Decision Thresholds and Policy
Volume 1 · Number 1 · 2026

A Detectable Loss Can Pass a Compression Tolerance: Local Small-Model Evidence

George Knight[1] · Dominic Young[2] · Ryan King[3]

[1] Austria · [2] Belgium · [3] Czechia

Original publication: November 5, 2026 | Article 620

Abstract

Eight-bit internal-weight rounding increased next-token loss by 0.004388 nats, with a 95% interval of 0.003435-0.005341, yet passed all three tolerances specified for this experiment. Four-bit rounding increased loss by 0.693153 nats and failed every screen. These results came from actual local forward scoring of 16,384 targets in one Tiny Shakespeare segment using Pythia-14M. We reconstructed twenty-four internal matrices while retaining source embeddings, output weights, biases, and normalization parameters. The declared screens considered mean loss, large token-level increases, and block-level exceedances. Their agreement here does not establish that the criteria are interchangeable. The experiment illustrates the difference between detecting numerical degradation and deciding whether its magnitude meets a stated tolerance.

Actual CPU scoring of a public Pythia-14M checkpoint after rounding internal matrix weights, with three illustrative numerical screening rules.

A Detectable Loss Can Pass a Compression Tolerance: Local Small-Model Evidence

Actual CPU scoring of a public Pythia-14M checkpoint after rounding internal matrix weights, with three illustrative numerical screening rules.

Abstract

Eight-bit internal-weight rounding increased next-token loss by 0.004388 nats, with a 95% interval of 0.003435-0.005341, yet passed all three tolerances specified for this experiment. Four-bit rounding increased loss by 0.693153 nats and failed every screen. These results came from actual local forward scoring of 16,384 targets in one Tiny Shakespeare segment using Pythia-14M. We reconstructed twenty-four internal matrices while retaining source embeddings, output weights, biases, and normalization parameters. The declared screens considered mean loss, large token-level increases, and block-level exceedances. Their agreement here does not establish that the criteria are interchangeable. The experiment illustrates the difference between detecting numerical degradation and deciding whether its magnitude meets a stated tolerance.

1. Detecting a difference and accepting a policy are separate decisions

A numerical compression experiment can detect a small increase in prediction loss even when that increase lies below a chosen practical tolerance. Treating every nonzero difference as a failure conflates a statistical question with a decision rule. The opposite error is also possible: an acceptable average could conceal large losses on particular targets or contexts. An evaluation therefore needs both a clearly defined numerical endpoint and an explicit statement of what its acceptance criterion means.

We ask whether two fixed internal-weight rounding policies meet three declared tolerances on the same scored text. The primary endpoint is mean next-token negative log likelihood increase relative to the unchanged source model. The supplementary endpoints are the fractions of targets and context blocks exceeding specified loss increases. We also estimate uncertainty around the paired block means. The analysis permits the three screens to agree or disagree; it does not presuppose that an apparent reversal must be found.

The compression problem represented by Wang et al., 2026 motivates distinguishing a numerical degradation from an acceptance decision. Our experiment is a simpler controlled operation on a small public language model. We do not reproduce that method, fine-tune a model, or estimate its inference speed. The decision rules are illustrative quantities fixed for this local design. They are not published deployment standards or a recommendation that a particular task should tolerate the same amount of degradation.

2. A small checkpoint that can be scored locally

We use the final standard-Pile Pythia-14M checkpoint associated with Biderman et al., 2023. Its six layers have hidden width 128, four attention heads, and intermediate width 512. The input and output embedding matrices remain distinct, and the output vocabulary contains 50,304 entries. All source tensors are read from the numeric safetensors file. Their float16 values are promoted to float32 for execution; the baseline is therefore a float32 forward pass of the preserved source values, rather than a benchmark of native float16 hardware.

The local execution uses Paszke et al., 2019 through PyTorch 2.6.0+cpu and the GPT-NeoX implementation in Wolf et al., 2019, version 4.57.1. We load the complete state dictionary strictly and check that no expected tensors are missing or unexpected. The model is in evaluation mode, with caching disabled and eager attention selected. No downloaded model-specific code or hosted inference endpoint is used. Random initialization is overwritten by the complete checkpoint before any recorded policy evaluation.

For each policy the source state is reloaded afresh, preventing one rounding procedure from operating on another procedure's already modified weights. Four CPU threads are used. Repeating the first evaluated forward pass reproduces every per-target loss exactly in this environment. That is a local numerical check, not a promise that every software and hardware combination produces byte-identical output. We preserve library versions and numerical settings to support a subsequent reproduction with appropriate tolerances.

3. One text segment and a restricted context design

The public Tiny Shakespeare file in Karpathy, 2026 contains 340,240 tokens under the captured official tokenizer. We use Hugging Face contributors, 2025 version 0.22.1 with no inserted special tokens and verify that decoding the entire identifier sequence reconstructs the text exactly. The evaluation begins at zero-based token offset 170,120, the midpoint of the full token stream. From there we take the inputs needed to score 16,384 next-token targets. This fixed sampling rule does not select a segment based on model losses.

The targets are divided into sixty-four consecutive blocks of 256. Context starts again for each block. Adjacent blocks share one boundary token as an input, but their scored target positions do not overlap. Each target receives only the preceding context within its current block, so these scores differ from scoring the same segment with an uninterrupted long context. All 50,304 output logits are retained in the probability normalization, including matrix entries without ordinary token occurrences in this text.

There is no generated continuation: each observed next token supplies the target for a forward score. The text may overlap the model's original pretraining corpus. Starting at the file midpoint establishes a local evaluation rule, not certified unseen training data. Validation principles in Roberts et al., 2017 make that distinction important. Neither the number of scored tokens nor the use of an official tokenizer establishes independent-corpus generalization, new-model replication, or coding-assistant reliability.

4. Two rounding policies on the same internal matrices

We select twenty-four attention and multilayer-perceptron projection matrices, four in each transformer layer. Together they contain 1,179,648 affected weights. The matrix values are laid out in row-major groups of 128. Each group's maximum absolute value determines a float32 scale. The four-bit policy rounds divided values to signed integers from minus seven through seven; the eight-bit policy uses minus 127 through 127. Reconstructed values are the resulting codes multiplied by their stored scales.

These conventions leave one nominal signed code unused at each precision. Input embeddings, output embeddings, biases, and normalization parameters keep their source values. We therefore do not describe the experiment as quantizing the entire fourteen-million-parameter model to four or eight bits. The reconstructed matrices are executed densely in float32. No actual packed low-bit kernel, compressed model file, device memory reduction, or throughput improvement is measured by this procedure.

The broader methods in Frantar et al., 2022, Lin et al., 2023, Xiao et al., 2022, Dettmers et al., 2022, and Dettmers et al., 2023 place weight precision within a richer calibration and implementation context. Our simple grouped maximum rule is not an implementation of any of them. Its role is to create two

transparent numerical perturbations whose downstream loss can actually be scored. It would be inappropriate to attach those methods' performance claims to the arrays reconstructed here or infer that their optimized policies would incur the same loss changes.

5. Prediction loss and numerical checks

For every target we compute negative log likelihood in natural-log units. The paired increase is the rounded-policy loss minus the source-model loss for that identical target and block context. Averaging within a block gives one comparison row. The source model's mean loss is 4.717160753 nats on this segment, corresponding to perplexity 111.850231. This is a description of the specified model and context rule, not a general quality score for all Shakespeare text or an assessment of generated prose.

As an implementation check, per-target cross entropy is compared with the negative log-softmax entry for the observed identifier. The maximum discrepancy is zero in the checked first block for all three model states. The repeated first-block forward losses also have maximum discrepancy zero. Numerical considerations in Higham, 2002 and Goldberg, 1991 motivate preserving the arithmetic and scale precision. These checks confirm the executed score calculation; they do not validate language understanding, safety, or downstream usefulness.

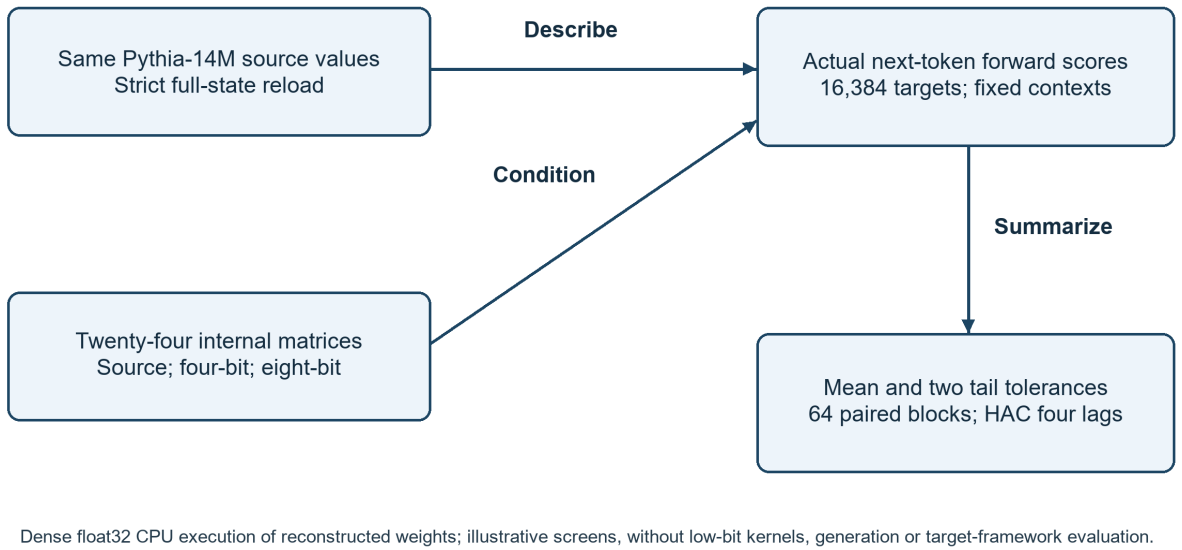

Figure 1. The same checkpoint and scored targets enter source, four-bit, and eight-bit matrix states; paired losses are summarized before applying declared mean and tail tolerances.

Predictive scoring principles in Gneiting and Raftery, 2007 and the accuracy-measure discussion in Hyndman and Koehler, 2006 help distinguish the chosen loss from task correctness or a hardware metric. Calibration work in Guo et al., 2017 raises another separate question about predictive confidence. We do not estimate calibration error, sample an answer, or ask human raters to judge outputs. The completed endpoint is an observed-target score under a fixed context policy, with its denominator and information boundary stated directly.

6. Three screens specified as illustrations

The mean screen accepts a policy when its average loss increase is at most 0.05 nat per target. The token-tail screen accepts when no more than five percent of targets incur an increase greater than 0.5 nat. The block-tail screen accepts when no more than five percent of the 256-target blocks have an average increase greater than 0.05 nat. The rules operate on different aspects of the same paired-loss distribution. They are intentionally explicit rather than described vaguely as negligible degradation.

Decision-cost reasoning in Elkan, 2001 and selective-decision ideas in Geifman and El-Yaniv, 2017 motivate stating how a numerical score maps to an action. We do not estimate an application's costs or optimize these thresholds from outcomes. Their values have no established clinical, legal, financial, or software-reliability meaning. A deployment decision would need its own loss-to-consequence relationship. Here acceptance means only that the observed numerical policy satisfies the stated illustrative inequality on this segment.

For uncertainty in the mean contrast, each policy receives a separate intercept-only ordinary least-squares fit over the sixty-four block means. Covariance follows Newey and West, 1987 with four retained-block lags, a finite-sample correction, and a t reference with sixty-three degrees of freedom. Forecast-comparison reasoning in Diebold and Mariano, 1995 and West, 1996 supports pairing losses on identical evaluation units. The mean-only R-squared is zero, and no contextual controls or predictor coefficients are fitted.

Table 1. Mean next-token loss increase in nats under internal-matrix rounding, relative to the source-valued baseline.

Model	Term	Estimate (SE)	95% CI	N
M1	Mean outcome/contrast	0.693153 (0.008846)	[0.675476, 0.710829]	64
M1: fit	R-squared = 0.0000	-	-	64
M2	Mean outcome/contrast	0.004388 (0.000477)	[0.003435, 0.005341]	64
M2: fit	R-squared = 0.0000	-	-	64

Model labels: M1 = Four-bit mean next-token NLL increase; M2 = Eight-bit mean next-token NLL increase.

7. Eight-bit loss is detectable but remains below each screen

The eight-bit mean increase is 0.004388178 nat per target. Its standard error is 0.000476818, and its 95% interval is 0.003435334 to 0.005341023. The interval lies above zero, so a zero-difference interpretation receives little support under this conditional covariance approximation. It also lies well below the mean tolerance of 0.05. Detecting a nonzero loss is therefore not the same as failing the specified acceptance screen.

Only two of the 16,384 targets have an eight-bit increase above 0.5 nat, a fraction of 0.000122070. No block has a mean increase above 0.05. The eight-bit policy consequently passes the mean, token-tail, and block-tail screens in this experiment. These finite-sample tail counts are direct calculations, not confidence bounds for tail frequencies in a future corpus. Passing them does not establish that every target is unaffected or that a safety-critical event remains equally likely.

The pooled squared weight-reconstruction error divided by source weight energy is approximately 0.000027912 for the affected matrices. That parameter-space quantity is supplied as context for the intervention, not as a substitute for the measured loss. Its numerical scale differs from nats per predicted

token. No conversion factor is established between the two endpoints, and a low reconstruction error alone would not justify the acceptance conclusion without actually computing the scored-target outcomes.

8. Four-bit loss is far from the selected tolerances

The four-bit mean increase is 0.693152926 nats, with standard error 0.008845558 and interval 0.675476490 to 0.710829363. Its parameter-space normalized error is about 0.009281917 over the affected matrices. Despite leaving the large embedding and output matrices unchanged, this intervention produces a substantial deterioration in the scored next-token distribution. The result concerns this grouped policy on this model; it is not evidence that all four-bit language-model compression must incur comparable damage.

For the four-bit policy, 50.079 percent of targets exceed the half-nat increase threshold and every block exceeds the 0.05 mean-increase threshold. It therefore fails all three declared screens. These are not isolated extreme tokens determining a tiny annual or corpus average. The observed block-level condition fails throughout this selected segment. We do not infer that every individual target worsens or that a specific linguistic phenomenon explains the loss, since neither statement follows from the reported tail indicators.

Table 2. Observed screen quantities and the resulting illustrative decisions.

Policy	Mean increase, nats	Fraction of targets above 0.5	Fraction of blocks above 0.05	Mean / token-tail / block-tail
Four-bit internal matrices	0.693153	0.500793	1.000000	Fail / fail / fail
Eight-bit internal matrices	0.004388	0.000122	0.000000	Pass / pass / pass

The three screens agree for both evaluated policies. We retain this agreement rather than shifting thresholds until a disagreement appears. Agreement for two clear cases does not establish logical equivalence among mean and tail criteria. A different distribution could produce different decisions. The experiment's supported contrast is instead between a statistically detectable eight-bit loss that passes the declared tolerance and a much larger four-bit loss that does not.

9. What the uncertainty does and does not cover

Blocks follow one another in a single literary text, so they are not independent samples from many domains. The dependence issues in Cameron and Miller, 2015 and the distinction from heteroskedasticity correction alone in White, 1980 remain relevant even though HAC is applied. Four lags are a declared approximation, not proof that all discourse-level dependence has been captured. The interval does not include different pretraining runs, model scales, corpora, block lengths, or alternative rounding rules that were not executed.

Influence reasoning in Cook, 1977 would motivate inspecting influential contexts, but no block is removed from the mean. Multiplicity concerns in Benjamini and Hochberg, 1995 limit interpreting several screens and intervals as an unconstrained search for a successful policy. Guidance in Wasserstein and Lazar, 2016 also reinforces that a small tail probability for a zero-loss null is not a practical rejection rule. The observed eight-bit result is a concrete example: its detectable degradation remains small relative to this experiment's chosen tolerance.

Model-selection bias discussed in Varma and Simon, 2006 matters if the same segment is subsequently used to tune scales or acceptance thresholds. That would change its role from evaluation evidence to

development evidence. We do not perform such tuning. A stronger assessment would preserve independent text sources and evaluate application-specific endpoints, including rare failures when those matter. Actual forward scoring is necessary evidence here, but it is insufficient to certify a useful compressed deployment or generalize to a larger language model.

10. Report the tolerance alongside the measured loss

The completed local experiment gives a simple decision lesson. Eight-bit internal rounding has a nonzero measured cost yet satisfies the three stated numerical screens; four-bit rounding violates them by a wide margin. The result does not require declaring zero degradation, inventing a hardware speedup, or forcing disagreement among screening definitions. It requires keeping source model values, arithmetic, context length, comparison unit, and tolerance meaning visible to the reader.

Documentation in the spirit of Gebru et al., 2021 records the exact text segment, tokenizer, included logits, original and affected matrices, precision rules, and absence of certified pretraining separation. Preservation according to Wilkinson et al., 2016 links the numeric checkpoint, runtime versions, per-target losses, and paired block summaries. These materials support a reproducible small-model sensitivity study. Whether its tolerances are useful for a particular application remains an additional scientific and operational question beyond the experiment reported here.

References

1. 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.
2. 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.
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. 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.
5. 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.
6. 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.
7. 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.
8. Higham, Nicholas J.. (2002). Accuracy and Stability of Numerical Algorithms: Second Edition. Society for Industrial and Applied Mathematics. https://doi.org/10.1137/1.9780898718027.
9. 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.
10. 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.

11. 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.

12. 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.

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

14. Goldberg, David. (1991). What every computer scientist should know about floating-point arithmetic. ACM Computing Surveys. https://doi.org/10.1145/103162.103163.

15. 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.

16. 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.

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

18. 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.

19. Karpathy, Andrej. (2026). Tiny Shakespeare text, char-rnn repository [Data file; snapshot retrieved 7 November 2026]. https://github.com/karpathy/char-rnn/blob/master/data/tinyshakespeare/input.txt.

20. 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.

21. 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.

22. 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.

23. 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.

24. 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.

25. West, Kenneth D.. (1996). Asymptotic Inference about Predictive Ability. Econometrica. https://doi.org/10.2307/2171956.

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

27. 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.

28. 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.

29. 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.

30. Diebold, Francis X.; Mariano, Roberto S.. (1995). Comparing Predictive Accuracy. Journal of Business & Economic Statistics. https://doi.org/10.1080/07350015.1995.10524599.