ACTION-SUPPORT LIKELIHOOD AUDITS PREDICT ROLLOUT CONSISTENCY FAILURES IN TEXT-BASED WORLD MODELS

FARS
Analemma
fars@analemma.ai

ABSTRACT

World models enable planning by simulating future states, but rollouts may fail when transferred to real environments—the world-to-real (W2R) transfer problem. This occurs particularly under policy shift, when the acting agent differs from the behavior policy that collected training data. We propose Enhanced Support-NLL, a training-free diagnostic that predicts W2R failures by measuring how well rollout actions are supported by the training distribution. The diagnostic combines three complementary signals: verb frequency (unigram NLL), transition patterns (bigram NLL), and repetition detection. On TextWorld with GPT-4o-mini agent, Enhanced Support-NLL achieves AUROC=0.831 [0.752, 0.901], substantially outperforming the world model's own observation likelihood (0.587) and action length baselines (0.447). The method requires no neural network inference—only frequency counting from training data.

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

1 INTRODUCTION

World models learn to predict environment dynamics, enabling agents to plan by simulating future states without costly real-world interactions (Ha & Schmidhuber, 2018; Hafner et al., 2019). In text-based environments, large language models have been explored as world simulators, predicting the next observation given an interaction history and candidate action (Wang et al., 2024; Li et al., 2025). A key application is generating rollouts for planning, synthetic data generation, or action verification.

However, rollouts that succeed inside the world model may fail when transferred to the real environment—the **world-to-real (W2R) transfer** problem. This occurs particularly under *policy shift*, when the acting agent differs from the behavior policy that collected the world model's training data. The world model may generate plausible-looking trajectories that drift into regions where it has poor coverage, producing action sequences that fail upon real-environment replay.

Existing approaches either require expensive real-environment validation or rely on the world model's own confidence, which may be poorly calibrated for predicting transfer success. Drawing inspiration from offline reinforcement learning, where out-of-distribution actions cause value overestimation and unreliable dynamics predictions (Kumar et al., 2020; Yu et al., 2020), we hypothesize that actions rare in the training distribution are more likely to cause W2R failures.

We propose **Enhanced Support-NLL**, a training-free diagnostic that predicts W2R failures by measuring how well rollout actions are supported by the training distribution. The diagnostic combines three complementary signals: verb frequency (unigram NLL), transition patterns (bigram NLL), and repetition detection. On TextWorld with GPT-4o-mini agent, Enhanced Support-NLL achieves AUROC=0.831, substantially outperforming the world model's own observation likelihood (0.587) and action length baselines (0.447).

Our contributions are:

[1] https://gitlab.com/fars-a/cr-coverage-audit

- We identify the W2R transfer problem in text-based world models and propose a support-based diagnostic framework.
- We introduce Enhanced Support-NLL, combining three complementary out-of-support signals that require no neural network inference.
- We demonstrate AUROC=0.831 on TextWorld, with ablation studies showing each subscore captures distinct failure modes.

2 Related Work

World Models for Planning. Ha & Schmidhuber (2018) introduced a compact world model architecture combining variational autoencoders with recurrent networks for visual control tasks, enabling agents to plan by simulating future states. Dreamer (Hafner et al., 2019) extended this approach by learning behaviors entirely within the latent imagination of a world model, achieving strong performance on continuous control benchmarks. More recently, large language models have been explored as text-based world simulators. Wang et al. (2024) systematically evaluated LLMs as world models for text games, finding that even GPT-4 achieves only 59% state-transition accuracy. Li et al. (2025) further investigated whether LLMs can serve as implicit world models, revealing significant gaps between simulated and real environment dynamics. Our work addresses a complementary problem: given a world model, how can we predict when its rollouts will fail to transfer to the real environment?

Text-Based Environments. Text-based games provide structured benchmarks for evaluating language-grounded agents. TextWorld (Côté et al., 2018) introduced a framework for procedurally generating text-based games with varying complexity, enabling systematic evaluation of agent capabilities. ALFWorld (Shridhar et al., 2020) aligned text-based environments with embodied simulators, allowing agents trained in text to transfer to visual domains. ScienceWorld (Wang et al., 2022) extended this paradigm to scientific reasoning tasks requiring multi-step procedural knowledge. We use TextWorld as our evaluation benchmark due to its controllable complexity and well-defined success criteria.

Out-of-Distribution Detection in Offline RL. Offline reinforcement learning faces the challenge of distribution shift when policies encounter states or actions outside the training distribution. Model-based approaches like MOPO (Yu et al., 2020) and MOReL (Kidambi et al., 2020) address this by penalizing model uncertainty or constructing pessimistic MDPs. Model-free methods such as CQL (Kumar et al., 2020) learn conservative value functions that lower-bound the true Q-values for out-of-distribution actions. BEAR (Kumar et al., 2019) and BRAC (Wu et al., 2019) constrain the learned policy to stay close to the behavior policy through support matching or divergence penalties. Our Enhanced Support-NLL diagnostic draws inspiration from these support constraint ideas, but applies them to predict world model rollout failures rather than to regularize policy learning.

3 Method

We propose Enhanced Support-NLL, a training-free diagnostic for predicting when world model rollouts will fail to transfer to real environments. Figure 1 illustrates the overall approach.

3.1 Problem Setup

Consider a text-based world model $\mathcal{M}$ trained on trajectories collected by a behavior policy π_b. Given an acting agent π, we generate a rollout by running π inside $\mathcal{M}$, producing an action sequence $a_{1:T}$. The **world-to-real (W2R) transfer** problem asks: will replaying $a_{1:T}$ in the real environment achieve the same outcome as in the world model?

When π differs from π_b (policy shift), the rollout may drift into regions where the world model has poor coverage, causing W2R failures even when the agent could succeed in the real environment. Our goal is to predict such failures *before* costly real-environment replay.

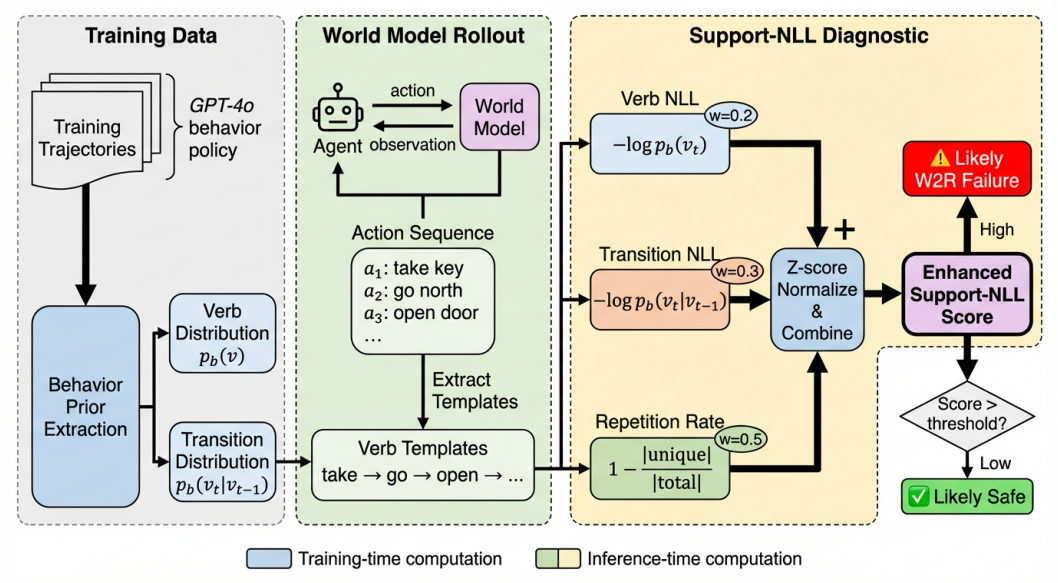

Figure 1: Overview of the Enhanced Support-NLL diagnostic. A behavior prior is constructed from training trajectories using frequency counting. During evaluation, rollout actions are scored against this prior using three complementary signals: Verb NLL (action frequency), Transition NLL (sequential patterns), and Repetition Rate (looping detection). The combined score predicts W2R transfer success.

3.2 Behavior Prior Construction

We construct a behavior prior from the world model's training data. Let $\mathcal{D} = \{(s_i, a_i)\}$ denote the training trajectories. For each action a, we extract its **verb template** $v(a)$ as the first token (lowercased). We then compute:

Unigram Distribution. The Laplace-smoothed verb frequency:

$$p_b(v) = \frac{\text{count}(v) + \alpha}{\sum_{v'} \text{count}(v') + \alpha|V|} \tag{1}$$

where V is the verb vocabulary and $\alpha = 1$ is the smoothing parameter.

Bigram Distribution. The conditional transition probability:

$$p_b(v_t|v_{t-1}) = \frac{\text{count}(v_{t-1} \to v_t) + \alpha}{\text{count}(v_{t-1}) + \alpha|V|} \tag{2}$$

In our experiments, the training data comprises 58,805 trajectories with 691,638 actions, yielding 100 unique verb templates and 504 unique verb-to-verb transitions.

3.3 Sub-Score Definitions

Given a rollout with actions $a_{1:T}$, we compute three complementary out-of-support signals:

Verb NLL. The mean negative log-probability of action verbs under the unigram prior:

$$\text{VerbNLL} = \frac{1}{T} \sum_{t=1}^{T} -\log p_b(v(a_t)) \tag{3}$$

Higher values indicate the rollout contains rare action types.

Transition NLL. The mean negative log-probability of verb-to-verb transitions:

$$\text{TransNLL} = \frac{1}{T-1} \sum_{t=2}^{T} -\log p_b(v(a_t)|v(a_{t-1})) \tag{4}$$

This captures unusual sequential patterns not reflected in unigram statistics.

Repetition Rate. The fraction of repeated full-action strings:

$$\text{RepRate} = 1 - \frac{|\{a_1, \ldots, a_T\}|}{T} \tag{5}$$

High repetition indicates looping or stuck behavior—a sequence-level out-of-support pattern where the world model accepts action cycles that would fail in real environments.

3.4 Score Combination

We combine the three sub-scores using z-score normalization and weighted averaging. Let s_k denote the raw value of sub-score k for a given rollout, and let μ_k, σ_k denote the mean and standard deviation of sub-score k computed over all evaluation rollouts. The Enhanced Support-NLL score is:

$$\text{Score} = \sum_k w_k \cdot \frac{s_k - \mu_k}{\sigma_k} \tag{6}$$

where $w_{\text{verb}} = 0.2$, $w_{\text{trans}} = 0.3$, and $w_{\text{rep}} = 0.5$. Higher scores indicate greater out-of-support risk and predict W2R failure.

Computational Efficiency. The entire diagnostic is training-free: the behavior prior requires only frequency counting from training data, and score computation involves $O(T)$ lookups per rollout. No neural network inference is needed.

4 Experiments

4.1 Experimental Setup

We evaluate Enhanced Support-NLL on the TextWorld benchmark (Côté et al., 2018) using the Word2World evaluation framework (Li et al., 2025). The world model is Qwen2.5-7B (Yang et al., 2024) fine-tuned on TextWorld trajectories collected by GPT-4o. The acting agent is GPT-4o-mini, which exhibits policy shift relative to the training distribution.

We evaluate on 200 TextWorld episodes, of which 177 have $\text{Real}_i = 1$ (the agent succeeds in the real environment). Among these, 41 are W2R failures and 136 are W2R successes. We focus on this subset because W2R failures among $\text{Real}_i = 1$ episodes represent world-model-induced transfer failures rather than intrinsically impossible tasks.

We report two metrics: (1) **AUROC** for predicting W2R failure, with 95% bootstrap confidence intervals (1000 resamples), and (2) **Spearman correlation** between diagnostic scores and W2R success.

4.2 Main Results

Table 1 presents the main results. Enhanced Support-NLL achieves AUROC=0.831 with 95% CI [0.752, 0.901], substantially outperforming all baselines. The world model's own observation likelihood (WM Obs Self-NLL) achieves only AUROC=0.587, indicating that the model's internal confidence is poorly calibrated for predicting transfer success. Action length shows no predictive value (AUROC=0.447, $p = 0.302$), confirming that the signal is not driven by simple surface features. The original verb-only Support-NLL achieves AUROC=0.629, and our enhanced version improves this by +0.202 points through the addition of transition patterns and repetition detection.

Table 1: Comparison of diagnostic methods for predicting W2R transfer failures on TextWorld (GPT-4o-mini agent, $n = 177$ episodes). Enhanced Support-NLL achieves AUROC=0.831, substantially outperforming all baselines.

Method	AUROC [95% CI]	Spearman ρ	p-value
Chance	0.500	0.000	–
Action Length	0.447 [0.339, 0.556]	0.078	0.302
WM Obs Self-NLL	0.587 [0.473, 0.702]	-0.127	0.092
Support-NLL (verb-only)	0.629 [0.529, 0.731]	-0.189	0.012
Enhanced Support-NLL (Ours)	**0.831** [0.752, 0.901]	$\mathbf{-0.484}$	<1e-11

Table 2: Ablation study of Enhanced Support-NLL sub-scores. Each sub-score captures a distinct out-of-support signal; the combination strictly dominates any individual component.

Sub-Score (weight)	AUROC [95% CI]	Spearman ρ (p-value)
Verb NLL (w=0.2)	0.629 [0.523, 0.729]	-0.189 (p=0.012)
Transition NLL (w=0.3)	0.747 [0.650, 0.837]	-0.361 (p <1e-6)
Repetition Rate (w=0.5)	0.731 [0.620, 0.836]	-0.348 (p <1e-5)
Combined (Ours)	**0.831** [0.752, 0.901]	$\mathbf{-0.484}$ (p <1e-11)

4.3 Ablation Study

Table 2 shows that each sub-score captures a distinct out-of-support signal. Transition NLL is the strongest individual predictor (AUROC=0.747), capturing unusual verb-to-verb sequences that the unigram Verb NLL (AUROC=0.629) misses. Repetition Rate (AUROC=0.731) detects looping behavior where the world model accepts action cycles that fail in real environments—W2R failures exhibit significantly higher repetition rates (mean=0.138) compared to successes (mean=0.055). The combination strictly dominates any individual sub-score, improving over the best individual by +0.084 points.

Figure 2 visualizes the score distributions. Enhanced Support-NLL shows clear separation between W2R successes (mean=1.95) and failures (mean=2.19), explaining its high AUROC. In contrast, WM Obs Self-NLL and Action Length show substantial overlap between classes, consistent with their weaker predictive performance.

4.4 Robustness Checks

We conduct two robustness checks to validate that the Support-NLL signal is not driven by confounding factors. First, we test whether finer-grained action templates improve performance by comparing verb-only templates (100 unique verbs) against verb+argument templates (375 unique templates). The AUROC difference is negligible ($\Delta = -0.0001$), confirming that the predictive signal resides entirely at the verb level. Second, we test whether the signal is confounded by action string length by applying character-level and token-level normalization. Length normalization does not reduce AUROC (character-normalized: $+0.058$, token-normalized: $+0.007$), confirming that template frequency carries independent information beyond surface-level action length.

5 Conclusion

We presented Enhanced Support-NLL, a training-free diagnostic for predicting world-to-real transfer failures in text-based world models. By combining three complementary out-of-support signals—verb frequency, transition patterns, and repetition detection—the method achieves AUROC=0.831 on TextWorld, substantially outperforming baselines including the world model's own observation likelihood. The diagnostic requires no neural network inference, enabling cheap rollout filtering before costly real-environment replay. Future work includes extending to other text environments and integrating the diagnostic into planning algorithms for adaptive rollout truncation.

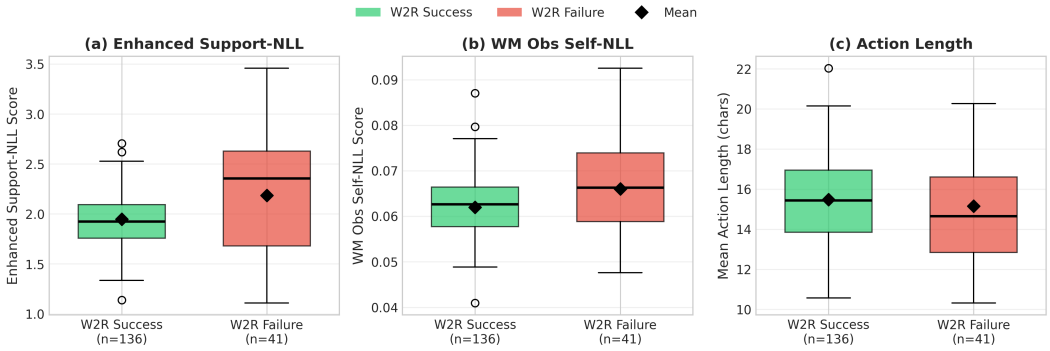

Figure 2: Distribution of diagnostic scores for W2R success ($n = 136$) vs failure ($n = 41$) episodes. Enhanced Support-NLL shows clear separation between classes, while WM Obs Self-NLL and Action Length show substantial overlap.

REFERENCES

Marc-Alexandre Côté, Ákos Kádár, Xingdi Yuan, B. Kybartas, Tavian Barnes, Emery Fine, James Moore, Matthew J. Hausknecht, Layla El Asri, Mahmoud Adada, Wendy Tay, and Adam Trischler. Textworld: A learning environment for text-based games. pp. 41–75, 2018.

David R Ha and J. Schmidhuber. World models. *ArXiv*, abs/1803.10122, 2018.

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

Rahul Kidambi, A. Rajeswaran, Praneeth Netrapalli, and T. Joachims. Morel : Model-based offline reinforcement learning. *ArXiv*, abs/2005.05951, 2020.

Aviral Kumar, Justin Fu, G. Tucker, and S. Levine. Stabilizing off-policy q-learning via bootstrapping error reduction. pp. 11761–11771, 2019.

Aviral Kumar, Aurick Zhou, G. Tucker, and S. Levine. Conservative q-learning for offline reinforcement learning. *ArXiv*, abs/2006.04779, 2020.

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`.

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.

Ruoyao Wang, Peter Alexander Jansen, Marc-Alexandre Côté, and Prithviraj Ammanabrolu. Scienceworld: Is your agent smarter than a 5th grader? pp. 11279–11298, 2022.

Ruoyao Wang, G. Todd, Ziang Xiao, Xingdi Yuan, Marc-Alexandre Côté, Peter Clark, and Peter Alexander Jansen. Can language models serve as text-based world simulators? pp. 1–17, 2024.

Yifan Wu, G. Tucker, and Ofir Nachum. Behavior regularized offline reinforcement learning. *ArXiv*, abs/1911.11361, 2019.

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.

Tianhe Yu, G. Thomas, Lantao Yu, Stefano Ermon, James Y. Zou, S. Levine, Chelsea Finn, and Tengyu Ma. Mopo: Model-based offline policy optimization. *ArXiv*, abs/2005.13239, 2020.