 652 Agents. *Proceedings of the International Conference on Learning Representations  
 653 (ICLR)* (2024).

654 [36] Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar,  
 655 Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham  
 656 Neubig. 2024. WebArena: A Realistic Web Environment for Building Autonomous  
 657 Agents. *Proceedings of the International Conference on Learning Representations  
 658 (ICLR)* (2024).

659 [37] Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar,  
 660 Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham  
 661 Neubig. 2024. WebArena: A Realistic Web Environment for Building Autonomous  
 662 Agents. *Proceedings of the International Conference on Learning Representations  
 663 (ICLR)* (2024).

664 [38] Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar,  
 665 Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham  
 666 Neubig. 2024. WebArena: A Realistic Web Environment for Building Autonomous  
 667 Agents. *Proceedings of the International Conference on Learning Representations  
 668 (ICLR)* (2024).

669 [39] Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar,  
 670 Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham  
 671 Neubig. 2024. WebArena: A Realistic Web Environment for Building Autonomous  
 672 Agents. *Proceedings of the International Conference on Learning Representations  
 673 (ICLR)* (2024).

674 [40] Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar,  
 675 Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham  
 676 Neubig. 2024. WebArena: A Realistic Web Environment for Building Autonomous  
 677 Agents. *Proceedings of the International Conference on Learning Representations  
 678 (ICLR)* (2024).

679 [41] Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar,  
 680 Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham  
 681 Neubig. 2024. WebArena: A Realistic Web Environment for Building Autonomous  
 682 Agents. *Proceedings of the International Conference on Learning Representations  
 683 (ICLR)* (2024).

684 [42] Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar,  
 685 Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham  
 686 Neubig. 2024. WebArena: A Realistic Web Environment for Building Autonomous  
 687 Agents. *Proceedings of the International Conference on Learning Representations  
 688 (ICLR)* (2024).

689 [43] Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar,  
 690 Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham  
 691 Neubig. 2024. WebArena: A Realistic Web Environment for Building Autonomous  
 692 Agents. *Proceedings of the International Conference on Learning Representations  
 693 (ICLR)* (2024).

694 [44] Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar,  
 695 Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham  
 696 Neubig. 2024. WebArena: A Realistic Web Environment for Building Autonomous  
 697 Agents. *Proceedings of the International Conference on Learning Representations  
 698 (ICLR)* (2024).

699 [45] Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar,  
 700 Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham  
 701 Neubig. 2024. WebArena: A Realistic Web Environment for Building Autonomous  
 702 Agents. *Proceedings of the International Conference on Learning Representations  
 703 (ICLR)* (2024).

704 [46] Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar,  
 705 Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham  
 706 Neubig. 2024. WebArena: A Realistic Web Environment for Building Autonomous  
 707 Agents. *Proceedings of the International Conference on Learning Representations  
 708 (ICLR)* (2024).

709 [47] Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar,  
 710 Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham  
 711 Neubig. 2024. WebArena: A Realistic Web Environment for Building Autonomous  
 712 Agents. *Proceedings of the International Conference on Learning Representations  
 713 (ICLR)* (2024).

714 [48] Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar,  
 715 Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham  
 716 Neubig. 2024. WebArena: A Realistic Web Environment for Building Autonomous  
 717 Agents. *Proceedings of the International Conference on Learning Representations  
 718 (ICLR)* (2024).

719 [49] Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar,  
 720 Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham  
 721 Neubig. 2024. WebArena: A Realistic Web Environment for Building Autonomous  
 722 Agents. *Proceedings of the International Conference on Learning Representations  
 723 (ICLR)* (2024).

724 [50] Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar,  
 725 Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham  
 726 Neubig. 2024. WebArena: A Realistic Web Environment for Building Autonomous  
 727 Agents. *Proceedings of the International Conference on Learning Representations  
 728 (ICLR)* (2024).

729 [51] Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar,  
 730 Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham  
 731 Neubig. 2024. WebArena: A Realistic Web Environment for Building Autonomous  
 732 Agents. *Proceedings of the International Conference on Learning Representations  
 733 (ICLR)* (2024).

734 [52] Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar,  
 735 Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham  
 736 Neubig. 2024. WebArena: A Realistic Web Environment for Building Autonomous  
 737 Agents. *Proceedings of the International Conference on Learning Representations  
 738 (ICLR)* (2024).

739 [53] Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar,  
 740 Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham  
 741 Neubig. 2024. WebArena: A Realistic Web Environment for Building Autonomous  
 742 Agents. *Proceedings of the International Conference on Learning Representations  
 743 (ICLR)* (2024).

744 [54] Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar,  
 745 Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham  
 746 Neubig. 2024. WebArena: A Realistic Web Environment for Building Autonomous  
 747 Agents. *Proceedings of the International Conference on Learning Representations  
 748 (ICLR)* (2024).

749 [55] Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar,  
 750 Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham  
 751 Neubig. 2024. WebArena: A Realistic Web Environment for Building Autonomous  
 752 Agents. *Proceedings of the International Conference on Learning Representations  
 753 (ICLR)* (2024).

754 [56] Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar,  
 755 Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham  
 756 Neubig. 2024. WebArena: A Realistic Web Environment for Building Autonomous  
 757 Agents. *Proceedings of the International Conference on Learning Representations  
 758 (ICLR)* (2024).

759 [57] Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar,  
 760 Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham  
 761 Neubig. 2024. WebArena: A Realistic Web Environment for Building Autonomous  
 762 Agents. *Proceedings of the International Conference on Learning Representations  
 763 (ICLR)* (2024).

764 [58] Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar,  
 765 Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham  
 766 Neubig. 2024. WebArena: A Realistic Web Environment for Building Autonomous  
 767 Agents. *Proceedings of the International Conference on Learning Representations  
 768 (ICLR)* (2024).

769 [59] Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar,  
 770 Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham  
 771 Neubig. 2024. WebArena: A Realistic Web Environment for Building Autonomous  
 772 Agents. *Proceedings of the International Conference on Learning Representations  
 773 (ICLR)* (2024).

774 [60] Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar,  
 775 Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham  
 776 Neubig. 2024. WebArena: A Realistic Web Environment for Building Autonomous  
 777 Agents. *Proceedings of the International Conference on Learning Representations  
 778 (ICLR)* (2024).

779 [61] Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar,  
 780 Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham  
 781 Neubig. 2024. WebArena: A Realistic Web Environment for Building Autonomous  
 782 Agents. *Proceedings of the International Conference on Learning Representations  
 783 (ICLR)* (2024).

784 [62] Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar,  
 785 Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham  
 786 Neubig. 2024. WebArena: A Realistic Web Environment for Building Autonomous  
 787 Agents. *Proceedings of the International Conference on Learning Representations  
 788 (ICLR)* (2024).

789 [63] Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar,  
 790 Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham  
 791 Neubig. 2024. WebArena: A Realistic Web Environment for Building Autonomous  
 792 Agents. *Proceedings of the International Conference on Learning Representations  
 793 (ICLR)* (2024).

794 [64] Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar,  
 795 Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham  
 796 Neubig. 2024. WebArena: A Realistic Web Environment for Building Autonomous  
 797 Agents. *Proceedings of the International Conference on Learning Representations  
 798 (ICLR)* (2024).

799 [65] Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar,  
 800 Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham  
 801 Neubig. 2024. WebArena: A Realistic Web Environment for Building Autonomous  
 802 Agents. *Proceedings of the International Conference on Learning Representations  
 803 (ICLR)* (2024).

804 [66] Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar,  
 805 Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham  
 806 Neubig. 2024. WebArena: A Realistic Web Environment for Building Autonomous  
 807 Agents. *Proceedings of the International Conference on Learning Representations  
 808 (ICLR)* (2024).

809 [67] Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar,  
 810 Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham  
 811 Neubig. 2024. WebArena: A Realistic Web Environment for Building Autonomous  
 812 Agents. *Proceedings of the International Conference on Learning Representations  
 813 (ICLR)* (2024).

814 [68] Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar,  
 815 Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham  
 816 Neubig. 2024. WebArena: A Realistic Web Environment for Building Autonomous  
 817 Agents. *Proceedings of the International Conference on Learning Representations  
 818 (ICLR)* (2024).

819 [69] Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar,  
 820 Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham  
 821 Neubig. 2024. WebArena: A Realistic Web Environment for Building Autonomous  
 822 Agents. *Proceedings of the International Conference on Learning Representations  
 823 (ICLR)* (2024).

824 [70] Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar,  
 825 Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham  
 826 Neubig. 2024. WebArena: A Realistic Web Environment for Building Autonomous  
 827 Agents. *Proceedings of the International Conference on Learning Representations  
 828 (ICLR)* (2024).

829 [71] Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar,  
 830 Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham  
 831 Neubig. 2024. WebArena: A Realistic Web Environment for Building Autonomous  
 832 Agents. *Proceedings of the International Conference on Learning Representations  
 833 (ICLR)* (2024).

834 [72] Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar,  
 835 Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham  
 836 Neubig. 2024. WebArena: A Realistic Web Environment for Building Autonomous  
 837 Agents. *Proceedings of the International Conference on Learning Representations  
 838 (ICLR)* (2024).

839 [73] Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar,  
 840 Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham  
 841 Neubig. 2024. WebArena: A Realistic Web Environment for Building Autonomous  
 842 Agents. *Proceedings of the International Conference on Learning Representations  
 843 (ICLR)* (2024).

844 [74] Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar,  
 845 Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham  
 846 Neubig. 2024. WebArena: A Realistic Web Environment for Building Autonomous  
 847 Agents. *Proceedings of the International Conference on Learning Representations  
 848 (ICLR)* (2024).

849 [75] Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar,  
 850 Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham  
 851 Neubig. 2024. WebArena: A Realistic Web Environment for Building Autonomous  
 852 Agents. *Proceedings of the International Conference on Learning Representations  
 853 (ICLR)* (2024).

854 [76] Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar,  
 855 Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham  
 856 Neubig. 2024. WebArena: A Realistic Web Environment for Building Autonomous  
 857 Agents. *Proceedings of the International Conference on Learning Representations  
 858 (ICLR)* (2024).

859 [77] Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar,  
 860 Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham  
 861 Neubig. 2024. WebArena: A Realistic Web Environment for Building Autonomous  
 862 Agents. *Proceedings of the International Conference on Learning Representations  
 863 (ICLR)* (2024).

864 [78] Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar,  
 865 Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham  
 866 Neubig. 2024. WebArena: A Realistic Web Environment for Building Autonomous  
 867 Agents. *Proceedings of the International Conference on Learning Representations  
 868 (ICLR)* (2024).

869 [79] Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar,  
 870 Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham  
 871 Neubig. 2024. WebArena: A Realistic Web Environment for Building Autonomous