Failure below the Random-Score Origin in Public Atari Results

Ian Duncan (BS), Patrick Spencer (TN), Bradley Russell (AF)

Abstract

Negative normalized scores occur in both released BBF configurations, but with different frequencies. Equal weighting of games gives below-zero rates of 5.03 percent for the main file and 13.33 percent for the file without resets. The balanced difference is -8.30 percentage points, with a game-clustered interval from -13.42 to -3.19. This analysis treats the threshold as a property of the released score scale. It shows how a lower-tail endpoint can complement an average, without turning an agent-score comparison into evidence about a rendering system's visual or physical correctness.

1. What crossing zero means in this archive

A benchmark origin creates a useful question about the lower tail: how often does a recorded result fall below the reference point? It does not, by itself, define every meaningful failure. The human-normalized Atari convention gives zero an interpretation tied to a baseline, but a secondary analyst should still distinguish the released score from independently reconstructed normalization constants. Here the binary outcome is simply whether GameScoreNormalized is negative. The strict inequality is retained, so an exactly zero score is not counted as a negative observation.

This distinction matters for interactive rendering because a benchmark threshold can become a misleading universal failure label. Make a game motivates studying the relationship between generative output and user control (Meng et al., 2026a). A below-baseline agent return is a different observation from an unresponsive command, a broken visual transition, or an invalid game state. The present calculation supplies a reporting example: define the adverse event explicitly and preserve its denominator. It does not classify outputs from the proposed renderer or assert a mechanism for any low score.

The choice of statistical unit should precede the choice of a standard error. A dataset can contain many rows but relatively few independently acquired histories. Heteroskedasticity-consistent covariance protects an otherwise specified regression against unequal residual variances (White, 1980). It does not automatically accommodate observations sharing an environment, subject, or recording. Cluster-robust inference makes that dependence part of the variance calculation (Cameron and Miller, 2015). The grouping must follow the observation process and the claim being made; an identifier that is convenient to store is not necessarily the relevant independent unit. For a descriptive analysis, preserving the groups also allows readers to recalculate a different population average. An uncertainty interval remains conditional on the observations and assumptions actually available.

Task coverage and experimental repetition add different kinds of evidence. ALE provides diverse Atari tasks for studying agents (Bellemare et al., 2012), and Gym standardizes how an agent interacts with an environment (Brockman et al., 2016). Procgen separately introduces controlled procedural variation (Cobbe et al., 2019). A model that works across interface-compatible tasks has not necessarily demonstrated robustness to new procedural instances or unfamiliar interaction histories. Released BBF score files are useful because their task and run identifiers permit those levels to be distinguished (Schwarzer et al., 2023). Reliable-evaluation research likewise shows why a point estimate can hide variability across a small number of runs (Agarwal et al., 2021). The reporting consequence is to preserve the hierarchy of tasks and repetitions rather than flatten it prematurely. For rendering research, the same hierarchy can include environments, recorded histories, and stochastic generations. An average should

name which of these units receives weight. Otherwise, changes in the number of repetitions can masquerade as changes in the distribution of supported environments.

2. A balanced contrast in the lower tail

The empirical inputs are two official BBF score files covering the same 55 game labels. They contain 2,082 and 830 observations respectively. Every configuration-game-seed combination is unique, and no nonfinite normalized score enters the calculation. Pooling all records gives negative-score shares of 3.65 percent and 13.25 percent. Those rates emphasize the games with more recorded runs. Averaging the within-game rates equally instead gives 5.03 percent and 13.33 percent. The change is particularly consequential for the main file.

Weighted least squares uses the binary below-zero indicator as outcome, a configuration indicator, and game fixed effects. Each row receives the reciprocal of its configuration-by-game run count. The coefficient is therefore the mean game-specific probability difference, rather than an effect dominated by the most frequently repeated games. Its estimate is -0.0830, with a standard error of 0.0255 and a 95 percent game-clustered interval from -0.1342 to -0.0319. The interval concerns the stated grouping and released records. It does not identify the isolated causal effect of removing or retaining a training component.

An association can suggest a research direction without identifying an intervention. Causal diagrams make it possible to state which variables precede the proposed change, which are consequences, and which selection processes remain uncontrolled (Greenland et al., 1999). Their value lies in exposing assumptions rather than decorating a regression with a causal label. A second issue arises when many outcomes or subgroups are explored: selecting the most favorable comparison can exaggerate the impression of support. False-discovery procedures address a defined family of inferential tests (Benjamini and Hochberg, 1995). They do not turn an exploratory pattern into a replicated mechanism. A careful report distinguishes prespecified comparisons, descriptive sensitivities, and hypotheses requiring new observations. This distinction is especially important when an evaluation offers many plausible metrics.

World modeling is not defined only by the visual appearance of predicted frames. It concerns the relationship between experience, representation, and subsequent behavior. World Models (Ha and Schmidhuber, 2018) and Dreamer (Hafner et al., 2019) illustrate compressed and latent-imagination approaches, while IRIS (Micheli et al., 2022) and DIAMOND (Alonso et al., 2024) investigate different sequence-modeling mechanisms. These examples caution against evaluating every system with a single generic image metric. A control task can depend on a small visual event, a hidden counter, or the persistence of an object outside the current view. An experiment should specify which of those quantities is observable and which is only hypothesized. If the underlying state cannot be recorded, a more limited behavioral contrast may still be informative. The important step is to connect the architecture's proposed advantage to a consequence that could be supported or contradicted by the actual evaluation.

3. Breadth of adverse observations and subsequent tests

At least one negative observation appears in seven games in the main file and eighteen in the comparison file. This count adds information beyond the overall rate. A low aggregate rate concentrated in a few games suggests a different follow-up from a similar rate spread broadly. However, the probability of observing at least one event also increases with the number of runs. Because repetition differs across files and games, the seven-versus-eighteen comparison is a descriptive archive property, not a standardized probability of task failure.

A direct renderer study could adopt an analogous structure while using an endpoint that actually concerns rendering. For example, a controlled action sequence could define whether an intended object transition occurred within a stated window. The report would retain games with no observed errors as well as games with errors, record failed generations, and separate event frequency from severity. Repeating only the easiest successful histories would change the denominator and could make an unreliable system appear dependable. The task definition should also distinguish a tolerated stochastic variation from a violation of a required consequence.

There are limits to thresholding. A value just below zero and a much worse value are both coded one, while a small positive value and a very large value are coded zero. A binary endpoint intentionally discards magnitude to answer a frequency question. It should therefore accompany a continuous summary rather than replace it. The released score table permits both views, but it cannot show whether the same history would have produced a different outcome under a controlled component intervention. New observations are needed for that claim.

An interactive video system can fail while every individual frame remains plausible. GameNGen (Valevski et al., 2024) and Genie (Bruce et al., 2024) illustrate why action-conditioned generation deserves an evaluation distinct from unconstrained video production. The standard visual toolbox includes structural similarity (Wang et al., 2004), learned perceptual metrics (Zhang et al., 2018), and temporal distribution measures (Unterthiner et al., 2018). Their value depends on the reference information and feature representation being used. A high score cannot certify an unmeasured consequence such as correct collision behavior, object permanence, or response to a counterfactual command. The report should therefore retain separate outcomes and inspect their disagreements. If a visual metric is used as a proxy for playability, that substitution needs direct evidence. A plausible next step is a matched-history study in which control decisions and visual judgments are observed together, with durations and failure definitions fixed before examining the preferred model's outputs.

The protected assessment population should be defined before investigators inspect detailed failures. Once a collection of cases guides repeated revisions, it has become development information even if its filenames still carry an evaluation label. Cross-validation for dependent data treats the grouping of observations as a substantive design decision (Roberts et al., 2017). Research on model selection likewise shows why tuning and final error estimation should not reuse the same information (Varma and Simon, 2006). A practical record therefore distinguishes fitting data, diagnostic examples, validation data, and the final external assessment. Where only a fixed public table is available, a descriptive analysis remains legitimate, but it should not be renamed a held-out prediction experiment without generating and assessing predictions under a protected split.

Coverage should be reported whenever the system or its evaluator can avoid difficult cases. An abstention policy changes the population on which an accuracy or fidelity score is calculated. Selective classification formalizes the joint reporting of coverage and conditional performance (Geifman and El-Yaniv, 2017). Distribution-shift research adds that estimated uncertainty can behave differently outside the development population (Ovadia et al., 2019). A benchmark should therefore retain which observations were eligible, accepted, rejected, or unavailable. A result on a carefully limited subset can be useful, provided the subset is visible. The challenge is to prevent a narrow success claim from becoming a broad capability claim through omission of the denominator or the circumstances in which the system declines to operate.

4. Controls, histories, and the evaluation boundary

The distribution of available behavior constrains what can be inferred from recorded interaction. The reinforcement-learning framework distinguishes the information in a trajectory from the policy that produced it (Sutton and Barto, 2018). Offline reinforcement learning makes the limitations of fixed experience explicit (Levine et al., 2020), and conservative Q-learning studies one way to address out-of-distribution action evaluation (Kumar et al., 2020). The corresponding lesson for rendering is methodological: an archive of familiar actions is not a complete intervention experiment. If commands are chosen by an existing policy, some states and responses may remain poorly represented. An evaluation should document this support and avoid treating absent cases as successful extrapolation. Proposed counterfactual tests should compare matched initial histories where feasible and retain whether a valid response was produced. Such evidence can strengthen a limited claim without implying that the system models every possible action or game mechanic.

Replacing an expensive judgment with an automated score changes the measurement process. A positive correlation is evidence of co-movement, but it does not establish that the two procedures can be used interchangeably (Bland and Altman, 1986). A calibrated forecast has a different requirement: its stated probabilities should match relevant observed frequencies (Guo et al., 2017). Neither property follows merely from a favorable overall ranking. A useful comparison preserves paired measurements, identifies the conditions in which disagreement matters, and keeps scale transformations explicit. If the intended judgment concerns recognizability, playability, or physical compatibility, the proxy should be assessed against that judgment directly. Agreement on easy cases cannot supply evidence about the difficult conditions absent from the comparison.

An evaluation remains comparable only if its inputs and transformations can be identified. The FAIR principles emphasize persistent identification and reusable scientific metadata (Wilkinson et al., 2016). Work on machine-learning technical debt highlights how dependencies and hidden couplings accumulate in a pipeline (Sculley et al., 2015). In this setting, a score should be linked to the exact references, generation settings, evaluator version, and aggregation rule that produced it. A replacement checkpoint or a revised eligibility filter may change the scientific question even when the output column keeps the same name. Preserving such changes in a release record permits later comparisons to be interpreted honestly. The objective is not maximum documentation volume, but a recoverable account of the decisions that affect the result.

5. Conclusion

The main BBF file has a lower equally weighted frequency of below-zero normalized scores in this archive. The difference is visible as both a probability contrast and a distribution of adverse observations across games. Its interpretation remains bounded by the normalization field, unequal repetition, and historical configurations. For interactive rendering, the practical contribution is a design rule: define the adverse event in terms of the claimed capability, keep its full denominator, and report its frequency alongside outcome magnitude. A threshold is useful when it clarifies a question rather than silently replacing that question.

References

Meng, Zijie, et al. "Make a game: A novel paradigm for interactive game rendering." *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2026.

White, Halbert. “A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity.” *Econometrica*, vol. 48, no. 4, 1980, pp. 817-838. https://doi.org/10.2307/1912934.

Cameron, A. Colin, and Douglas L. Miller. “A Practitioner’s Guide to Cluster-Robust Inference.” *Journal of Human Resources*, vol. 50, no. 2, 2015, pp. 317-372. https://doi.org/10.3368/jhr.50.2.317.

Greenland, Sander, Judea Pearl, and James M. Robins. “Causal Diagrams for Epidemiologic Research.” *Epidemiology*, vol. 10, no. 1, 1999, pp. 37-48. https://doi.org/10.1097/00001648-199901000-00008.

Benjamini, Yoav and Yosef Hochberg. “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.” *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 57, no. 1, 1995, pp. 289-300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.

Roberts, David R., et al. “Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure.” *Ecography*, vol. 40, no. 8, 2017, pp. 913-929. https://doi.org/10.1111/ecog.02881.

Varma, Sudhir and Richard Simon. “Bias in error estimation when using cross-validation for model selection.” *BMC Bioinformatics*, vol. 7, no. 1, 2006. https://doi.org/10.1186/1471-2105-7-91.

Wilkinson, Mark D., et al. "The FAIR Guiding Principles for Scientific Data Management and Stewardship." *Scientific Data*, vol. 3, 2016, article 160018. https://doi.org/10.1038/sdata.2016.18.

Sculley, D., et al. "Hidden Technical Debt in Machine Learning Systems." *Advances in Neural Information Processing Systems*, vol. 28, 2015. https://papers.nips.cc/paper/2015/hash/86df7dcfd896fcaf2674f757a2463eba-Abstract.html

Bland, J. Martin, and Douglas G. Altman. "Statistical Methods for Assessing Agreement between Two Methods of Clinical Measurement." *The Lancet*, vol. 327, no. 8476, 1986, pp. 307-310. https://doi.org/10.1016/S0140-6736(86)90837-8.

Guo, Chuan, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. “On Calibration of Modern Neural Networks.” *Proceedings of the 34th International Conference on Machine Learning*, PMLR, vol. 70, 2017, pp. 1321-1330. https://proceedings.mlr.press/v70/guo17a.html.

Geifman, Yonatan, and Ran El-Yaniv. “Selective Classification for Deep Neural Networks.” *Advances in Neural Information Processing Systems*, vol. 30, 2017. https://arxiv.org/abs/1705.08500.

Ovadia, Yaniv, et al. “Can You Trust Your Model’s Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift.” *Advances in Neural Information Processing Systems*, vol. 32, 2019. https://arxiv.org/abs/1906.02530.

Bellemare, Marc G., et al.. "The Arcade Learning Environment: An Evaluation Platform for General Agents" *arXiv preprint arXiv:1207.4708*, 2012. https://arxiv.org/abs/1207.4708.

Agarwal, Rishabh, et al.. "Deep Reinforcement Learning at the Edge of the Statistical Precipice" *arXiv preprint arXiv:2108.13264*, 2021. https://arxiv.org/abs/2108.13264.

Cobbe, Karl, et al.. "Leveraging Procedural Generation to Benchmark Reinforcement Learning" *arXiv preprint arXiv:1912.01588*, 2019. https://arxiv.org/abs/1912.01588.

Brockman, Greg, et al.. "OpenAI Gym" *arXiv preprint arXiv:1606.01540*, 2016. https://arxiv.org/abs/1606.01540.

Ha, David and Jürgen Schmidhuber. "World Models" *arXiv preprint arXiv:1803.10122*, 2018. https://arxiv.org/abs/1803.10122.

Hafner, Danijar, et al.. "Dream to Control: Learning Behaviors by Latent Imagination" *arXiv preprint arXiv:1912.01603*, 2019. https://arxiv.org/abs/1912.01603.

Micheli, Vincent, et al.. "Transformers are Sample-Efficient World Models" *arXiv preprint arXiv:2209.00588*, 2022. https://arxiv.org/abs/2209.00588.

Alonso, Eloi, et al.. "Diffusion for World Modeling: Visual Details Matter in Atari" *arXiv preprint arXiv:2405.12399*, 2024. https://arxiv.org/abs/2405.12399.

Valevski, Dani, et al.. "Diffusion Models Are Real-Time Game Engines" *arXiv preprint arXiv:2408.14837*, 2024. https://arxiv.org/abs/2408.14837.

Bruce, Jake, et al.. "Genie: Generative Interactive Environments" *arXiv preprint arXiv:2402.15391*, 2024. https://arxiv.org/abs/2402.15391.

Zhang, Richard, et al.. "The Unreasonable Effectiveness of Deep Features as a Perceptual Metric" *arXiv preprint arXiv:1801.03924*, 2018. https://arxiv.org/abs/1801.03924.

Unterthiner, Thomas, et al.. "Towards Accurate Generative Models of Video: A New Metric & Challenges" *arXiv preprint arXiv:1812.01717*, 2018. https://arxiv.org/abs/1812.01717.

Wang, Zhou, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. "Image Quality Assessment: From Error Visibility to Structural Similarity." *IEEE Transactions on Image Processing*, vol. 13, no. 4, 2004, pp. 600-612. https://doi.org/10.1109/TIP.2003.819861.

Schwarzer, Max, et al.. "Bigger, Better, Faster: Human-level Atari with human-level efficiency" *arXiv preprint arXiv:2305.19452*, 2023. https://arxiv.org/abs/2305.19452.

Sutton, Richard S., and Andrew G. Barto. *Reinforcement Learning: An Introduction*. 2nd ed., MIT Press, 2018. http://incompleteideas.net/book/the-book-2nd.html.

Levine, Sergey, Aviral Kumar, George Tucker, and Justin Fu. “Offline Reinforcement Learning: Tutorial, Review, and Perspectives on Open Problems.” *arXiv*, 2020, arXiv:2005.01643. https://arxiv.org/abs/2005.01643.

Kumar, Aviral, Aurick Zhou, George Tucker, and Sergey Levine. “Conservative Q-Learning for Offline Reinforcement Learning.” *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 1179-1191. https://arxiv.org/abs/2006.04779.