Influential Games and Stability of a Balanced Performance Contrast

Daniel Wright (CL), Connor Carter (BG), Michael Hoffman (IR)

Abstract

Removing one game at a time changes the balanced transformed BBF contrast from 0.466 to values between 0.422 and 0.477. Boxing contributes the largest positive task difference, while Bank Heist has the most negative one. No single omission reverses the average comparison in the released files. This is an influence analysis, not a confidence interval or a new experiment. It illustrates how an interactive evaluation can retain unusual environments, expose their contribution to an aggregate, and avoid confusing stability within a fixed benchmark with transfer beyond it.

A benchmark can be stable and still incomplete

A result that disappears when one task is removed deserves a different interpretation from a result that persists across every such omission. Neither is automatically correct or incorrect. The first may identify a genuine specialization; the second may demonstrate breadth within a chosen collection. Both remain conditional on which tasks entered the collection. Influence analysis addresses the contribution of observed tasks, not the relevance of tasks that were never observed.

Make a game motivates examining the range of environments in which a generative interface responds usefully to control (Meng et al., 2026a). Its interactive framing makes task-level influence a relevant evaluation question. No output from the proposed renderer is analyzed here, and unavailable publisher details are not reconstructed. The actual numeric exercise uses two released BBF score files to ask whether their equally weighted game comparison is driven by a single represented task. The separation between that observation and a future rendering test is maintained throughout.

A leave-one-game-out calculation

The archived records cover 55 games under both configurations. Within each configuration-game group, the supplied normalized score is transformed using the inverse hyperbolic sine and then averaged. The difference between configuration means gives one contrast per game. Averaging those 55 contrasts equally produces 0.4662. This transformation preserves negative and zero observations and compresses large positive outcomes; the final number is expressed in transformed score units.

The influence calculation removes each game in turn and averages the remaining 54 contrasts. No runs are regenerated and no model is refitted to an emulator. The minimum reduced-set mean is 0.4216 and the maximum is 0.4765. Thus every single-game omission leaves the average positive. These limits are the range of 55 related sensitivity calculations. They are not a percentile interval, a sampling confidence interval, or a probability statement about future games. They also do not treat the 55 reduced collections as independent replications.

The role of the strongest and weakest tasks

Boxing has the largest positive transformed contrast, 2.8778. Removing it produces the lowest reduced-set average. Bank Heist has the most negative contrast, -0.0904, and removing it produces the largest average. The different sizes show that the upper contribution is more influential than the most negative contribution on this scale. That is consistent with a right-skewed distribution of task differences. It does not identify why either game behaves that way or establish that a particular mechanic explains the result.

The practical interpretation is twofold. The main comparison is not solely a consequence of one positive task, because it remains positive after that task is excluded. At the same time, the size of the reported average is measurably affected by the strongest contrast. A report can state both facts without choosing between an overconfident universal claim and a claim that the benchmark is meaningless. Retaining the full task table allows the specialization to be studied rather than removed from view.

For interactive rendering, a similar analysis could be applied to a declared environment-level endpoint, such as the probability of satisfying an action-response requirement. The endpoint should be meaningful before influence is calculated. Removing a difficult environment merely because it lowers a preferred model's mean would change the study's question. A legitimate sensitivity analysis keeps that environment in the primary record and shows what its presence contributes. The resulting explanation should distinguish a change in the average from evidence about a specific failure mechanism.

There is also a limit to single-task omission. Two tasks may jointly dominate a result, and related tasks may share a weakness that becomes visible only when an entire environment family is considered. An investigator should justify any grouped omission through a prespecified taxonomy or scientific question rather than search through subsets for a desired conclusion. The available archive supports additional descriptive work, but an exhaustive subset search would create another model-selection problem. A useful next experiment would test unfamiliar tasks chosen to probe the hypothesized specialization.

An influence plot for a finite game collection

A near-constant leave-one-out series can mask a correlated block of several games with similar dynamics. The issue becomes concrete in a hypothetical assessment. If dropping one unusually large positive contrast changes a benchmark headline, a reader needs to see which title carries that leverage. Use omitted game to connect all-game estimator and leave-one-game-out estimator to the range of contrasts under single-game omission. For a new evaluation, define families before seeing results and extend the omission display to those families.

In a leave-one-out display, the omitted title should be named and its original contrast should appear next to the recomputed estimate. If a family of similar titles contributes several related favorable outcomes, individually small changes after omission can still conceal family-level leverage. Define the candidate families from mechanics or documented lineage before reading their scores. Then inspect joint omissions. This separates a robustness result for the finite release from a hypothesis about transfer to a different population of games.

Before estimating the range of contrasts under single-game omission, fix the conditions under which a case counts as eligible and decide how exclusions will appear in the summary. For the omitted game, log interruptions and timeouts separately so readers can see whether a response was observed.

Record attempts in their original order and store state, input and response separately for omitted game. To evaluate the range of contrasts under single-game omission, keep paired observations together during resampling, since independent resampling would lose their shared source of variation. For leave-one-game-out estimator, compare conclusions after varying the permitted response window within a declared range.

After evaluating leave-one-game-out estimator, list the cases whose result changes when the measurement definition is adjusted within a declared range. For all-game estimator, report whether agreement concerns game progress, visible state or the timing of the response. Document the sampling boundary for the range of contrasts under single-game omission and reserve a new cohort of cases for a distinct replication.

Analysis and interpretation

Because several changes coexist in one system, isolating a component requires an intervention or a comparison with the other choices held fixed. A curriculum and a reusable skill library can change what a system attempts, so log the attempted tasks alongside completed ones. For implementation, a methodological precedent combines an automatic curriculum with a reusable Minecraft skill library (Wang et al., 2023). One published study combines several extensions of deep Q-learning in a common Atari agent (Hessel et al., 2017).

Matching can improve comparability on recorded attributes; an overlap display should identify which observations are lost before the outcome contrast is computed. For a prospective comparison, earlier research uses causal diagrams to articulate assumptions about adjustment and selection (Greenland et al., 1999). One published study reviews matching as a design step for comparing observational units (Stuart, 2010). A proposed causal diagram can distinguish decisions made before a run from variables created by the system and observed only after it responds.

Another study examines false discovery rate control when tests are dependent (Benjamini and Yekutieli, 2001). The archived rows support comparisons within recorded game and configuration cells, with substantial differences in how many seeds contribute observations to each cell. When diagnostic endpoints are dependent, the correction should reflect that dependence and use a defined family rather than whichever comparisons looked notable. In a controlled test, related work investigates data-efficient Atari agents and supplies the comparison's released score archive (Schwarzer et al., 2023).

Earlier research describes estimation and implementation of linear mixed-effects models (Bates et al., 2015). In a controlled test, reusing a validation split for repeated configuration selection can make the final error estimate optimistic; retain a separate assessment for the chosen rule. Mixed-effects models account for repeated observations of the same unit, while the choice of grouping still needs a design rationale. A relevant contribution shows bias from reusing validation for selection and error estimation (Varma and Simon, 2006).

One approach investigates reliable comparisons of reinforcement learning under limited runs (Agarwal et al., 2021). In a controlled test, unequal residual variance changes uncertainty even under a common mean model, which matters when a few conditions are much noisier than others. Uncertainty intervals and performance profiles help show whether an apparent advantage persists across titles with very different return distributions. Earlier research allows regression covariance estimates to account for unequal error variances (White, 1980).

Seed choices and implementation details can create measurable variation, making independent repetitions and version records part of the evidence rather than ancillary information. In an evaluation, serial dependence makes neighboring measurements count differently for uncertainty, even if the average contrast itself is calculated using all observed time points. Another study allows covariance estimates to account for heteroskedasticity and serial dependence (Newey and West, 1987). Earlier research examines how experimental choices affect reproducibility in deep reinforcement learning (Henderson et al., 2017).

Samples and evidence

Related work organizes open-ended Minecraft tasks and external knowledge resources (Fan et al., 2022). An existing method connects validation splits to temporal, spatial and hierarchical dependence (Roberts et al., 2017). Validation should respect shared identities, scenes or capture times when units near one another would otherwise leak into nominally held-out data. For a prospective comparison, the task and knowledge inventories help separate performance on well-represented activities from evidence about broader transfer to unusual interactions.

Measurement and error

Report which raters, measurements and agreement definition enter the coefficient, since different designs give superficially similar statistics different meanings. Another study compares choices of forecast error scale across series (Hyndman and Koehler, 2006). During validation, related work distinguishes forms of intraclass correlation according to the rating design (Koo and Li, 2016). Scaling choices determine how units of different magnitude contribute to an aggregate error; a transparent report retains the original-scale results where interpretable.

Systems and comparators

One approach investigates data-efficient Atari learning using a learned planning model (Ye et al., 2021). One published study uses a transformer world model for sample-efficient reinforcement learning (Micheli et al., 2022). Data-efficiency comparisons benefit from matching environment-interaction budgets and specifying whether any additional planning or pretrained information differs between systems. For a prospective comparison, an ablation can hold histories fixed and alter the predictive model, separating changes to rollout quality from changes in task selection.

A sequential decision formulation keeps reward, action, observation and state distinct, which helps determine where a claimed capability can actually be measured. Cross-domain reporting is most informative when the analysis preserves environment-specific scores rather than treating differently scaled rewards as interchangeable observations. In practical terms, a relevant contribution uses world models across several control domains (Hafner et al., 2023). One approach defines sequential decision problems in terms of state, action and return (Sutton and Barto, 2018).

A methodological precedent defines an emulator-based environment for testing general game agents (Bellemare et al., 2012). Tests on unseen scenes and seen control types can separate transfer to new appearance from transfer to new dynamics and commands. For implementation, the emulator protocol keeps the task definition observable, including how action sets and episode boundaries affect which returns can legitimately be compared. A relevant contribution investigates action-conditioned game video and transfer to new scenes (Yu et al., 2025).

The distribution that generated an archive constrains follow-up evaluation, particularly where difficult states or rare recovery actions are sparsely represented. A relevant contribution uses imagined trajectories in learned latent dynamics for control (Hafner et al., 2019). Latent rollout and task reward are different outputs, and their errors can be isolated by comparing predictions over a common initial history. Earlier research reviews the problem of learning policies from a fixed set of experience (Levine et al., 2020).

Reconstructing the evaluation

A reproducible result requires identifiers for both the released inputs and the exact intermediate tables from which the paper derives its figures. A methodological precedent presents single-file implementations that make training details inspectable (Huang et al., 2021). An existing method specifies principles for discoverable and reusable research artifacts (Wilkinson et al., 2016). In a controlled test, compact training files can expose update order and defaults that would otherwise be difficult to reconstruct from a named algorithm and final results.

Linking evidence to the evaluation

One approach uses procedurally generated environments to study generalization (Cobbe et al., 2019). In an evaluation, related work contrasts precision-recall and receiver operating characteristic summaries under class imbalance (Saito and Rehmsmeier, 2015). An existing method controls the false discovery rate for a declared family of hypotheses (Benjamini and Hochberg, 1995). Low prevalence can make a seemingly small false-positive rate operationally costly, so report precision at plausible recall levels and retain class counts. Holding out level generation seeds permits a cleaner transfer test than adding repetitions of environments that the system already encountered. A family of endpoints should be declared before evaluation so that reported discoveries reflect every comparison eligible to appear in the result.

Conclusion

The balanced comparison in the BBF archive survives every one-game omission, with its size most affected by Boxing. This supports a limited statement about stability within the represented game set. It does not establish coverage of new mechanics or causal explanations for individual task differences. Interactive-rendering reports can use influence analysis productively when they preserve their primary task population, identify consequential specializations, and keep benchmark sensitivity separate from evidence of external generalization.

References

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.

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.

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.

Henderson, Peter, et al. "Deep Reinforcement Learning that Matters." *arXiv preprint arXiv:1709.06560*, 2017. https://arxiv.org/abs/1709.06560.

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

Koo, Terry K., and Mae Y. Li. "A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research." *Journal of Chiropractic Medicine*, vol. 15, no. 2, 2016, pp. 155-163. https://doi.org/10.1016/j.jcm.2016.02.012.

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.

Newey, Whitney K., and Kenneth D. West. “A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix.” *Econometrica*, vol. 55, no. 3, 1987, pp. 703-708. https://doi.org/10.2307/1913610.

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.

Wang, Guanzhi, et al. "Voyager: An Open-Ended Embodied Agent with Large Language Models." *arXiv preprint arXiv:2305.16291*, 2023. https://arxiv.org/abs/2305.16291.

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

Hessel, Matteo, et al. "Rainbow: Combining Improvements in Deep Reinforcement Learning." *arXiv preprint arXiv:1710.02298*, 2017. https://arxiv.org/abs/1710.02298.

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.

Huang, Shengyi, et al. "CleanRL: High-quality Single-file Implementations of Deep Reinforcement Learning Algorithms." *arXiv preprint arXiv:2111.08819*, 2021. https://arxiv.org/abs/2111.08819.

Saito, Takaya and Marc Rehmsmeier. “The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets.” *PLOS ONE*, vol. 10, no. 3, 2015, pp. e0118432. https://doi.org/10.1371/journal.pone.0118432.

Stuart, Elizabeth A. "Matching Methods for Causal Inference: A Review and a Look Forward." *Statistical Science*, vol. 25, no. 1, 2010, pp. 1-21. https://doi.org/10.1214/09-sts313.

Hyndman, Rob J. and Anne B. Koehler. “Another look at measures of forecast accuracy.” *International Journal of Forecasting*, vol. 22, no. 4, 2006, pp. 679-688. https://doi.org/10.1016/j.ijforecast.2006.03.001.

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

Fan, Linxi, et al. "MineDojo: Building Open-Ended Embodied Agents with Internet-Scale Knowledge." *arXiv preprint arXiv:2206.08853*, 2022. https://arxiv.org/abs/2206.08853.

Ye, Weirui, et al. "Mastering Atari Games with Limited Data." *arXiv preprint arXiv:2111.00210*, 2021. https://arxiv.org/abs/2111.00210.

Hafner, Danijar, et al. "Mastering Diverse Domains through World Models." *arXiv preprint arXiv:2301.04104*, 2023. https://arxiv.org/abs/2301.04104.

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.

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.

Benjamini, Yoav and Daniel Yekutieli. “The control of the false discovery rate in multiple testing under dependency.” *The Annals of Statistics*, vol. 29, no. 4, 2001, pp. 1165-1188. https://doi.org/10.1214/aos/1013699998.

Yu, Jiwen, et al. "GameFactory: Creating New Games with Generative Interactive Videos." *arXiv preprint arXiv:2501.08325*, 2025. https://arxiv.org/abs/2501.08325.

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.

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.

Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker. “Fitting Linear Mixed-Effects Models Using lme4.” *Journal of Statistical Software*, vol. 67, no. 1, 2015. https://doi.org/10.18637/jss.v067.i01.

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.