Journal of Measurement and Evidence
Volume 1 · Number 1 · 2026

Weight Reconstruction Is Not Projection Fidelity: A Four-Bit Grouping Audit

Stephen Anderson[1] · Samuel Nelson[2] · Matthew Fisher[3]

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

Original publication: October 2, 2026 | Article 707

Abstract

Five of twenty-four Pythia-14M projection matrices preferred different four-bit grouping orientations depending on whether error was measured in weights or in outputs on observed inputs. Row grouping had lower average error under both criteria, but that aggregate agreement concealed matrix-specific reversals. Inputs came from 2,048 actual model positions in one public text segment. The result identifies a limitation of weight-only selection: a parameter discrepancy does not specify which input directions it perturbs. Local projection fidelity also remains distinct from the behavior of the fully modified language model.

Actual reference-model inputs with local numerical weight reconstruction. The backbone remains unchanged; no compressed-model next-token accuracy, low-bit kernel, or target compression architecture is evaluated.

Weight Reconstruction Is Not Projection Fidelity: A Four-Bit Grouping Audit

> Actual reference-model inputs with local numerical weight reconstruction. The backbone remains unchanged; no compressed-model next-token accuracy, low-bit kernel, or target compression architecture is evaluated.

Abstract

Five of twenty-four Pythia-14M projection matrices preferred different four-bit grouping orientations depending on whether error was measured in weights or in outputs on observed inputs. Row grouping had lower average error under both criteria, but that aggregate agreement concealed matrix-specific reversals. Inputs came from 2,048 actual model positions in one public text segment. The result identifies a limitation of weight-only selection: a parameter discrepancy does not specify which input directions it perturbs. Local projection fidelity also remains distinct from the behavior of the fully modified language model.

1. A small weight error can act in the wrong direction

Numerical compression replaces a parameter matrix with an approximation. An elementary check measures the squared discrepancy between the original and reconstructed weights. That check is useful for detecting reconstruction problems, yet a matrix is used through multiplication by inputs. Two equally sized weight discrepancies can act differently when input directions are unevenly excited or correlated. A decision based on weight error alone may therefore choose a different approximation from a decision based on the resulting projection error.

We investigate whether the preferred grouping orientation agrees under these two measurements for actual small-language-model inputs. The comparison uses two transparent four-bit reconstructions: consecutive groups in the matrix's row-major order and consecutive groups after transposition. Both use 128 weights per scale and the same symmetric quantization rule. We measure relative squared error in parameter space and in bias-free local projection outputs, retaining original-model inputs for both policies.

Wang et al., 2026 concerns dynamic fusion for large-language-model compression. That problem motivates checking whether a compressed representation's numerical fidelity corresponds to the function it is intended to preserve. Our experiment does not implement the paper's fusion mechanism, evaluate its released models, or estimate its quality-efficiency tradeoff. It offers a bounded diagnostic for the choice of a measurement used around a compression decision.

The central proposition is that a low weight error need not select the same policy as low projection error. A second question asks whether an average policy advantage hides individual matrices where the preferred orientation changes. These are finite comparisons within one checkpoint and one input segment. Neither proposition requires an inference that the sampled text represents all language-model uses.

2. Function-aware compression has several meanings

GPTQ Frantar et al., 2022 incorporates information beyond a uniform weight discrepancy in post-training quantization. AWQ Lin et al., 2023 emphasizes activation-aware choices, while SmoothQuant Xiao et al., 2022 addresses the relationship between weight and activation ranges. These methods motivate measuring the inputs relevant to a matrix. Our two policies do not reproduce their calibration, transformations, or optimization procedures. The comparison instead keeps the reconstruction rule deliberately simple and checks its consequences in two spaces.

LLM.int8() Dettmers et al., 2022, ZeroQuant Yao et al., 2022, and OmniQuant Shao et al., 2023 provide further examples of numerical compression strategies with different assumptions and implementation details. A conclusion about row-major groups in this study is not an empirical comparison against those systems. The source checkpoint, quantizer, group size, precision of the calculation, and error normalization must all remain attached to the reported result.

Pruning work such as SparseGPT Frantar and Alistarh, 2023 and Wanda Sun et al., 2023 also illustrates why a parameter change's function can matter. However, pruning and rounding are different interventions. We introduce no sparsity objective or removed parameter mask. FlashAttention Dao et al., 2022 and the Roofline perspective Williams et al., 2009 address other computational concerns, reinforcing that numerical fidelity, algorithmic efficiency, and measured hardware performance require different evidence. This study measures only the first of those concerns in two local forms.

Measurement analysis Carroll et al., 2006 helps distinguish the two outcomes. Parameter error is a property of a source matrix and its reconstruction. Projection error also depends on the inputs and on the output-energy denominator. Agreement between them is an empirical question. Agreement-analysis principles Bland and Altman, 1986 caution that a positive relationship between two summaries does not automatically make them interchangeable decision criteria.

3. A fixed checkpoint and actual reference inputs

We load the final public Pythia-14M checkpoint described by Biderman et al., 2023 and retain its original weights throughout input collection. The local source file and configuration have recorded checksums. The backbone runs as a float32 CPU model using PyTorch Paszke et al., 2019 and Transformers Wolf et al., 2019, with eager attention and caching disabled. No compressed parameter is inserted into the backbone during this stage.

The public Tiny Shakespeare text snapshot Karpathy, 2026 is encoded with the checkpoint's tokenizer into 340,240 tokens. We start at zero-based input position 170,120 and use the following 2,048 positions. The sequence is divided into sixteen contexts of 128 inputs, each evaluated independently. This context-reset rule is part of the workload, because the resulting activations need not equal those from one uninterrupted 2,048-token context.

Forward pre-hooks capture inputs to twenty-four internal linear modules: the attention query-key-value and attention output projections, and the two MLP projections in each of six layers. Each module therefore receives 2,048 captured input rows. The dimensions differ between module types. We retain matrix-specific denominators rather than treating all entries across modules as observations of one common scalar output.

This text segment is locally separate from the start of the source text, but it is not claimed to be absent from original pretraining. No document-level training audit is available here. The corpus contains one

literary domain, not repository code, tool interactions, or a representative sample of model use. The input collection establishes an actual executed workload while leaving broad transport claims unresolved.

4. Matched reconstructions and two error spaces

For row grouping, the source matrix is flattened in row-major order into groups of 128 consecutive values. Each group's scale is its maximum absolute weight divided by seven. We round scaled values to the nearest integer, restrict codes to -7 through +7, and reconstruct with the group scale. Column grouping applies exactly the same rule to the transposed matrix and then restores the original orientation. All compared dimensions are compatible with complete groups.

Both policies have the same number of groups and nominal code-and-scale budget for a given matrix. We do not pack codes into bytes or benchmark a low-bit implementation. Quantization calculations use float64 representations of the same source weight values. This choice supports a clear numerical audit but differs from a deployed low-precision kernel's rounding path. Numerical-stability guidance Higham, 2002 and floating-point analysis Goldberg, 1991 motivate preserving those details rather than labeling all implementations simply four-bit.

Relative weight error is the sum of squared reconstructed-minus-source weights divided by source weight energy. Relative projection error multiplies the same weight difference by the actual original inputs and divides by the energy of the original bias-free projections. Biases are excluded from both the projection discrepancy and its normalizer. Thus, the functional measurement is explicitly local to a linear map and its observed input rows.

Holding the inputs fixed is essential to the scope. If quantized parameters were inserted into earlier layers, later layers would generally receive different inputs. We do not propagate such changes. The calculation isolates how the two approximations act on reference-model inputs, without claiming that those inputs are the activations of a fully compressed model. A small local error can still produce a different downstream result, and several local errors can interact.

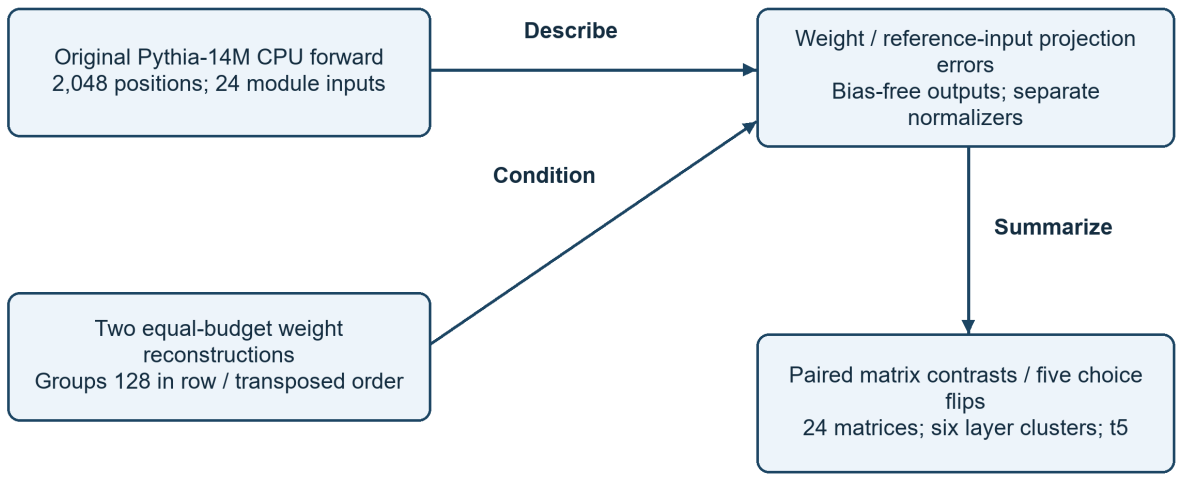

Figure 1. *Original-model inputs remain common to both grouping policies. Weight discrepancy and bias-free output discrepancy are normalized separately, then compared within each matrix and across six layer groups.*

5. Describing agreement without multiplying the sample size

For each matrix, we calculate column-minus-row error under each criterion. Two intercept-only OLS models summarize those paired differences with equal matrix weights. Standard errors are clustered by the six layers, using finite-sample correction and t intervals with five degrees of freedom. Robust covariance White, 1980 and clustered-inference guidance Cameron and Miller, 2015 inform the declared working summary. Six connected layers are not six randomly sampled independent compression systems.

The 2,048 input positions do not become the regression sample size. They define each matrix's projection error; the regressions contain twenty-four matrix observations. Likewise, a matrix with more parameters does not receive greater primary weight. A global parameter-weighted reconstruction objective or a token-loss objective would be a different aggregate. The equal-matrix summary answers whether the mean module-level relative contrast favors one orientation.

We also inspect strict orientation choices and descriptive score association. Spearman's rank association Spearman, 1904 and Kendall's ordering measure Kendall, 1938 summarize the relation between the forty-eight policy-matrix score pairs. Their values are descriptive only: the two policies within a matrix and modules within a layer are dependent. No correlation p-value is interpreted as evidence of measurement equivalence or replacement validity.

NumPy Harris et al., 2020 supports the finite summaries, while the statistical numerical stack uses SciPy Virtanen et al., 2020 for distribution calculations. The retained analysis contains matrix dimensions, each policy's two error values, and every orientation switch. This makes a mismatch inspectable at its source rather than reducing it to an unexplained global correlation.

6. Results: aggregate agreement, local reversals

Row grouping has mean relative weight error 0.014518, compared with 0.022117 for column grouping. Their mean column-minus-row difference is 0.007599. The corresponding projection errors are 0.009192 and 0.014201, giving a difference of 0.005009. Thus, the same orientation is favored in both equal-matrix averages. The two differences are in separately normalized quantities and should not be subtracted as if they measured an identical scale.

Table 1. Column-minus-row relative errors across twenty-four projections.

Model	Term	Estimate (SE)	95% CI	N
M1	Mean outcome/contrast	0.007599 (0.001495)	[0.003756, 0.011442]	24
M1: fit	R-squared = 0.0000	-	-	24
M2	Mean outcome/contrast	0.005009 (0.001698)	[0.000645, 0.009373]	24
M2: fit	R-squared = 0.0000	-	-	24

Model labels: M1 = Mean column-minus-row relative weight error; M2 = Mean column-minus-row reference-input projection error.

Model 1's layer-clustered 95% interval is [0.003756, 0.011442]; Model 2's is [0.000645, 0.009373]. Both models use twenty-four observations, one intercept, six clusters, and a t5 reference distribution. Their R-squared values are effectively zero. The intervals are conditional working summaries over related modules and do not quantify uncertainty across checkpoints, text domains, or training runs.

Weight error prefers column grouping in six matrices; projection error prefers it in three. Five matrices change their preferred orientation between criteria. Four switch from a column preference under weight error to a row preference under projection error, while one switches in the other direction. This finite count is not a five-out-of-twenty-four estimate of a universal failure probability for weight-based selection.

For a concrete example, layer two's MLP down-projection has row and column weight errors of approximately 0.016640 and 0.015609. The column policy therefore looks better in weight space. On the captured inputs, the projection errors are about 0.006027 and 0.015577, reversing the choice. In layer four's attention output projection, the reverse pattern occurs: row grouping has smaller weight error, but column grouping has smaller projection error.

Across all forty-eight policy-matrix pairs, descriptive Spearman association is 0.5058 and Kendall tau-b is 0.3741. These positive associations coexist with the five within-matrix choice reversals. They describe broad ordering across heterogeneous modules, while the practical selection question compares two approximations for the same module. The distinction explains why a correlation summary alone can conceal the decision disagreement of interest.

7. How the two criteria can disagree

Weight error treats source parameter entries uniformly within a matrix's energy normalization. Projection error weights combinations of those entries through the observed input geometry. An error aligned with a frequently excited input direction can matter more than an equally large error in a rarely excited direction. Correlated input features can also create cancellation or reinforcement. The experiment observes the resulting discrepancy without claiming to identify a single causal activation feature responsible for each switch.

This interpretation does not make local projection error a complete language-model quality metric. The reference-input design omits downstream activation changes, attention interactions after replacement, residual accumulation, and output-token decisions. A function-aware selection criterion would still require validation against the eventual application loss. The present evidence supports measuring more than one stage, rather than replacing one unqualified scalar with another.

Validation cautions Varma and Simon, 2006 are relevant if the same text segment is used repeatedly to choose quantization rules. Here the two policies are specified without searching the held projection results for new clipping choices. The ASA statement Wasserstein and Lazar, 2016 similarly supports emphasizing the finite mismatches and conditional scope rather than interpreting a conventional significance threshold as a compression-quality certificate. No operating tolerance or deployment policy is selected after seeing the intervals.

Dataset documentation Gebru et al., 2021, FAIR principles Wilkinson et al., 2016, and technical-debt arguments Sculley et al., 2015 motivate binding the result to the checkpoint, tokenizer, text interval, hook locations, quantizer, and precision conventions. Reproducing only the nominal bit count would leave too many conditions unspecified. Future studies should evaluate independent text domains and fully modified-model outputs before drawing a recommendation about grouping orientation.

There is a useful special case where the two relative errors coincide. If the observed inputs' uncentered second-moment matrix is proportional to the identity, multiplication weights every input direction equally. Under that condition, the proportional factor cancels from the numerator and denominator of the relative projection error, leaving the relative weight error. This is a mathematical condition for agreement, not an extra isotropic-input experiment or an assumption imposed on the captured activations.

The orientation switches demonstrate that the empirical criterion cannot be replaced by that special-case argument in every matrix here. They do not show that input anisotropy is the only practical obstacle to compression. Bias normalization, nonlinear operations, later layers, and application-specific losses can create additional distinctions. The value of the local audit is that its limited measurement is explicit: it tests one projection relationship while leaving those other questions for separate executed evaluations.

8. Conclusion

Relative weight error and reference-input projection error favor row grouping on average, yet select different orientations for five individual matrices. Parameter reconstruction is therefore informative but incomplete evidence for local functional fidelity. The actual-input audit makes that limitation visible while preserving a further boundary: local projection fidelity does not establish the end-to-end behavior or efficiency of a compressed language model.

References

1. 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.
2. 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.
3. 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.
4. 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.
5. Frantar, Elias; Alistarh, Dan. (2023). SparseGPT: Massive Language Models Can Be Accurately Pruned in One-Shot. https://arxiv.org/abs/2301.00774.
6. Sculley, D., et al. "Hidden Technical Debt in Machine Learning Systems." *Advances in Neural Information Processing Systems*, vol. 28, 2015. https://papers.nips.cc/paper/2015/hash/86df7dcfd896fcaf2674f757a2463eba-Abstract.html
7. Dao, Tri; Fu, Daniel Y.; Ermon, Stefano; Rudra, Atri; Ré, Christopher. (2022). FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness. https://arxiv.org/abs/2205.14135.
8. Shao, Wenqi; Chen, Mengzhao; Zhang, Zhaoyang; Xu, Peng; Zhao, Lirui; Li, Zhiqian; Zhang, Kaipeng; Gao, Peng; Qiao, Yu; Luo, Ping. (2023). OmniQuant: Omnidirectionally Calibrated Quantization for Large Language Models. https://arxiv.org/abs/2308.13137.
9. Goldberg, David. (1991). What every computer scientist should know about floating-point arithmetic. ACM Computing Surveys. https://doi.org/10.1145/103162.103163.
10. 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.
11. Williams, Samuel; Waterman, Andrew; Patterson, David. (2009). Roofline: an insightful visual performance model for multicore architectures. Communications of the ACM. https://doi.org/10.1145/1498765.1498785.
12. 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.
13. 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.

14. 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.

15. 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.

16. KENDALL, M. G.. (1938). A NEW MEASURE OF RANK CORRELATION. Biometrika. https://doi.org/10.1093/biomet/30.1-2.81.

17. 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.

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

19. Harris, Charles R.; Millman, K. Jarrod; van der Walt, Stéfan J.; Gommers, Ralf; Virtanen, Pauli; Cournapeau, David; Wieser, Eric; Taylor, Julian; Berg, Sebastian; Smith, Nathaniel J.; Kern, Robert; Picus, Matti; Hoyer, Stephan; van Kerkwijk, Marten H.; Brett, Matthew; Haldane, Allan; del Río, Jaime Fernández; Wiebe, Mark; Peterson, Pearu; Gérard-Marchant, Pierre; Sheppard, Kevin; Reddy, Tyler; Weckesser, Warren; Abbasi, Hameer; Gohlke, Christoph; Oliphant, Travis E.. (2020). Array programming with NumPy. Nature. https://doi.org/10.1038/s41586-020-2649-2.

20. Yao, Zhewei; Aminabadi, Reza Yazdani; Zhang, Minjia; Wu, Xiaoxia; Li, Conglong; He, Yuxiong. (2022). ZeroQuant: Efficient and Affordable Post-Training Quantization for Large-Scale Transformers. https://arxiv.org/abs/2206.01861.

21. Sun, Mingjie; Liu, Zhuang; Bair, Anna; Kolter, J. Zico. (2023). A Simple and Effective Pruning Approach for Large Language Models. https://arxiv.org/abs/2306.11695.

22. 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.

23. Spearman, C.. (1904). The Proof and Measurement of Association between Two Things. The American Journal of Psychology. https://doi.org/10.2307/1412159.

24. 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.

25. Bland, J. Martin, and Douglas G. Altman. "Statistical Methods for Assessing Agreement between Two Methods of Clinical Measurement." *The Lancet*, vol. 327, no. 8476, 1986, pp. 307-310. https://doi.org/10.1016/S0140-6736(86)90837-8.

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

27. 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.

28. 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.

29. 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.

30. 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.