Population Variation and Equity
Volume 1 · Number 1 · 2026

A Low Pilot Score Is Not a Reliable Map of Benchmark Advantage: Disjoint-Run Subgroups in Atari Reports

Jeremy Mitchell[1] · Simon Wells[2] · Ryan King[3]

[1] Sweden · [2] Finland · [3] Ireland

Original publication: November 3, 2026 | Article 677

Abstract

Which games should be expected to show the larger benchmark gap? We used disjoint no-reset run records to define lower- and higher-pilot-score groups across fifty-five Atari games. The BBF-minus-comparison gap averaged 0.583 in the lower group and 1.461 in the upper group. Their difference was uncertain, as was a continuous pilot-score association. Separating pilot records from comparison records avoids an exact shared-outcome artifact, but one noisy pilot split still provides a weak basis for targeting future evaluations.

Reanalysis of released BBF and no-reset run records. Subgroups use a separate pilot portion; no agents, game interactions, or rendered videos were produced.

A Low Pilot Score Is Not a Reliable Map of Benchmark Advantage: Disjoint-Run Subgroups in Atari Reports

> Reanalysis of released BBF and no-reset run records. Subgroups use a separate pilot portion; no agents, game interactions, or rendered videos were produced.

Abstract

Which games should be expected to show the larger benchmark gap? We used disjoint no-reset run records to define lower- and higher-pilot-score groups across fifty-five Atari games. The BBF-minus-comparison gap averaged 0.583 in the lower group and 1.461 in the upper group. Their difference was uncertain, as was a continuous pilot-score association. Separating pilot records from comparison records avoids an exact shared-outcome artifact, but one noisy pilot split still provides a weak basis for targeting future evaluations.

1. Subgroup explanations can begin with the outcome they seek to explain

A benchmark average often leaves a practical question unanswered: where does the reported condition difference occur? Researchers may divide games into easy and difficult groups, then compare the gains. If difficulty is defined using the same baseline scores that are subtracted from the new condition, a relationship can arise partly because the baseline outcome appears on both sides. The resulting subgroup story can look more informative than the underlying design permits.

This paper asks whether lower pilot scores in the no-reset condition identify games with a different BBF advantage when that advantage is computed from disjoint no-reset records. The task is descriptive. We use the released BBF resource Schwarzer et al., 2023 and its no-reset table to build an empirical grouping without assigning game genres or claiming to measure intrinsic task difficulty. The pilot does not supply any record used in the no-reset comparison mean.

Interactive rendering research such as Meng et al., 2026 motivates broad evaluation across game settings, where a single pooled result can obscure varied operating conditions. The present records concern agent scores, not rendering quality or interaction fidelity. We do not run the target paradigm or infer how it would perform on these games. The connection is the need for defensible subgroup definitions when explaining heterogeneous benchmark outcomes.

Two analytical propositions follow. The subgroup definition should be separated from the exact baseline observations entering the evaluated gap. After making that separation, the observed subgroup difference should retain its uncertainty rather than being translated into a targeting rule automatically. A numerical lower-score group is not assumed to receive a larger improvement, and an uncertain association is reported without selecting a different split to strengthen it.

2. Benchmark diversity and the scope of a score gap

Deep reinforcement-learning evaluation has raised substantial concerns about aggregate summaries and uncertainty Agarwal et al., 2021, while differences in experimental practice can affect apparent performance Henderson et al., 2017. Those concerns motivate an explicit accounting of the run records and game-level units. They do not make the released runs a random sample from all environments, nor do they establish that two CSV conditions differ only in the reset mechanism named in their labels.

Atari learning systems Mnih et al., 2013 and algorithmic combinations such as Rainbow Hessel et al., 2017 provide context for why benchmark scores can differ markedly by game. Procedural generation

Cobbe et al., 2019 adds another axis of environmental variation, but it is not evaluated here. The current grouping uses only released no-reset scores. It does not incorporate visual complexity, exploration requirements, action-space size, or any unobserved measure of how difficult a game is for humans.

World-model approaches Ha and Schmidhuber, 2018 and latent imagination Hafner et al., 2019 also illustrate the breadth of game-related evaluation questions. Transformer world models Micheli et al., 2022 and diffusion world models Alonso et al., 2024 change the modeling object, not the meaning of a pilot subgroup in this dataset. Generative interactive environments Bruce et al., 2024 and neural game engines Valevski et al., 2024 similarly require their own evidence. Their inclusion in the discussion identifies boundaries beyond agent-score analysis rather than suggesting a shared empirical performance comparison.

3. Keeping every game that meets the record-count gate

The captured BBF table has 2,082 run records and the no-reset table 830. Both contain the same fifty-five game identifiers, and every shared game has at least ten records in each condition. We retain all fifty-five, including games whose condition means or held gaps are tied. An outcome-dependent requirement for a nonzero difference would alter the subgroup population and is unnecessary for the current question.

For each game, no-reset records are first sorted by their source seed identifier, then permuted using one fixed pseudorandom stream initialized at 20261107. The first floor half forms the pilot and the remainder the comparison portion. Across games, this yields 388 pilot records and 442 disjoint no-reset comparison records. All 2,082 BBF records enter their game-specific comparison means. Source seeds are identifiers, not historical dates, and conditions are not paired by equal seed values.

Different record counts across conditions affect the precision of their means but do not change the equal-game definition of the analysis. We do not equalize counts by choosing outcomes, invent additional runs, or claim a controlled training-budget comparison. The CSV files alone do not independently establish every environment and training detail behind their records. The relevant empirical object is the gap between captured condition summaries under the documented selection rule.

Each game's pilot score is the mean released normalized score among its pilot records. Sorting those means and breaking exact ties by game identifier assigns the lowest twenty-seven games to the lower group and the remaining twenty-eight to the upper group. The two values adjacent to the split are 0.171301 and 0.173298. This is a near-equal-size empirical partition, without a universal threshold for difficult games.

4. Disjoint observations remove one artifact, not every dependency

The game-level outcome subtracts the held no-reset mean from the BBF mean. Because no exact no-reset observation appears in both the pilot and the gap, an unusually low pilot record is not mechanically subtracted again from that game's evaluated advantage. This design addresses a specific shared-observation problem. It does not remove persistent game characteristics that can influence both the pilot level and the gap, or uncertainty in the pilot's estimated rank.

Model 1 regresses the gap on an intercept and the lower-pilot indicator. The indicator coefficient is therefore the lower-group average minus the upper-group average. Model 2 replaces that binary predictor with the pilot score centered at its fifty-five-game mean. The continuous model avoids treating the empirical split as a natural scientific boundary, but it still imposes a linear descriptive relationship and can be sensitive to unusual score levels.

Both models use fifty-five game rows and HC3 standard errors White, 1980, with a t reference based on fifty-three residual degrees of freedom. No run-level row is treated as an independent subgroup observation. Dependence-aware inference Cameron and Miller, 2015 motivates this aggregation, although game-level robust covariance cannot recover uncertainty from unobserved training histories or dependence common to the source evaluation protocol. The intervals condition on the one pilot split and captured source cohort.

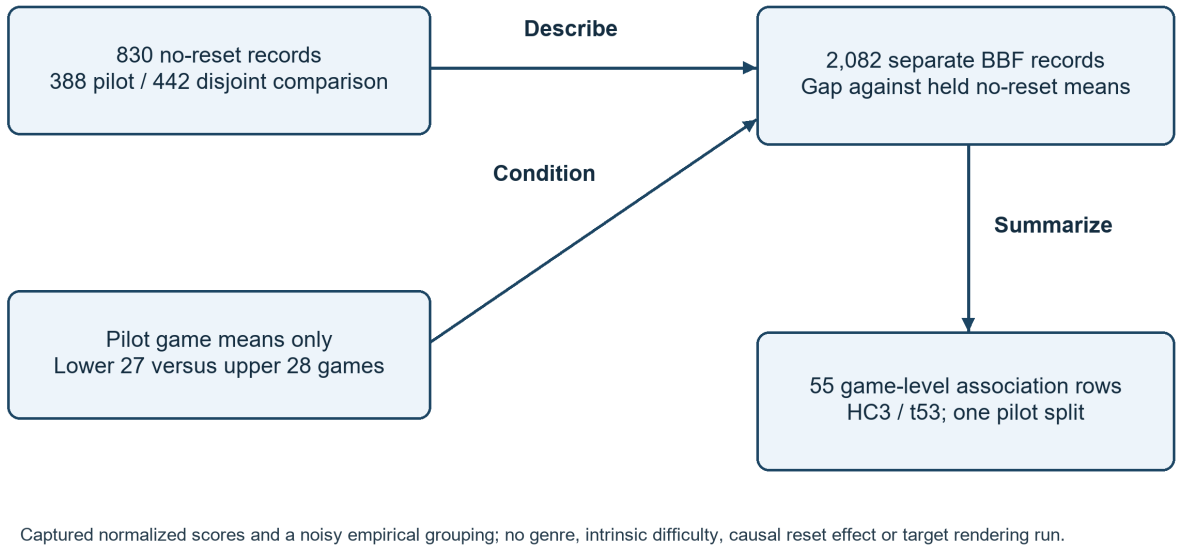


Figure 1. Disjoint no-reset pilot and comparison records define empirical game groups and an evaluated BBF gap, with two game-level association summaries.

Figure 1 highlights a measurement choice rather than a randomized treatment. Causal reasoning Greenland et al., 1999 prevents interpreting the lower-group coefficient as the effect of making a game easier or harder. Measurement-error analysis Carroll et al., 2006 also matters because pilot means are estimated from limited records. Their classification uncertainty is not propagated by the conditional HC3 calculation, and the binary group can change under a different pilot draw.

5. The lower-score group does not show the larger point estimate

The lower group's pilot mean is 0.006475, and its disjoint no-reset comparison mean is 0.025667. The upper group's corresponding values are 0.883089 and 0.868496. These differences show that the split identifies persistent level differences in this captured record. They do not establish a comprehensive difficulty scale or prove that the pilot estimates are sufficiently precise for prospective game selection.

The BBF means are 0.608216 in the lower group and 2.329034 in the upper group. Subtracting their respective held no-reset means gives average gaps of 0.582549 and 1.460539. BBF has a positive held gap in twenty-three of twenty-seven lower-group games and twenty-six of twenty-eight upper-group games. Positive-game counts and group-average magnitude are retained as separate summaries because a few large values can influence a mean without changing many signs.

Table 1. Conditional game-level regressions for the held normalized-score gap. *M1 uses the lower-pilot indicator, with the upper group as reference. M2 uses mean-centered continuous pilot score. Each has N55 and HC3 standard errors with t53 intervals; the pilot split, condition records, and source normalization are treated as fixed for this inference.*

Model	Term	Estimate (SE)	95% CI	N
M1	Upper-pilot mean gap	1.4605 (0.4645)	[0.5289, 2.3922]	55
M1	Lower-pilot minus upper-pilot group	-0.8780 (0.5458)	[-1.9727, 0.2167]	55
M1: fit	R-squared = 0.0474	-	-	55
M2	Gap at mean pilot score	1.0295 (0.3223)	[0.3831, 1.6759]	55
M2	Centered pilot normalized score	0.3473 (1.2187)	[-2.0970, 2.7917]	55
M2: fit	R-squared = 0.0232	-	-	55

Model labels: M1 = Lower-pilot subgroup difference in held BBF gap; M2 = Continuous pilot-score association with held BBF gap.

The lower-minus-upper coefficient is -0.877990, with standard error 0.545784 and interval -1.972695 to 0.216714. Its point estimate opposes the idea that lower pilot performance necessarily identifies a larger absolute normalized-score gap. The interval nevertheless crosses zero and does not establish a stable difference between these groups. Model 1 has R-squared 0.047431, leaving most observed game-level variation outside this simple partition.

The continuous pilot-score slope is 0.347315, with standard error 1.218680 and interval -2.097046 to 2.791677. Model 2 has R-squared 0.023155. The very wide interval supports a restrained interpretation: the captured pilot level is not a precise linear map of the held gap. The intercept, 1.029525, equals the overall average gap because the predictor is centered, rather than representing the expected result for a hypothetical game with zero pilot score.

The group averages also address magnitude rather than the value of another evaluation run. A game with a large expected gap may need little additional evidence to determine its sign, whereas a game with a small but noisy gap may remain difficult to classify. Pilot score level does not directly measure that uncertainty. Accordingly, our results should not be turned into a rule that prioritizes whichever subgroup has the larger point estimate. No marginal information gain, monetary run cost, or decision loss was measured, and the saved records do not establish an optimal allocation of new experiments.

6. Normalization and imbalance affect the story a subgroup can tell

Pilot means range from -0.987499 to 5.828130 in the released normalized units. We retain those values and do not clip them into a nominal zero-to-one range. The file's normalized-score field supplies the measurement scale; our analysis does not estimate new human or random-agent normalization baselines. Comparing scores across games therefore inherits the source convention, including its consequences for large positive and negative values.

Evaluation metrics can weight errors and comparisons differently Hyndman and Koehler, 2006, and scoring-rule analysis Gneiting and Raftery, 2007 underscores that a summary must match the question. The current outcome is an absolute difference in released normalized units, not a percentage gain over the no-reset mean. The latter could behave poorly near zero or for negative pilot values. We do not switch to

relative gains after seeing the subgroup pattern or interpret our coefficient as a universal percentage improvement.

The lower and upper groups contain twenty-seven and twenty-eight games because the retained total is odd. This slight imbalance is stated rather than repaired by deleting an otherwise eligible game. Missing-data reasoning Little and Rubin, 2002 is relevant to any extension with incomplete conditions, but no missing-value imputation is needed for these fifty-five shared games. The primary record-count gate and the later subgroup assignment remain distinct operations.

7. Can a pilot guide the next evaluation budget?

This evidence does not justify allocating future runs solely to games with low pilot scores. A targeting rule would require a prospectively specified objective, repeated pilot uncertainty, and evaluation on records independent of its selection process. Validation principles for structured observations Roberts et al., 2017 are relevant because new runs on the same games address a different generalization question from performance on previously unseen games.

There is a further distinction between constructing a pilot from an existing report and collecting one before an actual benchmark decision. Our split is performed after the source records exist, and the seed establishes reproducibility rather than prospective registration. We do not know that the designated pilot runs were available earlier in the original workflow. The experiment demonstrates a feasible separation of observations, not a historical claim about the sequence in which a researcher learned the results. For a new game, the proposed grouping would itself require an initial no-reset sample and an assessment of whether that sample represents the intended operating condition.

The comparison was not selected from many candidate partitions. We keep the one seeded split, its near-equal groups, and the continuous companion model. Evaluation-selection concerns Varma and Simon, 2006 and multiple-testing principles Benjamini and Hochberg, 1995 motivate that restraint. A future search across thresholds, transformations, and subsets would require its own validation record; it could not reuse the current intervals as if the preferred analysis had been fixed beforehand.

Large values could make influence diagnostics useful Cook, 1977, but no observation was removed to obtain a cleaner subgroup pattern. The source scores, pilot assignment, and game-level outcomes are saved for examination. Statistical interpretation should reflect design and magnitude rather than a threshold alone Wasserstein and Lazar, 2016. In particular, an interval crossing zero does not show that all game-specific effects are the same or that every potential subgroup is uninformative.

The engineering literature offers reusable evaluation infrastructure, including CleanRL Huang et al., 2021, Gymnasium Towers et al., 2024, and Stable-Baselines3 Raffin et al., 2021. None was executed to produce new outcomes here. They indicate feasible directions for a fresh, controlled follow-up with common budgets and recorded environment settings. Such a follow-up could test an actual subgroup-targeted resource policy rather than infer its value from the present secondary table analysis.

Documentation Gebru et al., 2021 and reusable provenance Wilkinson et al., 2016 are especially important when pilot and evaluation records originate in the same downloaded file. The saved assignment lists identify each record's role and make the disjointness check reproducible. That transparency supports the narrow empirical finding without turning a convenient internal split into evidence of independent environment replication, causal reset benefit, or rendering-model transfer.

8. Conclusion

Disjoint no-reset records permit a subgroup analysis without reusing the exact baseline outcomes inside the evaluated gap. In the resulting fifty-five-game comparison, lower pilot scores do not identify a clearly larger BBF gap, and both subgroup and continuous associations remain uncertain. The design offers a defensible way to describe heterogeneity, while the result limits what one small pilot can support as a rule for selecting future benchmark effort.

References

1. 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.
2. Ha, David and Jürgen Schmidhuber. "World Models" *arXiv preprint arXiv:1803.10122*, 2018. https://arxiv.org/abs/1803.10122.
3. 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.
4. 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.
5. Micheli, Vincent, et al. "Transformers are Sample-Efficient World Models" *arXiv preprint arXiv:2209.00588*, 2022. https://arxiv.org/abs/2209.00588.
6. 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.
7. 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.
8. 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.
9. 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.
10. Little, Roderick J. A., and Donald B. Rubin. *Statistical Analysis with Missing Data*. Wiley, 2002. https://doi.org/10.1002/9781119013563.
11. 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.
12. 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.
13. Cobbe, Karl, et al. "Leveraging Procedural Generation to Benchmark Reinforcement Learning" *arXiv preprint arXiv:1912.01588*, 2019. https://arxiv.org/abs/1912.01588.
14. 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.
15. Cook, R. Dennis. (1977). Detection of Influential Observation in Linear Regression. Technometrics. https://doi.org/10.1080/00401706.1977.10489493.
16. 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.
17. 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.

18. 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.

19. 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.

20. Mnih, Volodymyr; Kavukcuoglu, Koray; Silver, David; Graves, Alex; Antonoglou, Ioannis; Wierstra, Daan; Riedmiller, Martin. (2013). Playing Atari with Deep Reinforcement Learning. arXiv:1312.5602. https://arxiv.org/abs/1312.5602.

21. Hessel, Matteo; Modayil, Joseph; van Hasselt, Hado; Schaul, Tom; Ostrovski, Georg; Dabney, Will; Horgan, Dan; Piot, Bilal; Azar, Mohammad; Silver, David. (2017). Rainbow: Combining Improvements in Deep Reinforcement Learning. arXiv:1710.02298. https://arxiv.org/abs/1710.02298.

22. 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.

23. 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.

24. Carroll, Raymond J., David Ruppert, Leonard A. Stefanski, and Ciprian M. Crainiceanu. “Measurement Error in Nonlinear Models.” 2006. https://doi.org/10.1201/9781420010138.

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

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

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

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. 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.

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