Operational Evidence and Outcomes
Volume 1 · Number 1 · 2026

One Success in How Many Attempts? Threshold Attainment in Captured Game Runs

Oliver Wells[1] · Aaron Coleman[2] · Evan Anderson[3]

[1] Denmark · [2] Norway · [3] Italy

Original publication: November 8, 2026 | Article 708

Abstract

An “at least once” result depends on how many opportunities a report permits. Holding 2,082 captured BBF records fixed, we calculated exact threshold-attainment chances for one, five and ten uniformly selected records within each of 55 games. The equal-game chance rose from 35.51% to 46.43% and then 49.61%. These changes required no improvement in any stored score. They changed the event being reported. Twenty-five games had no captured threshold-meeting record, while nine met the threshold in every record. An attempt budget therefore belongs beside a success claim, particularly when the reader might otherwise interpret it as typical single-run performance.

Exact finite-record selection analysis. An attempt means one captured run-score record; no agent is retrained and no new gameplay or future-seed success probability is measured.

One Success in How Many Attempts? Threshold Attainment in Captured Game Runs

> Exact finite-record selection analysis. An attempt means one captured run-score record; no agent is retrained and no new gameplay or future-seed success probability is measured.

Abstract

An “at least once” result depends on how many opportunities a report permits. Holding 2,082 captured BBF records fixed, we calculated exact threshold-attainment chances for one, five and ten uniformly selected records within each of 55 games. The equal-game chance rose from 35.51% to 46.43% and then 49.61%. These changes required no improvement in any stored score. They changed the event being reported. Twenty-five games had no captured threshold-meeting record, while nine met the threshold in every record. An attempt budget therefore belongs beside a success claim, particularly when the reader might otherwise interpret it as typical single-run performance.

Success is an event with an opportunity budget

A benchmark report can summarize a typical recorded run or ask whether any available run reaches a target. Those questions have different answers even when they use exactly the same scores. Allowing more attempts gives an at-least-one event more opportunities to occur. The resulting improvement should not automatically be interpreted as better performance from an individual run or a stronger learned policy.

Our research question is how large that reporting change becomes in a captured multi-game collection when the permitted attempt count increases from one to five and then ten. The event is a normalized score of at least one. We use fixed budgets that every captured game supports and calculate the finite-record event probability exactly rather than generate a favorable set of example draws.

Meng et al., 2026 propose a paradigm for interactive game rendering. The current work does not evaluate rendering or its interaction mechanism. The connection concerns the reporting of benchmark success: a statement that an interactive system achieved a target should identify the opportunities, retries or candidate runs available before success was declared. Our empirical example addresses that accounting in policy-score records, where the event can be reconstructed transparently.

There is no mathematical surprise in the direction of the change. An at-least-one event cannot become less likely as a uniform subset is enlarged within a fixed collection. The empirical question is its size across games with different observed numbers of successful and unsuccessful records. We distinguish that size from evidence that the evaluated agent itself improved.

The captured score collection

The input is the public BBF run table associated with its benchmark research (Schwarzer et al., 2023). It contains 2,082 records over 55 games. A record includes a source seed label and an aggregate game score, including the supplied GameScoreNormalized field. We retain the provided values and define one attempt operationally as selecting one such record, rather than identifying it with a single environment episode, action or human play session.

The fixed threshold is one in that supplied normalized coordinate. This convention is linked to the source's human-normalized score interpretation, but the present event is a comparison with a stored benchmark reference. We do not recruit human players, estimate a human performance distribution or claim that crossing the threshold demonstrates human-like behavior or general game competence.

Every game has at least sixteen captured records, so budgets of one, five and ten are feasible throughout. Games have unequal record counts; we do not let the larger collections receive greater weight in the final average. All 55 games enter every budget summary. There is no game exclusion based on whether a budget produces an appealing success rate.

The source run records are already available when the hypothetical selections are calculated. We do not execute fresh training attempts or use the calculation to estimate their elapsed cost. If stored runs share implementation, evaluation or training dependencies, the finite-file event remains well defined; extrapolating it to new independent random seeds would require additional assumptions not supplied by this selection exercise.

Reliable reinforcement-learning evaluation distinguishes empirical summaries and the uncertainty around them (Agarwal et al., 2021). Reproducibility research also emphasizes the choices behind a reported comparison (Henderson et al., 2017). Here the specific choice is how many captured opportunities are allowed before a threshold success is counted. It remains relevant even if the input file and every individual score are reproduced perfectly.

Exact subset probabilities

Within a game, let the record count be the available inventory and let the success count be the number of records meeting the threshold. For a budget of five, for example, we consider every unordered five-record subset as equally likely. The failure event is that all five selected records come from the threshold-failing inventory. Dividing its subset count by the total subset count gives the exact failure probability.

The reported success chance is one minus that probability. If fewer than five failing records exist, every five-record subset contains a success and the probability is one. The same calculation applies at budgets one and ten. For a single attempt it reduces to the observed fraction of threshold-meeting records, which we verify numerically for every game.

This is sampling without replacement. Selecting the same failed stored record repeatedly is not permitted. The familiar independent-repeat formula based on a constant success probability describes a different experiment. We do not substitute it for the finite-subset calculation or assume that the captured record fraction is the true success probability of a future training process.

Finite-population sampling principles make the selection design part of the estimand (Neyman, 1934). The numerical work is executed in the recorded local scientific environment (Virtanen et al., 2020), using exact integer combination counts before converting the resulting probability to floating point. No Monte Carlo simulation error or choice of a lucky random draw is present in the reported curves.

The result also differs from reporting the maximum score itself. A very large score and a score just above one both satisfy the binary event. Conversely, a game can have a relatively high mean score while its threshold-meeting frequency remains limited. The chosen event intentionally discards magnitude above the target, so it should accompany rather than replace other performance summaries.

Averaging across games

At each budget, we average the 55 game-specific success chances with equal game weight. This is the chance of an at-least-one event after first selecting a game uniformly from this captured collection and then selecting the allowed records. It is not the fraction of all pooled records that meet the threshold, since pooling would weight games according to their record counts.

Weighting should follow the intended target rather than occur implicitly through the available row counts (Solon et al., 2015). We additionally sum the 55 probabilities to obtain the expected number of games with an at-least-one success under the hypothetical per-game selection plan. That expectation need not be an integer. It is not a newly observed achievement count from a separately executed evaluation.

For each game, we calculate two paired gains: five attempts minus one and ten minus five. The table reports intercept regressions on these gains, with HC3 standard errors, t54 intervals and no controls. N is 55 game rows. Reusing captured records across budgets creates paired outcomes rather than independent experiments with thousands of new observations.

Heteroskedasticity-consistent variance supplies a declared working calculation across those game-row quantities (White, 1980). It does not propagate uncertainty about the stored success counts, missing runs or a population of games beyond the benchmark. The confidence intervals are conditional summaries of game-level variation, while the finite-file probabilities themselves are exact under the stated subset selection.

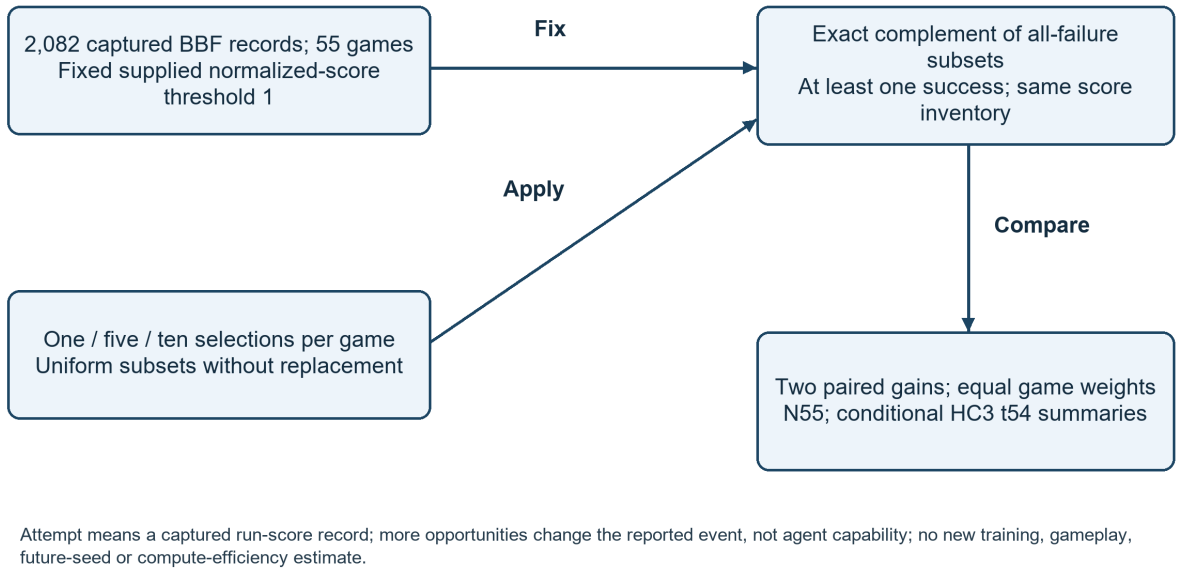


Figure 1. *A fixed per-game score inventory and threshold generate exact any-success probabilities for three attempt budgets. Paired gains summarize additional opportunities, without changing a recorded score or executing a new agent.*

Table 1. *Equal-game gain in at-least-one threshold-attainment chance, N=55, intercept-only OLS, HC3 covariance and t54 intervals. R-squared is zero to rounding. Gains are nonnegative by construction; their observed magnitudes and heterogeneity are the empirical quantities of interest.*

Model	Term	Estimate (SE)	95% CI	N
M1	Mean outcome/contrast	0.109209 (0.022752)	[0.063595, 0.154824]	55
M1: fit	R-squared = 0.0000	-	-	55
M2	Mean outcome/contrast	0.031739 (0.010800)	[0.010086, 0.053392]	55
M2: fit	R-squared = 0.0000	-	-	55

Model labels: M1 = Any-threshold success chance gain from one to five captured attempts; M2 = Any-threshold success chance gain from five to ten captured attempts.

Additional opportunities produce a substantial change

The equal-game success chance is 0.355112 for one attempt, 0.464321 for five and 0.496060 for ten. The first expansion adds 0.109209, standard error 0.022752 and interval [0.063595, 0.154824]. Expressed in percentage points, that is approximately 10.92. The second expansion adds 0.031739, standard error 0.010800 and interval [0.010086, 0.053392], about 3.17 percentage points.

The expected number of games with an at-least-one success rises from 19.5311 to 25.5377 and 27.2833. These values refer to a plan that offers the same budget separately to every game. Such a plan selects 55 records at budget one, 275 at budget five and 550 at budget ten. Record opportunities increase; a fixed total computational-cost comparison is not being performed.

Twenty-five games have no threshold-meeting record in the captured inventory. Increasing the subset budget cannot make those games succeed within this finite collection. Nine games meet the threshold in every stored record, so they already have success probability one at the smallest budget. The remaining 21 games contain a mixture and generate all of the improvement in the event probability.

Across the full inventory, thirty games have at least one threshold-meeting record, while 22 have an average normalized score of at least one. The first count is an all-available-records event with unequal maximum opportunities across games. The second is a mean-score criterion. Neither equals the equal-budget probability summary, and they should not be compared as though they were repeated measurements of the same success definition.

These distinctions require explicit scale language (Stevens, 1946). A ten-percentage-point gain in the chance of at least one selected record meeting the threshold is not a ten-percent improvement in mean reward. It also is not evidence that each attempted run became more reliable. The source score distribution has not changed between the three budgets.

The smaller second increment is an observed aggregate pattern, not a general statement that every extra attempt has equal or declining value under every reporting scheme. The finite counts and threshold determine each game's curve. A future inventory with different success counts could produce different aggregate increments, even while the at-least-one event remains monotone in the allowed number of selections.

Selection, significance and actual capability

Optimization and selection can change which outcomes are highlighted. Work on the optimizer's curse studies related consequences of choosing alternatives using uncertain value estimates (Smith and Winkler, 2006). Our calculation does not estimate that paper's selection bias or apply its Bayesian correction. It demonstrates a separate, explicit change of event: a maximum-over-opportunities success statement versus a single-record success statement.

Multiple testing is another setting where opportunities must be named. False-discovery procedures address families of statistical hypotheses under their respective conditions (Benjamini and Hochberg, 1995; Benjamini and Yekutieli, 2001). Crossing a game-score threshold is not a p-value test, and no false-discovery adjustment is applied here. The distinction prevents a loose analogy about repeated tries from becoming an incorrect claim that the present event has a controlled statistical false-positive rate.

Statistical reporting guidance further separates a thresholded declaration from the magnitude and context of a result (Wasserstein and Lazar, 2016). We do not use the positive working intervals to prove the known monotonicity. They summarize how much the captured games vary in the added opportunity

benefit. The practical comparison remains between clearly specified success events and budgets, rather than a binary statement that more attempts are significant.

Decision costs help determine whether an at-least-one objective is appropriate (Elkan, 2001). Some applications can tolerate retries and genuinely value one successful attempt; others require dependable performance on an arbitrary attempt. Predictive-scoring theory likewise emphasizes matching a score to the target quantity (Gneiting and Raftery, 2007). Our binary threshold summary cannot decide that application preference on its own.

How the broader game literature informs the boundary

EfficientZero and Dreamer examine policy learning with modeled dynamics (Ye et al., 2021; Hafner et al., 2019). DIAMOND, Genie and GameFactory address other learned-world settings (Alonso et al., 2024; Bruce et al., 2024; Yu et al., 2025), while GameNGen supplies another context for generated interactive environments (Valevski et al., 2024). Their model capabilities do not follow from the source record-selection calculation, and none is executed as an experimental arm.

Benchmark and software conventions also affect the meaning of an attempt. Gymnasium, CleanRL and Stable-Baselines3 specify different interfaces or implementation resources (Towers et al., 2024; Huang et al., 2021; Raffin et al., 2021). DeepMind Control Suite and Procgen organize task collections with their own evaluation demands (Tassa et al., 2018; Cobbe et al., 2019). We do not assume that their run counts, episode structures or success criteria match this captured Atari inventory.

A valid comparison of two systems should therefore identify what is held fixed: training effort, evaluation episodes, selected candidates, retries or total compute. Our finite-file exercise holds the inventory and per-game threshold fixed while changing permitted record selections. It measures no wall time, energy or environment-interaction budget, so the larger chance cannot be presented as an efficiency improvement.

A rule that stops as soon as a success appears would also have a distinction between its maximum permitted attempts and the number actually consumed. We calculate the event probability under a budget cap, not the expected stopping cost. Those quantities would require an ordered selection policy and an explicit accounting of attempt costs. Treating the cap as measured processing time would add an interpretation absent from the data.

Similarly, retrying evaluation of one fixed policy and independently training several policies are different sources of variation. The captured score record is our operational unit, and its source aggregation is retained. We do not separate training-seed variability from evaluation-episode variability or assume that either can be eliminated by allowing more record selections. The exact probability is conditional on the supplied inventory rather than a model of how new attempts would be generated.

Influential games and unusual run inventories should remain visible (Cook, 1977). Validation for new tasks or future training seeds would require another design (Roberts et al., 2017). In particular, one should not choose the displayed attempt budget after inspecting which budget creates the most attractive comparison and then describe that choice as an externally established capability level.

The saved table retains each game's record count, threshold-meeting count and three exact event probabilities. Dataset documentation makes the operational attempt definition inspectable (Gebru et al., 2021), and reusable provenance supports tracing the calculation to the captured records (Wilkinson et al., 2016). A reader can therefore see how much the headline changes through additional opportunities before deciding what it says about the performance of a single run.

References

1. 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.
2. Henderson, Peter; Islam, Riashat; Bachman, Philip; Pineau, Joelle; Precup, Doina; Meger, David. (2017). Deep Reinforcement Learning that Matters. arXiv:1709.06560. https://arxiv.org/abs/1709.06560.
3. Bruce, Jake, et al. "Genie: Generative Interactive Environments" *arXiv preprint arXiv:2402.15391*, 2024. https://arxiv.org/abs/2402.15391.
4. 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.
5. 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.
6. Hafner, Danijar, et al. "Dream to Control: Learning Behaviors by Latent Imagination" *arXiv preprint arXiv:1912.01603*, 2019. https://arxiv.org/abs/1912.01603.
7. Wilkinson, Mark D.; Dumontier, Michel; Aalbersberg, IJsbrand Jan; Appleton, Gabrielle; Axton, Myles; Baak, Arie; Blomberg, Niklas; Boiten, Jan-Willem; da Silva Santos, Luiz Bonino; Bourne, Philip E.; Bouwman, Jildau; Brookes, Anthony J.; Clark, Tim; Crosas, Mercè; Dillo, Ingrid; Dumon, Olivier; Edmunds, Scott; Evelo, Chris T.; Finkers, Richard; Gonzalez-Beltran, Alejandra; Gray, Alasdair J.G.; Groth, Paul; Goble, Carole; Grethe, Jeffrey S.; Heringa, Jaap; ’t Hoen, Peter A.C; Hooft, Rob; Kuhn, Tobias; Kok, Ruben; Kok, Joost; Lusher, Scott J.; Martone, Maryann E.; Mons, Albert; Packer, Abel L.; Persson, Bengt; Rocca-Serra, Philippe; Roos, Marco; van Schaik, Rene; Sansone, Susanna-Assunta; Schultes, Erik; Sengstag, Thierry; Slater, Ted; Strawn, George; Swertz, Morris A.; Thompson, Mark; van der Lei, Johan; van Mulligen, Erik; Velterop, Jan; Waagmeester, Andra; Wittenburg, Peter; Wolstencroft, Katherine; Zhao, Jun; Mons, Barend. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data. https://doi.org/10.1038/sdata.2016.18.
8. Valevski, Dani, et al. "Diffusion Models Are Real-Time Game Engines" *arXiv preprint arXiv:2408.14837*, 2024. https://arxiv.org/abs/2408.14837.
9. Stevens, S. S.. (1946). On the Theory of Scales of Measurement. Science. https://doi.org/10.1126/science.103.2684.677.
10. Meng, Zijie; Che, Jinming; Wei, Bingcai; Cao, Xixin. (2026). Make a game: A novel paradigm for interactive game rendering. ICASSP 2026, 1026-1030. https://doi.org/10.1109/ICASSP55912.2026.11461927.
11. Tassa, Yuval, et al. "DeepMind Control Suite." *arXiv preprint arXiv:1801.00690*, 2018. https://arxiv.org/abs/1801.00690.
12. Schwarzer, Max; Obando-Ceron, Johan; Courville, Aaron; Bellemare, Marc; Agarwal, Rishabh; Castro, Pablo Samuel. (2023). Bigger, Better, Faster: Human-level Atari with human-level efficiency. arXiv:2305.19452. https://arxiv.org/abs/2305.19452.
13. Towers, Mark; Kwiatkowski, Ariel; Terry, Jordan; Balis, John U.; De Cola, Gianluca; Deleu, Tristan; Goulão, Manuel; Kallinteris, Andreas; Krimmel, Markus; KG, Arjun; Perez-Vicente, Rodrigo; Pierré, Andrea; Schulhoff, Sander; Tai, Jun Jet; Tan, Hannah; Younis, Omar G.. (2024). Gymnasium: A Standard Interface for Reinforcement Learning Environments. arXiv:2407.17032. https://arxiv.org/abs/2407.17032.
14. Solon, Gary; Haider, Steven J.; Wooldridge, Jeffrey M.. (2015). What Are We Weighting For?. Journal of Human Resources. https://doi.org/10.3368/jhr.50.2.301.
15. Ye, Weirui, et al. "Mastering Atari Games with Limited Data." *arXiv preprint arXiv:2111.00210*, 2021. https://arxiv.org/abs/2111.00210.
16. Cook, R. Dennis. (1977). Detection of Influential Observation in Linear Regression. Technometrics. https://doi.org/10.1080/00401706.1977.10489493.
17. Gneiting, Tilmann; Raftery, Adrian E. (2007). Strictly Proper Scoring Rules, Prediction, and Estimation. Journal of the American Statistical Association. https://doi.org/10.1198/016214506000001437.
18. Smith, James E.; Winkler, Robert L.. (2006). The Optimizer’s Curse: Skepticism and Postdecision Surprise in Decision Analysis. Management Science. https://doi.org/10.1287/mnsc.1050.0451.
19. 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.
20. Yu, Jiwen, et al. "GameFactory: Creating New Games with Generative Interactive Videos." *arXiv preprint arXiv:2501.08325*, 2025. https://arxiv.org/abs/2501.08325.

21. Neyman, Jerzy. (1934). On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection. Journal of the Royal Statistical Society,97(4),558-606. https://doi.org/10.1111/j.2397-2335.1934.tb04184.x.

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

23. Gebru, Timnit, et al. "Datasheets for datasets" *Communications of the ACM*, 2021. https://doi.org/10.1145/3458723.

24. Virtanen, Pauli; Gommers, Ralf; Oliphant, Travis E.; Haberland, Matt; Reddy, Tyler; Cournapeau, David; Burovski, Evgeni; Peterson, Pearu; Weckesser, Warren; Bright, Jonathan; van der Walt, Stéfan J.; Brett, Matthew; Wilson, Joshua; Millman, K. Jarrod; Mayorov, Nikolay; Nelson, Andrew R. J.; Jones, Eric; Kern, Robert; Larson, Eric; Carey, C J; Polat, İlhan; Feng, Yu; Moore, Eric W.; VanderPlas, Jake; Laxalde, Denis; Perktold, Josef; Cimrman, Robert; Henriksen, Ian; Quintero, E. A.; Harris, Charles R.; Archibald, Anne M.; Ribeiro, Antônio H.; Pedregosa, Fabian; van Mulbregt, Paul; SciPy 1.0 Contributors; Vijaykumar, Aditya; Bardelli, Alessandro Pietro; Rothberg, Alex; Hilboll, Andreas; Kloeckner, Andreas; Scopatz, Anthony; Lee, Antony; Rokem, Ariel; Woods, C. Nathan; Fulton, Chad; Masson, Charles; Häggström, Christian; Fitzgerald, Clark; Nicholson, David A.; Hagen, David R.; Pasechnik, Dmitrii V.; Olivetti, Emanuele; Martin, Eric; Wieser, Eric; Silva, Fabrice; Lenders, Felix; Wilhelm, Florian; Young, G.; Price, Gavin A.; Ingold, Gert-Ludwig; Allen, Gregory E.; Lee, Gregory R.; Audren, Hervé; Probst, Irvin; Dietrich, Jörg P.; Silterra, Jacob; Webber, James T; Slavič, Janko; Nothman, Joel; Buchner, Johannes; Kulick, Johannes; Schönberger, Johannes L.; de Miranda Cardoso, José Vinícius; Reimer, Joscha; Harrington, Joseph; Rodríguez, Juan Luis Cano; Nunez-Iglesias, Juan; Kuczynski, Justin; Tritz, Kevin; Thoma, Martin; Newville, Matthew; Kümmerer, Matthias; Bolingbroke, Maximilian; Tartre, Michael; Pak, Mikhail; Smith, Nathaniel J.; Nowaczyk, Nikolai; Shebanov, Nikolay; Pavlyk, Oleksandr; Brodtkorb, Per A.; Lee, Perry; McGibbon, Robert T.; Feldbauer, Roman; Lewis, Sam; Tygier, Sam; Sievert, Scott; Vigna, Sebastiano; Peterson, Stefan; More, Surhud; Pudlik, Tadeusz; Oshima, Takuya; Pingel, Thomas J.; Robitaille, Thomas P.; Spura, Thomas; Jones, Thouis R.; Cera, Tim; Leslie, Tim; Zito, Tiziano; Krauss, Tom; Upadhyay, Utkarsh; Halchenko, Yaroslav O.; Vázquez-Baeza, Yoshiki. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. Nature Methods. https://doi.org/10.1038/s41592-019-0686-2.

25. 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.

26. 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.

27. Raffin, Antonin; Hill, Ashley; Gleave, Adam; Kanervisto, Anssi; Ernestus, Maximilian; Dormann, Noah. (2021). Stable-Baselines3: Reliable Reinforcement Learning Implementations. Journal of Machine Learning Research 22(268), 1-8. https://jmlr.org/papers/v22/20-1364.html.

28. Wasserstein, Ronald L.; Lazar, Nicole A.. (2016). The ASA Statement on p -Values: Context, Process, and Purpose. The American Statistician. https://doi.org/10.1080/00031305.2016.1154108.

29. 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.

30. Elkan, Charles. “The Foundations of Cost-Sensitive Learning.” *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence*, 2001, pp. 973-978. https://cseweb.ucsd.edu/~elkan/rescale.pdf.