Sparse Binary Reward Degrades Agent Uncertainty Quantification in Multi-Seed Evaluation

Teganmosibineba Oluwatofarati Jegede [1]

Abstract

Uncertainty quantification (UQ) for RL-trained agents, including conformal prediction, calibration, and risk control, implicitly require that the agent produces consistent behaviour across training conditions. We present empirical evidence that this consistency requirement is violated when agents are trained with sparse binary reward signals. A Standard REINFORCE agent (binary task reward) trained on LLaVA-1.5-7B achieves Task Success Rate (TSR) $= 0.360$ at seed $= 42$ but exhibits severe output degeneration at seed $= 123$, yielding a cross-seed standard deviation of 0.129 and a 95% confidence interval of $[-0.13, 0.65]$, an interval that extends outside the parameter space $[0, 1]$, indicating that the three-seed sample is insufficient to characterise agent performance. We identify *reward density collapse* as the likely mechanism: binary reward provides zero gradient to approximately 82–88% of training steps, creating a high-variance optimisation landscape where policy trajectories diverge across seeds. A Hybrid Reward Architecture (HRA-RL) incorporating continuous automated factuality grounding, cosine similarity between sentence-embedding representations of agent outputs and ground-truth references, reduces cross-seed TSR standard deviation by $6.8\times$ (0.019 vs. 0.129) and produces a CI of $[0.281, 0.396]$, narrow enough to support meaningful UQ. We argue that dense, continuous reward signals are a practical design condition for agents intended to be used with downstream UQ methods, and that multi-seed CI reporting should be a prerequisite before applying conformal or calibration methods to RL-trained agents. Code and results are available at https://github.com/teganmosibineba/Mscproject.

[1]Department of Computer Science, Nile University of Nigeria, Abuja, Nigeria. Correspondence to: Teganmosibineba Oluwatofarati Jegede <tegan.jegede@yahoo.com>.

ICML 2026 Workshop on Statistical Frameworks for Uncertainty in Agentic Systems (AgenticUQ), Seoul, South Korea, 2026. Copyright 2026 by the author(s).

1. Introduction

Conformal prediction, calibration, and risk control have emerged as principled tools for providing coverage guarantees over language model outputs (Vovk et al., 2005; Angelopoulos & Bates, 2023; Quach et al., 2023). Applying these methods to RL-trained agents requires an assumption that is rarely stated explicitly: the agent must produce consistent behaviour across the conditions under which it is evaluated and deployed. Without this consistency, calibration sets computed from one training run may not represent the agent's behaviour in deployment, and coverage guarantees may not hold.

This paper asks: do RL-trained multimodal agents satisfy this consistency requirement? We evaluate a standard policy gradient agent trained with binary reward (Standard REINFORCE) across three random seeds on ScienceQA visual question answering using LLaVA-1.5-7B. The answer is, in at least one case, clearly not: seed $= 42$ produces a well-functioning agent (TSR $= 0.360$), while seed $= 123$ produces degenerate token repetition, outputs such as "Impalaiellement everybody everybody nobody nobody", in response to factual science questions. The resulting cross-seed confidence interval for TSR extends outside the parameter space $[0, 1]$, indicating that the sample is insufficient to characterise agent behaviour.

We make a single primary claim: sparse binary reward signals produce agent behaviour that is inconsistent enough across seeds to undermine standard UQ guarantees. Training stability and reward design are the mechanism and proposed remedy, respectively. We do not claim to solve UQ for agents; rather, we identify a concrete prerequisite, training-time behavioural consistency, that must be satisfied before UQ methods can be meaningfully applied.

Contributions:

1. Empirical documentation of cross-seed TSR variance (std $= 0.129$) and a CI extending outside $[0, 1]$ for Standard REINFORCE with binary reward, with qualitative characterisation of the seed $= 123$ degeneration failure.

2. Mechanistic hypothesis, *reward density collapse*, sup-

ported by training curve analysis comparing binary vs. continuous reward dynamics across seeds.

3. Demonstration that continuous factuality grounding (HRA-RL) reduces cross-seed std by $6.8\times$ under identical training conditions, producing a CI contained within $[0, 1]$.

4. Practical guideline: **multi-seed CI reporting** as a prerequisite before applying conformal or calibration methods to RL-trained agents.

2. Background

2.1. UQ for Language Agents and the Consistency Requirement

Conformal prediction (Vovk et al., 2005; Angelopoulos & Bates, 2023) provides distribution-free coverage guarantees under exchangeability. Applied to language agents, conformal methods have been used for selective prediction (Quach et al., 2023), abstention (Fisch et al., 2022), and risk-controlled generation (Ren et al., 2023). These methods share a structural requirement: the agent's score function must be stable across calibration and deployment conditions. When agent behaviour varies substantially across training seeds, this stability cannot be assumed. Our work empirically characterises the degree to which binary reward policy gradient training violates this stability requirement.

2.2. Reward Sparsity and Training Instability

Sparse reward signals are a known source of training instability in deep RL (Henderson et al., 2018; Skalse et al., 2022). For binary reward policy gradient training applied to language models, the dominant paradigm in RLHF (Ouyang et al., 2022) and its successors (Shao et al., 2024; DeepSeek-AI, 2025), sparsity is intrinsic: a single binary signal at the end of each episode provides no gradient to partially correct, semantically relevant, or near-threshold responses. Henderson et al. (2018) showed that reward sparsity creates seed-sensitive policy trajectories in robotics RL; we document the same phenomenon in LLM fine-tuning with direct UQ implications.

Note on terminology: we use *Standard REINFORCE* to refer to policy gradient training with a binary reward signal and EMA baseline, following Williams (1992). We do *not* implement Proximal Policy Optimisation (PPO) (Schulman et al., 2017): PPO uses a clipped surrogate objective $\min\big(r_t(\theta)\hat{A}_t,\ \mathrm{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t\big)$ with an importance sampling ratio $r_t(\theta) = \pi_\theta(a_t \mid s_t)/\pi_{\theta_{\text{old}}}(a_t \mid s_t)$. Our implementation uses the standard REINFORCE gradient $\nabla_\theta \log \pi_\theta(r \mid q, I) \cdot A_t$ with no clipping. We use the label "Standard REINFORCE" consistently throughout.

Table 1. Reward configurations. Both use identical model, dataset, and training steps. FGM uses all-MiniLM-L6-v2 cosine similarity.

Config	Reward Signal	Density	Key Property
Std REINFORCE	$R = \mathbb{1}[\mathrm{FGM}(r, g) > 0.7]$	Sparse	Zero grad $\sim$85% steps
HRA-RL	$R = W_{\text{task}} r_{\text{task}} + W_{\text{fact}} \mathrm{FGM}(r, g) - \delta \cdot \mathbb{1}[\mathrm{FGM} < \tau]$	Dense	Non-zero grad every step

HRA-RL params: $W_{\text{task}} = 1.0$, $W_{\text{fact}} = 0.5$, $\delta = 0.5$, $\tau = 0.6$.

2.3. Hallucination as Measurable Agent Risk

Hallucination, factually incorrect but fluent generation, is a primary reliability concern for knowledge-grounded agents (Ji et al., 2023; Yin et al., 2024). Cosine similarity between sentence-embedding representations of agent outputs and ground-truth references provides a continuous, automated, annotation-free correctness signal (Wang et al., 2020). We use this signal as both a training reward component and the primary evaluation metric, enabling consistency between training objective and UQ measurement.

3. Experimental Setup

3.1. Model and Training

Base model: LLaVA-1.5-7B (Liu et al., 2023) (`llava-hf/llava-1.5-7b-hf`), INT4 quantised via BitsAndBytes (Dettmers et al., 2022), $\sim$8.6 GB VRAM on a single NVIDIA T4. **Adapter:** LoRA (Hu et al., 2022), $r = 16$, $\alpha = 32$, dropout $= 0.05$, targeting `q_proj` and `v_proj` layers only; 9.96M trainable parameters (0.14% of total); vision tower fully frozen.

Algorithm: REINFORCE with EMA baseline (Williams, 1992). Loss: $L = -\log \pi_\theta(r \mid q, I) \cdot A_t$, where $A_t = R_t - b_t$ and $b_t = 0.9 \cdot b_{t-1} + 0.1 \cdot R_t$. **Optimiser:** AdamW, $\eta = 10^{-5}$, weight decay $= 0.01$, grad clip $= 0.5$. **Training:** 400 steps, 400 ScienceQA training items (Lu et al., 2022). **Evaluation:** frozen 500-item test set, identical across all agents and seeds $\{42, 123, 7\}$.

3.2. Reward Configurations

3.3. Statistical Protocol and Scope

We evaluate across seeds $\{42, 123, 7\}$, reporting mean $\pm$ standard deviation and 95% CIs via t-distribution (df $=$ 2). We acknowledge upfront that $n = 3$ seeds is a small sample: the t-distribution at df $= 2$ requires $|d| > 4.3$ for significance at $\alpha = 0.05$, so no pairwise comparison reaches statistical significance. We therefore do not claim that HRA-RL is statistically superior to Standard REINFORCE in the classical sense.

Our claim is narrower and does not require statistical significance: that Standard REINFORCE produces a CI extending outside the parameter space $[0, 1]$, which indicates that the

Table 2. Single-seed results (seed $= 42$, $n = 500$ ScienceQA). Standard REINFORCE ranks first at this seed.

Agent	TSR↑	SCS↑	HR↓	Rank
Vanilla LLaVA	0.336	0.448	0.664	3rd
RAG Baseline	0.354	0.455	0.646	2nd
Std REINFORCE	**0.360**	0.465	**0.640**	1st
HRA-RL (ours)	0.238	0.398	0.762	4th

Table 3. Multi-seed results (seeds $\{42, 123, 7\}$, $n = 500$). Standard REINFORCE CI extends outside $[0, 1]$; HRA-RL CI is $6.8\times$ narrower.

Agent	TSR Mean±Std	95% CI	Width	In $[0, 1]$?
Vanilla	0.329 ± 0.006	$[0.312, 0.347]$	0.035	Yes
RAG	0.359 ± 0.008	$[0.336, 0.382]$	0.046	Yes
Std REINFORCE	0.263 ± 0.129	$[-0.13, 0.65]$	0.783	**No**
HRA-RL	0.339 ± 0.019	$[0.281, 0.396]$	0.115	Yes

sample is insufficient to characterise agent behaviour. The seed $= 123$ degeneration is self-evident from the outputs and does not require a p-value. A power analysis indicates that $n = 8$–10 seeds would be sufficient to detect Cohen's $d \approx 0.67$ at $\alpha = 0.05$, power $= 0.80$. We present current results as a preliminary study motivating that larger evaluation.

4. Results

4.1. Single-Seed Evaluation: A Potentially Misleading Signal

Table 2 shows seed $= 42$ results. Standard REINFORCE achieves the highest TSR (0.360) and lowest HR (0.640). Without multi-seed evaluation, a practitioner might apply conformal methods to this agent under the assumption that TSR ≈ 0.36 is representative. Table 3 shows why this assumption deserves scrutiny.

4.2. Multi-Seed Analysis: CI Outside the Parameter Space

Table 3 shows multi-seed results. The Standard REINFORCE TSR CI of $[-0.13, 0.65]$ extends below zero, outside the parameter space for a rate metric. This does not mean TSR is literally negative; it means the three-seed sample is insufficient to constrain the interval to the feasible range, reflecting high cross-seed variance. The CI width of 0.783 spans 78% of the feasible range $[0, 1]$.

By contrast, HRA-RL achieves a CI of $[0.281, 0.396]$ (width $= 0.115$), contained within $[0, 1]$ and $6.8\times$ narrower. Vanilla and RAG are similarly well-constrained (widths ≤ 0.046).

Table 4. Standard REINFORCE training statistics by seed. Similar overall means mask final-window divergence.

Seed	Mean R	Final 50 R	Eval TSR	Outcome
42	$+0.143$	$+0.120$	0.360	Stable
123	$+0.113$	$+0.020$	≈ 0	Collapsed
7	$+0.150$	$+0.060$	0.189	Degraded

4.3. Qualitative Failure at Seed $= 123$

At seed $= 123$, Standard REINFORCE generates outputs consistent with reward collapse: incoherent token repetition unrelated to the input question or image. Three representative examples:

- Q: "What type of rock is dolerite?" → "…neither a mineral nor a rock everybody nobody…"
- Q: "What is the equator?" → "Equatorial Guinea lets nobody ultimately benefit FROM…"
- Q: "Animal in same genus as great blue heron?" → "Impalaiellement everybody everybody nobody nobody…"

Training reward appears normal (overall mean $= 0.11$), but the final 50-step mean drops to 0.020, revealing late-stage collapse invisible to aggregate training metrics.

4.4. Training Dynamics

Table 4 shows training statistics across seeds. Similar overall mean rewards (0.11–0.15) mask final-window divergence, suggesting the collapse is not detectable from aggregate monitoring alone.

4.5. Reward Ablation

Table 5 isolates the effect of feedback density. Configurations with continuous factuality signal (B, C, D) consistently achieve higher Semantic Consistency Scores (SCS) than pure binary reward (A, E). Config C (task binary reward plus continuous factuality, no penalty) achieves the best single-seed profile (TSR $= 0.358$, SCS $= 0.470$). The penalty term in Config D requires more training steps to converge, consistent with the density collapse mechanism.

5. Mechanism: Reward Density Collapse

We propose the following account, consistent with the training dynamics in Table 4, though we note it is a hypothesis rather than a proven causal claim.

Binary reward provides zero gradient to approximately 82–88% of training steps. The EMA baseline $b_t = 0.9 \cdot b_{t-1} +$

Table 5. Reward ablation (seed $= 42$, $n = 500$). $\star$ = best config. Dense continuous feedback consistently improves SCS.

Cfg	Signal	TSR↑	SCS↑	HR↓
A	Task binary only	0.344	0.460	0.656
B	Factuality continuous only	0.342	0.453	0.658
C*	Task + continuous fact.	**0.358**	**0.470**	**0.642**
D	Full HRA (+ penalty)	0.340	0.453	0.660
E	Std REINFORCE (binary)	0.356	0.456	0.644

$0.1 \cdot R_t$ converges slowly under sparse reward and is sensitive to the random sequence of rewarded steps early in training, which is determined by the seed. Under favourable seeds (e.g., seed $= 42$), early rewarded steps produce calibrated advantage estimates and stable policy updates. Under unfavourable seeds (e.g., seed $= 123$), zero-reward early steps cause the baseline to converge near zero; when a rewarded step eventually occurs, the large positive advantage drives the policy to overshoot onto that specific token pattern.

Continuous feedback (HRA-RL) interrupts this at the first stage: every step receives a non-zero gradient proportional to semantic alignment, keeping the baseline and advantage estimates calibrated regardless of seed. This is our proposed explanation for the $6.8\times$ variance reduction: a testable prediction that future work with larger seed samples could confirm or refute.

6. Implications for Uncertainty Quantification

6.1. Training Consistency as a UQ Prerequisite

The conformal prediction guarantee $P(Y \in C(X)) \geq 1-\alpha$ requires exchangeability between calibration and test data. For RL-trained agents, an implicit prerequisite is that the agent used at calibration time and deployment time are drawn from the same behavioural distribution. When cross-seed variance is high, as for Standard REINFORCE, this prerequisite is not guaranteed: a different training run might produce an agent with qualitatively different behaviour. We do not claim that conformal methods will necessarily fail; we claim that high variance warrants caution and motivates multi-seed validation before applying any calibration-based method.

6.2. Multi-Seed CI as a Deployment Gate

We propose a practical heuristic: before applying conformal prediction or calibration to an RL-trained agent, compute the multi-seed CI (minimum $n = 5$ seeds) for the primary performance metric. If the CI extends outside the parameter space, or if the CI width exceeds a pre-specified threshold (e.g., 0.20 for a rate metric), treat this as a signal that additional training seeds are required before UQ methods can be reliably applied.

6.3. Practical Recommendations

For UQ practitioners: Before applying conformal prediction or calibration to an RL-trained agent, conduct multi-seed evaluation (minimum $n = 5$) and verify that the CI is contained within the parameter space and has interpretable width.

For RL practitioners: Monitor the final 50-step training reward as an early instability indicator. A drop relative to the overall mean can signal reward density collapse before degenerate outputs appear at evaluation time.

For system designers: Dense, continuous feedback signals (factuality grounding, entailment scores, code execution pass rates) are preferred over binary task completion for agents intended for downstream UQ, as they produce more consistent behaviour across training conditions.

6.4. Limitations

This study has four acknowledged limitations. First, $n = 3$ seeds ($\mathrm{df} = 2$) limits statistical power; no pairwise comparison is significant at $\alpha = 0.05$, and effect sizes (Cohen's $d = 0.54$–0.75) require $n = 8$–10 for significance. Second, the reward density collapse mechanism is a hypothesis supported by training curve analysis, not a proven causal claim. Third, all experiments use a single model (LLaVA-1.5-7B), benchmark (ScienceQA), and training budget (400 steps). Fourth, the connection between reward density and conformal coverage validity is stated as a practical argument rather than a formal theorem.

7. Conclusion

We present preliminary evidence that sparse binary reward policy gradient training produces agent behaviour inconsistent enough across training seeds to undermine standard UQ guarantees: Standard REINFORCE yields a cross-seed TSR confidence interval extending outside $[0, 1]$, while a hybrid reward incorporating continuous factuality grounding produces a CI that is $6.8\times$ narrower and contained within $[0, 1]$. We propose reward density collapse as the likely mechanism and multi-seed CI reporting as a practical deployment gate before applying conformal or calibration methods to RL-trained agents.

We emphasise that these are preliminary findings based on $n = 3$ seeds and a single model–benchmark combination. The central contribution is not a definitive empirical result but a concrete, testable claim: that feedback signal density is a prerequisite for agent consistency, and agent consistency is a prerequisite for reliable UQ, with supporting evidence that motivates larger-scale study.

Code and results are available at https://github.com/teganmosibineba/Mscproject.

Acknowledgements

We thank the anonymous AgenticUQ reviewers for constructive feedback that improved this paper. This work was conducted as part of the author's MSc thesis at Nile University of Nigeria. All experiments were run on freely available NVIDIA T4 GPU resources.

References

Angelopoulos, A. N. and Bates, S. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023.

DeepSeek-AI. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv:2501.12948*, 2025.

Dettmers, T., Lewis, M., Belkada, Y., and Zettlemoyer, L. LLM.int8(): 8-bit matrix multiplication for transformers at scale. In *NeurIPS*, 2022.

Fisch, A., Schuster, T., Jaakkola, T., and Barzilay, R. Conformal prediction sets with limited false positives. In *ICML*, 2022.

Henderson, P., Islam, R., Bachman, P., Pineau, J., Precup, D., and Meger, D. Deep reinforcement learning that matters. In *AAAI*, 2018.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., and Fung, P. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12), 2023.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning. In *NeurIPS*, 2023.

Lu, P., Mishra, S., Xia, T., Qiu, L., Chang, K.-W., Zhu, S.-C., Tafjord, O., Clark, P., and Kalyan, A. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *NeurIPS*, 2022.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., and Lowe, R. Training language models to follow instructions with human feedback. In *NeurIPS*, 2022.

Quach, V., Fisch, A., Schuster, T., Yala, A., Sohn, J. H., Jaakkola, T. S., and Barzilay, R. Conformal language modeling. In *ICLR*, 2023.

Ren, A., Dixit, A., Bodrova, A., Singh, S., Tu, S., Brown, N., Xu, P., Takayama, L., Xia, F., Varley, J., Xu, Z., Sadigh, D., Zeng, A., and Majumdar, A. Robots that ask for help: Uncertainty alignment for large language model planners. In *CoRL*, 2023.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. *arXiv:1707.06347*, 2017.

Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y. K., Wu, Y., and Guo, D. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv:2402.03300*, 2024.

Skalse, J., Howe, N., Krasheninnikov, D., and Krueger, D. Defining and characterizing reward gaming. In *NeurIPS*, 2022.

Vovk, V., Gammerman, A., and Shafer, G. *Algorithmic Learning in a Random World*. Springer, 2005.

Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., and Zhou, M. MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *NeurIPS*, 2020.

Williams, R. J. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8:229–256, 1992.

Yin, S., Fu, C., Zhao, S., Xu, T., Wang, H., Sui, D., Shen, Y., Li, K., Sun, X., and Chen, E. Woodpecker: Hallucination correction for multimodal large language models. *Science China Information Sciences*, 2024.