What Seed Labels Can and Cannot Establish in Released Agent Results

Connor Brooks (TV), Nathan Simmons (MZ), Dennis Powell (UY)

Abstract

All 2,912 configuration-game-seed combinations in the two public BBF files are unique. The main and comparison files also share no seed labels. These checks support an unambiguous record count, but they do not certify independent random streams, matched evaluation histories, or a paired component experiment. The article distinguishes identifier integrity from experimental comparability. For interactive rendering, repeatable histories, randomness controls, and output identifiers should be recorded separately so that a label named seed does not carry scientific assumptions that the release never establishes.

1. A reproducibility field is not an experimental design

A seed field is often interpreted as a compact guarantee of repeatability. In practice, its meaning depends on which random processes it controls, which software consumes it, and which other conditions are held fixed. A numerical identifier can be unique even when experiments share substantial data or evaluation histories. Conversely, the same integer can be used in different environments without implying a shared realization of every stochastic process. Identifier checks are useful, but their scientific meaning must follow the documented implementation.

Make a game raises a related question for generated interaction (Meng et al., 2026a). A replay might depend on an initial state, action history, model checkpoint, sampling configuration, and several sources of randomness. The available public description does not establish a complete replay protocol, so one is not inferred here. Instead, an audit of released BBF identifiers provides a concrete example of what can be checked directly and which assumptions remain external to a score table.

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.

Common evaluation platforms make comparisons convenient, but convenience should not determine the scientific estimand. ALE organizes many distinct games under a reusable platform (Bellemare et al., 2012), while Gym supplies a widely used interaction abstraction (Brockman et al., 2016). Procgen emphasizes that procedural instances can themselves define an important assessment boundary (Cobbe et al., 2019). BBF's public result tables give a concrete example in which tasks and runs can be counted separately (Schwarzer et al., 2023). The broader statistical lesson is that aggregate performance depends on uncertainty, task heterogeneity, and the chosen summary (Agarwal et al., 2021). A rendering evaluation should therefore retain environment-level outcomes before constructing a total. Matching the number of generated frames does not necessarily match the diversity of actions or histories. Nor does a large number of repeated observations make the benchmark representative of every interactive setting. The relevant claim is determined by which tasks and variations were observed, under what controls, and with what weighting.

2. The record-level checks actually performed

The main official CSV contains 2,082 observations, while the file without resets contains 830. Combining configuration, game, and seed produces 2,912 distinct keys, exactly matching the total row count. Within the main file there are 2,082 unique seed values, and the comparison contains 830. The two sets of seed labels have no intersection. There are therefore no shared game-seed keys across configurations either. The audit does not remove or merge any records through this key definition.

These properties mean a reader can identify each archived row without ambiguity using the recorded fields. They do not establish how many random initializations, environment episodes, or evaluation samples contributed to each score. GameScoreAverage is a released summary rather than a trajectory log. The file does not provide a complete enumeration of the lower-level observations behind it. Consequently, the unique-row count should not be translated into a larger unobserved episode count or a claim of statistically independent experimental units.

A proposed explanation should be distinguishable from alternative explanations using observations the study can actually collect. Diagramming the relationships among selection, exposure, measurement, and outcome makes that requirement explicit (Greenland et al., 1999). The diagram does not establish its own correctness, but it prevents adjustment choices from remaining invisible. The same transparency is needed when investigators inspect many endpoints before choosing a narrative. Controlling false discovery concerns the behavior of a specified testing family (Benjamini and Hochberg, 1995); it is not a general certificate of scientific validity. Primary endpoints, secondary descriptions, and future mechanism tests should be named separately. Readers can then assess whether an attractive pattern was predicted, explored, or merely noticed after the available comparisons were examined.

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. Why the files should not be paired through their row order

There are different numbers of records under the two configurations and no common seed labels. Pairing their first rows, sorting within games and matching positions, or assigning an artificial common run index would manufacture a pairing not supplied by the release. A game-level comparison remains possible because both files include the same 55 task labels. A within-seed paired comparison does not follow from those labels. The distinction affects both the interpretation and the uncertainty calculation of any configuration contrast.

Nor would common seed integers alone establish a valid pairing. A controlled paired design should explain what is shared: initial observations, stochastic environment events, data order, or another source of variation. If a configuration changes how randomness is consumed, the same starting integer can produce different subsequent random events. The resulting experiment may still be useful, but its pairing

needs a justified interpretation. Reproducibility requires a recoverable execution record rather than an unsupported inference from a column name.

A rendering release could record a hierarchy of identifiers: task definition, initial-state snapshot, action history, model version, generation invocation, and scoring invocation. The randomness configuration would be linked to the invocation it actually governs. Such a record would allow users to distinguish a repeated history with new sampling noise from a newly selected history. It would also identify whether the evaluator itself is stochastic. These are proposed records, not properties established about Make a game's unpublished implementation or about the BBF source experiments beyond the fields examined here.

The audit also clarifies a useful limitation of metadata. Hashing an input or assigning a unique identifier can show that the same bytes were reused. It cannot by itself show that an outcome measurement is valid, a population is representative, or a model has learned the intended dynamics. Those claims require their own observations. A careful report can value identifier integrity without allowing it to substitute for experimental control or scientific validation.

The ability to accept commands changes the interpretation of a generated sequence. GameNGen's interactive setting (Valevski et al., 2024) and Genie's action-controllable worlds (Bruce et al., 2024) make this distinction visible. A command-response pair needs to be judged for its consequence as well as its visual continuity. Image structure (Wang et al., 2004), learned perceptual correspondence (Zhang et al., 2018), and video-distribution similarity (Unterthiner et al., 2018) supply useful but incomplete observations. The evaluator should explain whether a reference trajectory exists and whether the same initial condition can be replayed. Without those records, a visually different outcome might reflect a valid stochastic alternative or a failure to obey control. The experiment must distinguish those interpretations through its task specification. Matching output length alone is insufficient, because a longer sequence can contain more opportunities for a delayed or persistent control error. The reported endpoint should remain connected to the history that made the interaction meaningful.

A validation split defines which form of transfer has been observed. Related observations can share subjects, scenes, prompts, or acquisition conditions, so randomly splitting rows may leave important information on both sides of the boundary. Structured cross-validation addresses these dependencies (Roberts et al., 2017). Model selection introduces another boundary: choosing a representation, threshold, or tuning parameter using the final assessment data makes the resulting error estimate optimistic (Varma and Simon, 2006). The appropriate held-out unit follows the intended future use. Generalization to a new frame, a new trajectory, and a new subject are not equivalent claims. A study should record those units before fitting or choosing a scoring rule, and preserve an assessment set that has not become part of iterative development.

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. From a task contrast to a controlled study

An action-dependent observation has meaning only within a specified interaction process. Reinforcement learning provides a formal distinction among states, actions, rewards, and policies (Sutton and Barto, 2018). Offline learning further emphasizes that a fixed collection of histories has limited behavioral support (Levine et al., 2020). Conservative value learning addresses a particular problem arising when actions are evaluated beyond that support (Kumar et al., 2020). These sources do not validate a rendering architecture, but they clarify why familiar trajectories can be an incomplete test. A renderer may look reliable when replaying common histories while responding poorly to interventions that the archive rarely contains. A future evaluation should state which commands and transitions are supported by its data, which are deliberately held out, and which remain untested. Reweighting observed histories can alter their importance but cannot create an observation of a missing control consequence. That boundary should stay visible in the interpretation of any broad capability claim.

An automated measurement should not inherit the authority of its training labels without a direct assessment of use. The relevant checks include systematic disagreement with the intended procedure (Bland and Altman, 1986) and calibration of any probability claims (Guo et al., 2017). Ranking performance, absolute accuracy, and confidence reliability are distinct outcomes. A study can improve one while leaving the others unchanged. Keeping the measurements separate permits examination of those tradeoffs and avoids implying that one favorable statistic establishes all aspects of validity. When independent judgments are unavailable, the limitation should remain part of the result. The absence of a target observation cannot be repaired by combining several correlated proxies into a more elaborate total score.

A scientific release should make its most consequential choices recoverable. Persistent identifiers and explicit metadata support reuse of observations and results (Wilkinson et al., 2016), while attention to dependencies helps prevent an evaluation from changing unnoticed as software evolves (Sculley et al., 2015). Relevant choices include the source version, preprocessing, accepted input set, metric implementation, and weighting convention. The record should also separate an observed output from a proposed future test. Saving that distinction prevents a plan from later being mistaken for a completed experiment. A small collection of well-linked records can be more useful than a large folder of unexplained scores. The important property is whether an independent reader can reconstruct what was measured and why its reported number has that meaning.

5. Conclusion

The BBF files support a clean count of unique released records and provide no recorded seed pairing across configurations. That is enough to prevent accidental duplicate counting and artificial row matching. It is not enough to certify independent random streams or an isolated component effect. Interactive rendering studies should treat seed information as one documented part of an execution record, preserve the histories and states needed for meaningful replay, and state explicitly which aspects of an experiment are actually shared.

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.