Operational Evidence and Outcomes
Volume 1 · Number 1 · 2026

Nearly Equal Accuracy, Different Token Choices: A Numerical Compression Audit

Jeffrey Butler[1] · Daniel Ward[2] · James Harrison[3]

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

Original publication: November 8, 2026 | Article 687

Abstract

Eight-bit rounding changed 602 of 16,384 top token choices while producing only six fewer matches to the observed next tokens. The corresponding accuracy difference was -0.0366 percentage points, with a block-level 95% interval of -0.0960 to 0.0227 points. Most changed positions were mismatches under both model states. Choice changes concentrated in the original model's lowest-margin quarter of positions: 14.43 percent changed there, compared with 0.09 percent elsewhere. Four-bit rounding produced much larger disagreement and accuracy loss. This actual local forward experiment on WikiText-2 separates decision stability from observed-token accuracy; it does not measure free-running generation, factual correctness, or a deployed low-bit inference system.

Teacher-forced next-token decisions on a public WikiText-2 segment, using source and reconstructed Pythia-14M internal weights.

Nearly Equal Accuracy, Different Token Choices: A Numerical Compression Audit

Teacher-forced next-token decisions on a public WikiText-2 segment, using source and reconstructed Pythia-14M internal weights.

Abstract

Eight-bit rounding changed 602 of 16,384 top token choices while producing only six fewer matches to the observed next tokens. The corresponding accuracy difference was -0.0366 percentage points, with a block-level 95% interval of -0.0960 to 0.0227 points. Most changed positions were mismatches under both model states. Choice changes concentrated in the original model's lowest-margin quarter of positions: 14.43 percent changed there, compared with 0.09 percent elsewhere. Four-bit rounding produced much larger disagreement and accuracy loss. This actual local forward experiment on WikiText-2 separates decision stability from observed-token accuracy; it does not measure free-running generation, factual correctness, or a deployed low-bit inference system.

1. Similar accuracy can conceal changed decisions

Accuracy summarizes how many predictions match a reference. It does not identify which predictions change between two model states. Improvements and losses can cancel, and predictions can change while both remain different from the reference token. Consequently a small accuracy difference can coexist with many altered decisions. For numerical compression, that distinction matters when a researcher wants to preserve the original model's behavior rather than only one aggregate correctness score.

Our research question is how often fixed weight-rounding policies change the top next-token choice, and how much of that disagreement changes matching to the observed token. We also ask whether changed choices concentrate at positions where the source model's leading two logits are close. The comparison uses teacher forcing: every model sees the same actual text context at a scored position. It is an audit of local decisions, without feedback from one changed generated token into later contexts.

The language-model compression setting in Wang et al., 2026 gives context for the behavioral question. We do not implement that framework or test a compressed model served through optimized kernels. Instead, two explicitly defined internal-weight reconstructions are executed in a small public model. The purpose is to measure what an aggregate score can conceal in a transparent example. It provides no estimate of that target method's compression quality, speed, or operational reliability.

2. A separate public text source

The observations come from the raw-v1 test partition of WikiText-2, associated with Merity et al., 2016, obtained from the Salesforce dataset mirror. The preserved Parquet file contains 4,358 text rows, of which 2,891 are nonempty. We concatenate the stored strings without inserting additional separators, retaining their existing whitespace. The resulting UTF-8 text has 1,287,656 bytes. This rule matters because adding or removing line breaks would alter the token stream and its prediction contexts.

The captured Pythia tokenizer, run through Hugging Face contributors, 2025 version 0.22.1, produces 285,831 tokens. A complete decoding round trip reproduces the joined text exactly, without inserted special markers. The evaluation begins at zero-based input offset 142,915 and scores 16,384 subsequent targets. We divide those targets into sixty-four blocks of 256, resetting context in every block. All 50,304 vocabulary logits remain in score normalization and top-choice selection.

The source test label is a dataset-partition designation, not proof that the text was absent from Pythia pretraining. We make no certified unseen-document claim. Wikipedia-derived source attribution and the captured CC BY-SA 4.0 terms are retained with the local resource. No prose continuation is generated for publication here. The recorded outputs are token identifiers, loss values, and comparison indicators, with the original dataset remaining the documented evaluation reference.

3. A fixed model and two numerical interventions

We load the final standard-Pile checkpoint from Biderman et al., 2023 into a float32 CPU GPT-NeoX model. Numeric safetensors source values are loaded strictly, and the original state is restored before each policy. The local environment uses PyTorch 2.6.0+cpu, represented by Paszke et al., 2019, and Wolf et al., 2019 version 4.57.1. Evaluation mode, eager attention, disabled caching, and four CPU threads define the execution. This is an actual forward-score experiment rather than a simulation of predicted language-model accuracy from weight norms.

Both reconstruction policies affect twenty-four internal attention and multilayer-perceptron projection matrices. Contiguous row-major groups contain 128 weights. Four-bit reconstruction uses a float32 maximum-based scale with integer codes from minus seven through seven; eight-bit reconstruction uses minus 127 through 127. Source input embeddings, output weights, biases, and normalization parameters remain unchanged. Reconstructed matrices execute as dense float32 arrays, so nominal code precision does not imply measured low-bit hardware performance.

Compression methods in Frantar et al., 2022, Lin et al., 2023, Xiao et al., 2022, Dettmers et al., 2022, and Dettmers et al., 2023 involve broader numerical and implementation choices. None is reproduced by this simple maximum-scale rule. Their relevance is the general need to assess the consequences of altered parameter precision. We do not attach their performance labels to the arrays used here, treat these policies as competitive implementations, or infer that a different calibrated procedure would change the same positions.

4. Define a changed choice before naming an error

At each target position we take the identifier with the largest output logit. A choice changes when this identifier differs between the source-valued and rounded model. Observed-token accuracy records whether it equals the actual next identifier in the dataset. These indicators are distinct. A changed choice can lose a match, gain a match, or move from one nonmatching token to another. A nonmatch also does not establish that the proposed token would be semantically unacceptable as an alternative continuation.

Subword segmentation, motivated by Sennrich et al., 2015, explains why the unit is a token identifier rather than necessarily a complete word. The embedding perspectives in Mikolov et al., 2013 and Pennington et al., 2014 further place representation choices within a broader language-modeling context, without supplying vectors to this experiment. We do not count a partial word as a complete lexical mistake, interpret a changed punctuation identifier as a factual error, or treat all mismatches as equally consequential to a reader.

We retain paired transitions instead of subtracting two accuracies and discarding their origin. For each policy we count source-match to nonmatch changes, nonmatch to match changes, and changes between two nonmatching choices. Both model states cannot match the same observed identifier while choosing different identifiers. This simple accounting provides a direct consistency check: these three changed categories must add up to total choice disagreement, while the first two determine the net accuracy difference.

5. Source margins provide a declared subgroup

The source margin is the largest source logit minus the second largest at a position. We identify exactly 4,096 positions with the smallest margins, using stable source-position ordering for ties. The other 12,288 form the comparison group. Membership uses source-model logits only, without looking at which quantized choices change. These groups describe a local decision boundary in the original score vector; they are not labels for difficult semantic topics or risky requests.

Calibration concerns in Guo et al., 2017 limit the interpretation of that margin. We fit no mapping from a margin to a probability of correctness, and no operational reject threshold is selected. Decision-cost reasoning in Elkan, 2001 and selective-decision methods in Geifman and El-Yaniv, 2017 would be relevant if an application used uncertainty to defer an action. The present quartile is a descriptive grouping that locates changes in this sample. It is not a validated recommendation to abstain on one quarter of future inputs.

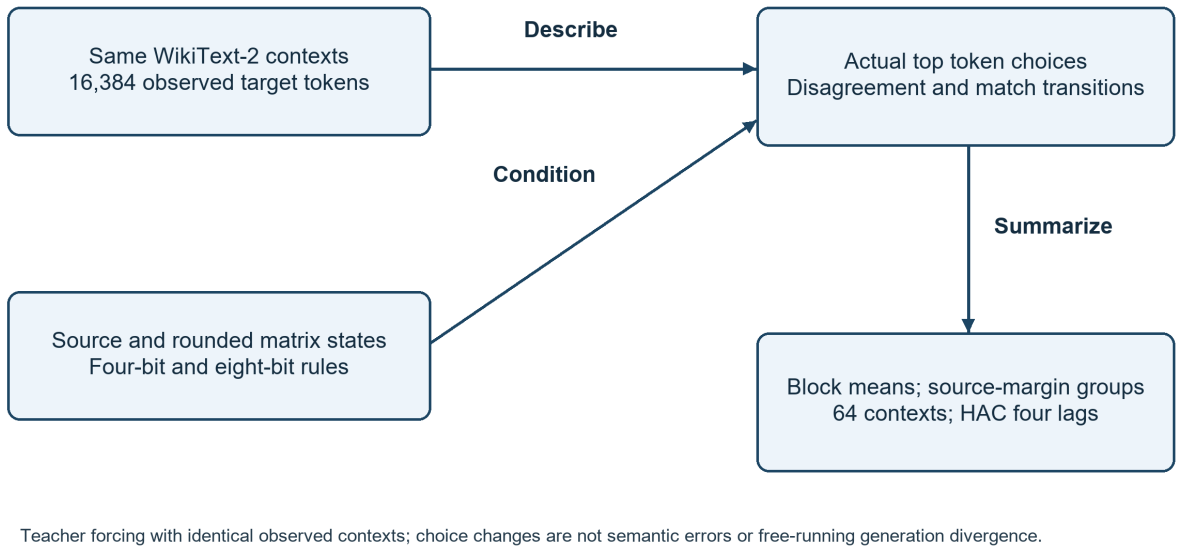

Figure 1. Identical observed contexts produce source and rounded-model choices; disagreement, observed-token matching, and a source-margin subgroup are summarized separately.

Metric distinctions associated with Fawcett, 2006, Saito and Rehmsmeier, 2015, and Matthews, 1975 reinforce why a named summary should not stand for every aspect of prediction behavior. We compute neither ROC curves, precision-recall curves, nor a multiclass correlation statistic. The figure reports the two executed endpoints and their transition decomposition. Adding an unmeasured performance metric to its final box would imply an evaluation that the source records do not support.

6. Block means and serially robust intervals

For each 256-target block we calculate choice disagreement and rounded-minus-source observed-token accuracy separately for the two rounding policies. Each of the resulting four outcome series has sixty-four rows. Intercept-only ordinary least squares estimates their means. Equal block lengths ensure that these means also equal the corresponding token-level fractions. We do not treat 16,384 targets as independent texts or use the number of vocabulary entries as the inferential sample size.

Covariance follows Newey and West, 1987 with four retained-block lags, a finite-sample correction, and a t reference on sixty-three degrees of freedom. Heteroskedasticity-consistent reasoning in White, 1980

alone would omit adjacency among text blocks. The dependence principles discussed in Cameron and Miller, 2015 remain pertinent because all observations come from one segment and one checkpoint. HAC is a declared approximation to serial dependence, rather than evidence that the corpus or model has been independently replicated.

Table 1. Mean source-choice disagreement and observed-token accuracy changes; all coefficients are proportions.

Model	Term	Estimate (SE)	95% CI	N
M1	Mean outcome/contrast	0.528137 (0.004601)	[0.518942, 0.537332]	64
M1: fit	R-squared = 0.0000	-	-	64
M2	Mean outcome/contrast	-0.070190 (0.002333)	[-0.074853, -0.065527]	64
M2: fit	R-squared = 0.0000	-	-	64
M3	Mean outcome/contrast	0.036743 (0.001369)	[0.034008, 0.039478]	64
M3: fit	R-squared = 0.0000	-	-	64
M4	Mean outcome/contrast	-0.000366 (0.000297)	[-0.000960, 0.000227]	64
M4: fit	R-squared = 0.0000	-	-	64

Model labels: M1 = Four-bit source-choice disagreement; M2 = Four-bit top-one accuracy gain; M3 = Eight-bit source-choice disagreement; M4 = Eight-bit top-one accuracy gain.

All fits contain only a mean, giving R-squared values of zero. No covariates or margin effects are fitted in these regressions. The margin grouping is reported descriptively with its exact denominators. We do not claim that adding a source-margin control identifies a causal effect of uncertainty. The completed inference concerns the paired outcomes under the two specified numerical interventions, conditional on the original model, text segment, and context rule.

7. Eight-bit reconstruction changes many more choices than matches

Source-valued observed-token accuracy is 0.233398438. Eight-bit reconstruction yields 0.233032227, a difference of -0.000366211 in proportion units. Its block-level standard error is 0.000296896 and its 95% interval is -0.000959511 to 0.000227089. Expressed in percentage points, the change is approximately -0.0366, with an interval from -0.0960 to 0.0227. The interval does not precisely separate the accuracy difference from zero under this design.

Yet 602 top choices change, a disagreement fraction of 0.036743164. Its standard error is 0.001368722 and interval 0.034007990 to 0.039478338. Of those changes, forty-six lose a match to the observed token and forty gain one. The remaining 516 switch between two nonmatching identifiers. The transition sum is 602, and the net loss is six observed-token matches. A stable marginal accuracy therefore coexists with a visible difference in which token the model selects.

The source mean negative log likelihood is 5.071492389 nats, rising to 5.074478916 under eight-bit reconstruction. The increase is 0.002986527. Predictive scoring principles in Gneiting and Raftery, 2007 distinguish that full-distribution score from top-choice accuracy and disagreement. We report it as descriptive context, without fitting an additional loss test in this article. A small average score change, small accuracy change, and nonzero choice-change rate are compatible facts about different summaries of the same forward passes.

8. Low source margins locate most eight-bit changes

Among the lowest-margin 4,096 positions, eight-bit choices change at 14.4287 percent of targets. Among the other 12,288, the rate is 0.0895 percent. The corresponding counts are 591 and eleven, adding to the 602 changed choices. This concentration is measured on actual model decisions. It does not establish that source margin alone predicts semantic damage, nor that every low-margin choice is unimportant or safely interchangeable with another continuation.

Four-bit reconstruction produces a much larger disagreement fraction, 0.528137207, with interval 0.518942237 to 0.537332177. Observed-token accuracy falls to 0.163208008, a difference of -0.070190430 with interval -0.074853380 to -0.065527480. There are 1,413 source matches lost, 263 matches gained, and 6,977 changes between two nonmatches. The total disagreement thus includes a substantial number of altered decisions whose observed-token status remains unchanged, even in this more severely degraded policy.

The four-bit change rates are 74.6338 percent in the lowest-margin quarter and 45.5404 percent elsewhere. Its mean loss increase is 0.924891630 nats on the same segment. The severe numerical intervention changes many choices beyond the closest source boundaries. We do not compare that quantity with an optimized four-bit implementation or infer a universal precision threshold. It is the outcome of this particular uncalibrated grouped policy on a small model.

9. Behavioral stability remains narrower than deployment reliability

A teacher-forced comparison holds the input context fixed. In free-running generation, an early changed token would alter later inputs and could amplify, cancel, or redirect subsequent differences. We have not executed that process here. A 3.67 percent local choice-change rate is therefore neither a generated-sequence divergence rate nor a probability of a harmful or factually incorrect response. The operational relevance of a changed decision requires a task and consequence model beyond exact next-token matching.

Validation principles in Roberts et al., 2017 and model-selection concerns in Varma and Simon, 2006 also restrict reuse of this segment. Choosing a compression policy after inspecting these outcomes would turn the segment into development evidence for that decision. The current study uses fixed rounding policies and a declared source-margin quartile. Influence reasoning associated with Cook, 1977 could motivate examining blocks with unusually many changes; no block is removed or downweighted because of an inconvenient result.

Interpretation following Wasserstein and Lazar, 2016 separates the uncertain net accuracy difference from the directly observed transition counts. Neither an interval containing zero nor a small p-value on a disagreement mean answers whether changed choices matter to an application. We also avoid reporting the WikiText test designation as a guarantee against pretraining overlap. One public model and one selected encyclopedic text segment provide an inspectable behavior audit, with broader claims about domains, scales, or user-facing tasks still unsupported.

10. Preserve the transition record alongside the aggregate score

The eight-bit example shows why a comparison should retain more than a difference between two accuracies. Hundreds of local choices change while beneficial and harmful match transitions nearly balance, and most remaining switches preserve nonmatch status. A source-margin breakdown helps locate the changes but does not validate their consequences. The four-bit example confirms that the same accounting remains useful when the average degradation is large rather than nearly canceled.

Documentation under Gebru et al., 2021 includes the exact source split, string joining, tokenizer, segment offsets, context resets, unchanged parameter families, and margin-quartile tie rule. Preservation following Wilkinson et al., 2016 retains per-target identifiers, scores, transition indicators, and block summaries with source attribution and runtime versions. These materials support reproduction of the reported choice audit. They do not certify identity with original behavior in generated text or establish a generally acceptable numerical compression policy without further task-specific evidence.

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. Fawcett, Tom. (2006). An introduction to ROC analysis. Pattern Recognition Letters. https://doi.org/10.1016/j.patrec.2005.10.010.
3. 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.
4. 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.
5. Lin, Ji; Tang, Jiaming; Tang, Haotian; Yang, Shang; Chen, Wei-Ming; Wang, Wei-Chen; Xiao, Guangxuan; Dang, Xingyu; Gan, Chuang; Han, Song. (2023). AWQ: Activation-aware Weight Quantization for LLM Compression and Acceleration. https://arxiv.org/abs/2306.00978.
6. Merity, Stephen; Xiong, Caiming; Bradbury, James; Socher, Richard. (2016). Pointer Sentinel Mixture Models. https://arxiv.org/abs/1609.07843.
7. Dettmers, Tim; Svirschevski, Ruslan; Egiazarian, Vage; Kuznedelev, Denis; Frantar, Elias; Ashkboos, Saleh; Borzunov, Alexander; Hoefler, Torsten; Alistarh, Dan. (2023). SpQR: A Sparse-Quantized Representation for Near-Lossless LLM Weight Compression. https://arxiv.org/abs/2306.03078.
8. 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.
9. 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.
10. Frantar, Elias; Ashkboos, Saleh; Hoefler, Torsten; Alistarh, Dan. (2022). GPTQ: Accurate Post-Training Quantization for Generative Pre-trained Transformers. https://arxiv.org/abs/2210.17323.
11. Sennrich, Rico; Haddow, Barry; Birch, Alexandra. (2015). Neural Machine Translation of Rare Words with Subword Units. https://arxiv.org/abs/1508.07909.
12. 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.
13. Mikolov, Tomas; Chen, Kai; Corrado, Greg; Dean, Jeffrey. (2013). Efficient Estimation of Word Representations in Vector Space. https://arxiv.org/abs/1301.3781.
14. 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.
15. Xiao, Guangxuan; Lin, Ji; Seznec, Mickael; Wu, Hao; Demouth, Julien; Han, Song. (2022). SmoothQuant: Accurate and Efficient Post-Training Quantization for Large Language Models. https://arxiv.org/abs/2211.10438.
16. Dettmers, Tim; Lewis, Mike; Belkada, Younes; Zettlemoyer, Luke. (2022). LLM.int8(): 8-bit Matrix Multiplication for Transformers at Scale. https://arxiv.org/abs/2208.07339.

17. Wang, Junhao; Xu, Guoxi; Tong, Huimin; Xu, Gaochao; Chen, Yiwei. (2026). Dynamic Fusion for Large Language Models Compression. ICASSP 2026, 4711-4715. https://doi.org/10.1109/icassp55912.2026.11464073.

18. Biderman, Stella; Schoelkopf, Hailey; Anthony, Quentin; Bradley, Herbie; O'Brien, Kyle; Hallahan, Eric; Khan, Mohammad Aflah; Purohit, Shivanshu; Prashanth, USVSN Sai; Raff, Edward; Skowron, Aviya; Sutawika, Lintang; van der Wal, Oskar. (2023). Pythia: A Suite for Analyzing Large Language Models Across Training and Scaling. https://arxiv.org/abs/2304.01373.

19. Wolf, Thomas; Debut, Lysandre; Sanh, Victor; Chaumond, Julien; Delangue, Clement; Moi, Anthony; Cistac, Pierric; Rault, Tim; Louf, Rémi; Funtowicz, Morgan; Davison, Joe; Shleifer, Sam; von Platen, Patrick; Ma, Clara; Jernite, Yacine; Plu, Julien; Xu, Canwen; Scao, Teven Le; Gugger, Sylvain; Drame, Mariama; Lhoest, Quentin; Rush, Alexander M.. (2019). HuggingFace's Transformers: State-of-the-art Natural Language Processing. https://arxiv.org/abs/1910.03771.

20. Hugging Face contributors, . (2025). Tokenizers v0.22.1 [Software]. https://github.com/huggingface/tokenizers/releases/tag/v0.22.1.

21. 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.

22. 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.

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

24. Matthews, B.W.. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochimica et Biophysica Acta (BBA) - Protein Structure. https://doi.org/10.1016/0005-2795(75)90109-9.

25. Pennington, Jeffrey; Socher, Richard; Manning, Christopher D.. (2014). GloVe: Global Vectors for Word Representation. https://aclanthology.org/D14-1162/.

26. 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.

27. Paszke, Adam; Gross, Sam; Massa, Francisco; Lerer, Adam; Bradbury, James; Chanan, Gregory; Killeen, Trevor; Lin, Zeming; Gimelshein, Natalia; Antiga, Luca; Desmaison, Alban; Köpf, Andreas; Yang, Edward; DeVito, Zach; Raison, Martin; Tejani, Alykhan; Chilamkurthy, Sasank; Steiner, Benoit; Fang, Lu; Bai, Junjie; Chintala, Soumith. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. https://arxiv.org/abs/1912.01703.

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. Cook, R. Dennis. (1977). Detection of Influential Observation in Linear Regression. Technometrics. https://doi.org/10.1080/00401706.1977.10489493.

30. 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.