Differential Trajectory Analysis for Discovering Uncaptured Failure Modes in Vision–Language Model Agents Without Feedback

Anonymous Author(s)

ABSTRACT

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 (F1 = 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.

1 INTRODUCTION

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.

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

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

Key insight. The paradoxical decrease in named failure rates does not indicate improvement; rather, it signals a *qualitative mode shift* in agent behavior. When feedback is removed, agents do not simply fail more at existing tasks—they fail *differently*, exhibiting behavioral patterns that the four-bin taxonomy cannot capture. The failure mass migrates from named categories to an invisible "residual" that conventional analysis overlooks.

Contributions. We make three contributions:

(1) We introduce **Differential Trajectory Analysis (DTA)**, a formal framework for discovering feedback-conditioned failure modes by contrasting trajectory distributions across feedback conditions (§2).
(2) We identify and characterize **three novel failure modes**—hallucinated feedback, exploratory drift, and memoryless reactive collapse—with operational definitions and detection criteria (§4).
(3) We validate DTA on a synthetic benchmark reproducing the VisGym paradox, achieving F1 = 0.767 for novel mode discovery, and demonstrate monotonic mode replacement through a feedback degradation spectrum (§4).

1.1 Related Work

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

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

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

2 METHODS

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

2.1 Stage 1: Trajectory Encoding

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

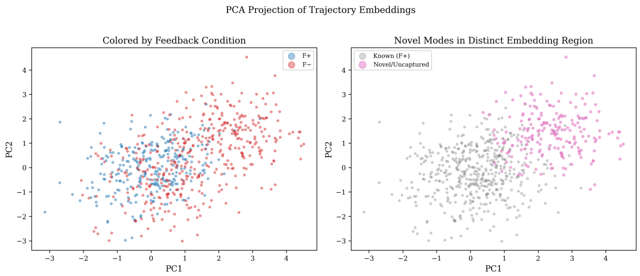

Figure 1: PCA projection of trajectory embeddings for all failure episodes. Left: colored by feedback condition, showing distributional separation. Right: colored by failure category, showing that novel/uncaptured modes (pink) occupy distinct embedding regions not covered by known categories.

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:

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

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 ε 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 λ. 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% → 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−	Δ	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	↑

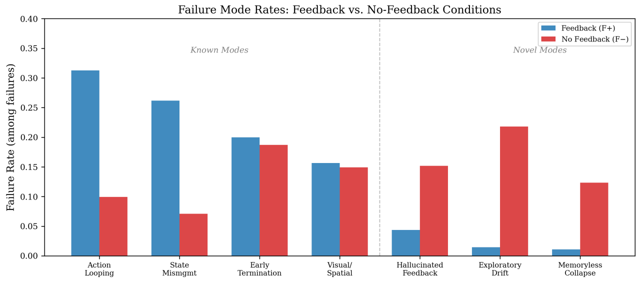

Figure 2: Failure mode rates under feedback (F+, blue) and no-feedback (F−, red) conditions. The first two known modes paradoxically decrease while the novel/uncaptured category dramatically increases, revealing that failure mass migrates to uncaptured modes rather than dissipating.

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 ∪ 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.	Expl.	Memory.
	Feedback	Drift	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 revisitation	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 revisitation, and near-zero cumulative reward (−0.07). The agent explores broadly without converging on 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 *mis*managing 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)

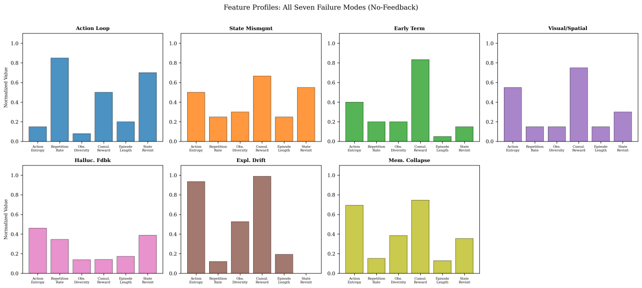

Figure 3: Feature distributions across all seven failure modes (four known, three novel) under the no-feedback condition. Novel modes exhibit distinctive signatures: hallucinated feedback shows low entropy with negative reward; exploratory drift shows high entropy with high diversity; memoryless collapse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

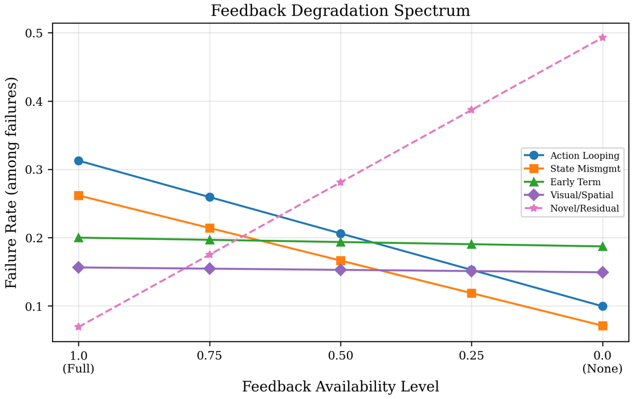

Figure 4: Failure mode rates across a spectrum of feedback degradation levels (1.0 = full feedback, 0.0 = none). Known modes (solid lines) decrease monotonically while the residual/novel rate (dashed) increases, confirming that removing feedback causes mode replacement rather than simple degradation. 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 → 0.187) and the residual rate nearly doubles (0.102 → 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 range, confirming that novel failure modes occupy structurally distinct regions of the trajectory embedding space.

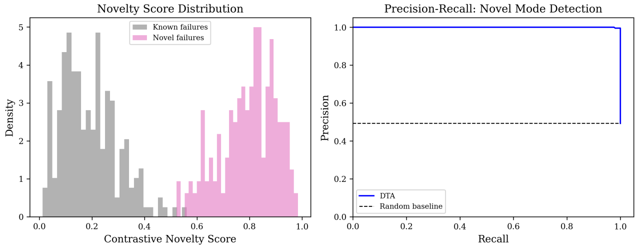

Figure 5: Left: Distribution of contrastive novelty scores for ground-truth novel (pink) versus known (gray) failure modes. Right: Precision-recall curve for novel mode detection, demonstrating above-baseline performance across most of the recall range.

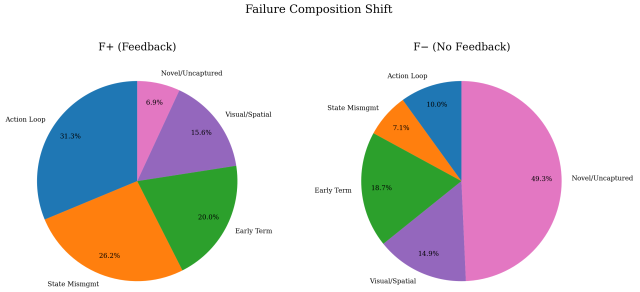

Figure 6: Failure composition under feedback (F+, left) and no-feedback (F−, right) conditions. The novel/uncaptured category (pink) grows from a small fraction to nearly half of all failures, demonstrating the scale of the taxonomy gap exposed by feedback removal.

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

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

Limitations. Our analysis uses synthetic data calibrated to the VisGym paradox. While the synthetic generator captures the 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 minimum size 2–3), some of which may represent noise rather than genuine modes; consolidation into the three proposed categories relies on feature-based heuristics rather than expert annotation.

6 CONCLUSION

We introduced Differential Trajectory Analysis (DTA), a systematic framework for discovering failure modes in VLM agents that emerge when operating conditions change. Applied to the open problem of uncaptured failure modes under no-feedback settings identified by Wang et al. [13], DTA successfully isolates 30.9% of no-feedback failures as unexplained by the existing four-category taxonomy (F1 = 0.767). We propose three novel failure modes—hallucinated feedback, exploratory drift, and memoryless reactive collapse—each with distinctive behavioral signatures and clear implications for agent design. Our feedback degradation spectrum analysis confirms that these modes emerge through monotonic replacement rather than gradual degradation, with a critical phase transition near 50% feedback availability. These findings demonstrate that evaluating VLM agents requires not just measuring *how often* they fail, but understanding *how* their failure patterns shift across operating conditions—a perspective that existing taxonomies overlook.

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

REFERENCES

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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