Resource Allocation and Optimization
Volume 1 · Number 1 · 2026

The Same Bit Budget, Different Errors: Grouping Orientation in a Small Language-Model Checkpoint

Robert Cooper[1] · Jeremy Hayes[2] · Ryan King[3]

[1] Canada · [2] Germany · [3] New Zealand

Original publication: October 31, 2026 | Article 561

Abstract

Grouping weights by storage order is a compression choice with measurable consequences. We quantized twenty-four projection matrices from the final Pythia-14M checkpoint using equally sized groups of 128 weights, arranged along either matrix orientation. Both policies used the same four-bit code convention and scale budget. Column-oriented groups had a mean log distortion ratio of 0.363 relative to row-oriented groups, with a six-layer-clustered interval of 0.231 to 0.495. The direction nevertheless varied across matrix families. These are weight-reconstruction errors in one small trained model, not token-quality, speed, or large-model compression results. Equal nominal precision does not specify an equivalent numerical approximation.

Exploratory numerical study of a public checkpoint. Functional model evaluation and human publication review remain pending.

The Same Bit Budget, Different Errors: Grouping Orientation in a Small Language-Model Checkpoint

Exploratory numerical study of a public checkpoint. Functional model evaluation and human publication review remain pending.

Abstract

Grouping weights by storage order is a compression choice with measurable consequences. We quantized twenty-four projection matrices from the final Pythia-14M checkpoint using equally sized groups of 128 weights, arranged along either matrix orientation. Both policies used the same four-bit code convention and scale budget. Column-oriented groups had a mean log distortion ratio of 0.363 relative to row-oriented groups, with a six-layer-clustered interval of 0.231 to 0.495. The direction nevertheless varied across matrix families. These are weight-reconstruction errors in one small trained model, not token-quality, speed, or large-model compression results. Equal nominal precision does not specify an equivalent numerical approximation.

1. Precision labels leave a grouping decision unresolved

A statement that a model is compressed to four bits does not completely define the approximation. Quantized values share scales over groups, and the grouping determines which values compete for the same representable range. When a matrix contains unequal extremes across directions, grouping along one orientation can produce different rounding errors from grouping along the other. We ask whether that difference is visible under a genuinely equal group-and-scale budget in a public trained checkpoint.

Pythia provides research checkpoints intended to support inspection of language-model behavior and scaling (Biderman et al., 2023). We use its small 14M model to keep the numerical experiment accessible and reproducible. The retrieved model card identifies the current checkpoint as trained on the standard Pile, distinguishes it from an earlier deduplicated model at the same location, and defines the final checkpoint. The local hash, rather than an unqualified live model name, identifies our input.

The compression problem addressed by Wang et al., 2026 motivates attention to such design choices. We have verified that work's official bibliographic record, but do not claim to reproduce its mechanism or reported performance. The present experiment is a simpler numerical component study. It can show that grouping matters for stored weights, while leaving open whether a more elaborate compression system preserves generation quality or realizes efficient hardware execution.

2. Approximation objectives are not interchangeable

GPTQ (Frantar et al., 2022) uses a post-training quantization framework, while activation-aware weight quantization (Lin et al., 2023) explicitly considers information beyond an unweighted reconstruction criterion. SmoothQuant (Xiao et al., 2022) and LLM.int8 (Dettmers et al., 2022) further illustrate how activation and weight behavior can influence practical low-precision computation. Our loss is deliberately narrower: mean squared error between reconstructed weights and the supplied checkpoint values, normalized by the original matrix's mean squared magnitude.

Sparse-quantized representations in Dettmers et al., 2023 and pruning approaches in Sun et al., 2023 and Ma et al., 2023 address other ways to allocate limited resources. They are not executed as competitors here. Removing weights, preserving exceptional values, altering scale assignments, and changing matrix

orientation can produce different storage and functional consequences. Treating all those decisions as a single bit-width label would hide the particular approximation being evaluated.

Adaptation methods such as LoRA (Hu et al., 2022) and preference-learning methods such as Rafailov et al., 2023 change the learning problem rather than establish the quality of a particular numerical reconstruction. We cite them to distinguish adaptation objectives from the stored-weight audit. No fine-tuning examples, preference labels, gradients, or additional training are used. The checkpoint is held fixed, and all reported differences arise from deterministic quantization operations on its supplied arrays.

3. What the checkpoint contains

The safetensors file contains seventy-six arrays and 14,067,712 parameters. We read its bounded JSON header and numeric half-precision payload directly, checking array dimensions, offsets, and finite values. No pickle object or downloaded model code is executed. Numeric values are converted to higher precision for the reconstruction calculation; that computational convenience does not mean that the original checkpoint was stored in higher precision than its recorded half-precision format.

The main analysis includes four projection families in each of six layers: attention output, attention query-key-value, feed-forward expansion, and feed-forward contraction. This yields twenty-four matrices. Embedding matrices, layer-normalization parameters, and biases are excluded. The restriction makes orientation comparisons coherent for two-dimensional projections, but also limits the scope. Most parameters in this tiny checkpoint belong to embeddings, so the reported projection result is not a whole-model storage-reduction estimate.

Matrix sizes differ across families. The attention output has 128 rows and columns, the query-key-value matrix stacks three output blocks, and feed-forward projections expand or contract the hidden dimension. We retain their actual recorded orientations rather than transposing selected families to obtain a preferred result. A family label records architectural role, while the experiment's row and column terminology always refers to the corresponding stored array.

4. Two equal-budget grouping policies

For the row-oriented policy, weights are traversed in row-major order and divided into consecutive groups of 128. For the column-oriented policy, the transpose is traversed in row-major order with the same group size. Every included matrix has a weight count divisible by 128. Consequently, both policies use exactly the same number of scales for that matrix, even when its row and column dimensions differ.

Each group uses its maximum absolute weight divided by seven as a symmetric scale, rounded to a representable float32 value before reconstruction. Values are rounded to the nearest integer code and reconstructed using that scale. The available signed codes run from minus seven to seven. Under a four-bit storage convention one code is unused. We assign one hypothetical float32 scale per group, giving 4.25 bits per weight before any unrelated file headers or alignment overhead. We do not claim to have produced an actual packed inference file.

A whole-tensor maximum-scale reconstruction is retained as a descriptive third reference. It uses fewer scales and therefore has a different budget from the two group policies. It cannot serve as an equal-cost competitor. This distinction is essential: a larger reconstruction error under one scale per tensor does not, by itself, prove that additional scales are an optimal use of real storage or hardware resources.

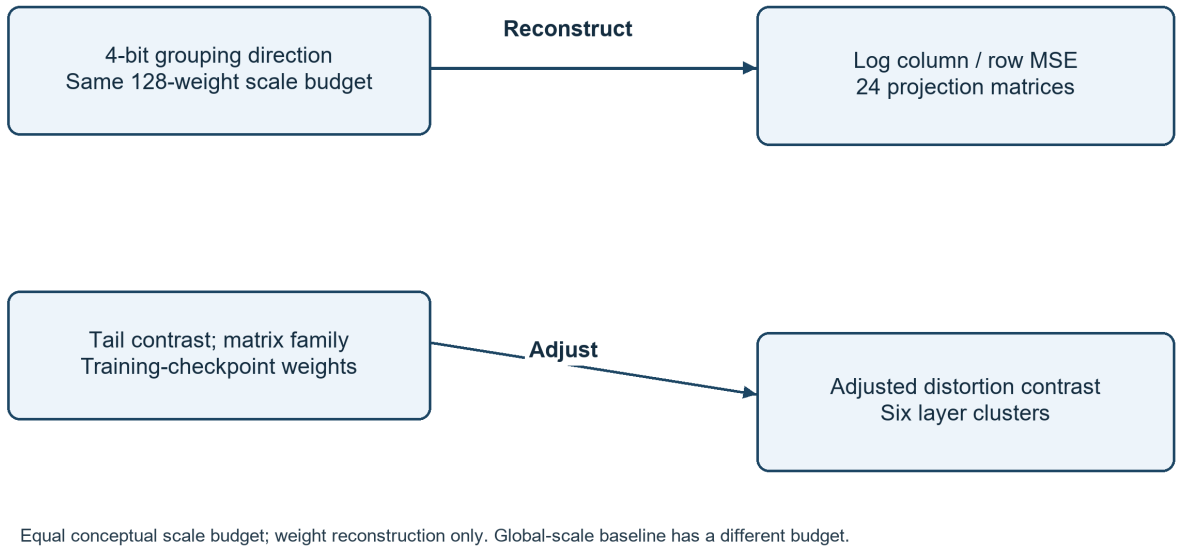


Figure 1. Equal-sized groups with different orientations produce paired matrix distortion ratios. A separate family-adjusted model examines their association with directional extreme-value summaries.

5. Distortion, directional tails, and inference

For each matrix and grouping policy, we calculate normalized mean squared weight error. Dividing by the original mean squared magnitude prevents a matrix with larger overall numerical scale from dominating solely for that reason. It does not make the criterion activation-aware. A large error in a frequently used direction may matter more functionally than a similar error elsewhere, and the current experiment contains no activation observations to assess that distinction.

The primary response is the natural logarithm of column-group distortion divided by row-group distortion. Positive values favor row grouping. Its intercept-only regression summarizes an equal-matrix mean rather than weighting matrices by parameter count. The second regression includes matrix-family indicators and a directional tail contrast. That contrast subtracts the average row maximum, relative to matrix root mean square, from the analogous column maximum. It is an observed matrix summary, not an intervention on outliers.

Covariance clusters by the six layers, with correction and t-based intervals. Cameron and Miller, 2015 explains why multiple matrices from one layer should not automatically count as independent architectural replications. White, 1980 supplies the broader heteroskedasticity-consistent foundation, but does not remove the limitation of only six clusters. These intervals describe a hypothetical layer-variation model conditional on one checkpoint. They do not quantify variation across model families, training seeds, or data corpora.

Another repeated-structure model would require an additional specification, which is not fitted here. Agreement concepts in Bland and Altman, 1986, Lin, 1989, and Koo and Li, 2016 distinguish reconstruction fidelity from agreement with a behavioral target. We report numerical error, not a reliability coefficient or proof of equivalent model behavior. The distinction remains important even when the original and reconstructed weights have similar overall magnitudes.

Table 1. Equal-budget grouping contrasts and the adjusted directional-tail association. Uncertainty clusters by six layers. The dependent variable is a log normalized weight-distortion ratio.

Model	Term	Estimate (SE)	95% CI	N
M1	Mean contrast	0.3631 (0.0512)	[0.2314, 0.4947]	24
M1: fit	R-squared = 0.0000	-	-	24
M2	Column minus row tail	0.1758 (0.0201)	[0.1240, 0.2275]	24
M2: fit	R-squared = 0.8823	-	-	24

Model labels: M1 = Mean log column-group over row-group distortion; M2 = Tail contrast association adjusted for matrix family.

6. An average orientation effect with family differences

Mean normalized matrix error is 0.01452 for row groups and 0.02212 for column groups. The mean paired log ratio is 0.3631, with standard error 0.0512 and interval 0.2314 to 0.4947. These statistics use different averaging operations: the mean of ratios on a logarithmic scale is not the ratio of the two arithmetic mean errors. We retain both summaries without treating them as interchangeable estimates.

Column grouping has lower distortion in six of the twenty-four matrices. Feed-forward contraction has similar average errors under the two orientations, with a slightly negative average log ratio. Query-key-value and feed-forward expansion matrices show larger positive average ratios. Thus the overall preference for row grouping does not establish that one orientation is best for every architectural role. A universal rule would ignore heterogeneity that is directly visible in the paired records.

After matrix-family adjustment, the directional-tail contrast has coefficient 0.1758, with standard error 0.0201 and interval 0.1240 to 0.2275. The adjusted model R-squared is 0.8823. The primary mean-only model has zero R-squared by construction. The second fit suggests that directional extremes accompany grouping sensitivity within this small collection, but the features and distortion share the same weights, so this is a diagnostic association rather than an independently validated predictor.

The whole-tensor reference has much larger mean normalized error, approximately 0.1763. Its lower scale budget is part of that comparison. We do not convert the difference into an accuracy gain, a memory-bandwidth saving, or a recommendation for a production quantizer. Serving-system technical debt discussed by Sculley et al., 2015 provides a useful reminder that a local numerical improvement can leave substantial implementation and deployment work unresolved.

7. What a matrix-error audit can and cannot support

Robust-estimation ideas in Huber, 1964, Hampel, 1974, and Rousseeuw, 1984 concern limiting observation influence in different statistical settings. Our maximum-based scale is not a robust scale estimator in that sense. Exceptional weights can expand the code range for an entire group. That mechanism makes directional tail summaries relevant, but replacing maxima with another scale rule would change the quantization policy and would require a separate declared experiment.

Influence diagnostics from Cook, 1977 could identify matrices that strongly affect the adjusted slope. Measurement-error considerations in Carroll et al., 2006 also warn that unweighted reconstruction error is a proxy for functional preservation. Missing-data reasoning in Little and Rubin, 2002 concerns the absent activation and behavior measurements, not missing numeric values in the selected matrices. We do not fill

those evidential gaps by assuming that a favorable reconstruction ratio implies preserved next-token probabilities.

Model-selection guidance in Varma and Simon, 2006 is relevant if the orientation were chosen separately after seeing each test outcome. Here both policies are evaluated on the same fixed matrices and no behavioral test set is used for selection. A future adaptive grouping policy should learn its rule on independent calibration information and evaluate the entire selection procedure on untouched behavior data. The present census of projection matrices cannot provide that external validation by itself.

The guidance in Wasserstein and Lazar, 2016 and concerns in Ioannidis, 2005 support a restrained interpretation of the small-cluster intervals. Multiple matrix-family summaries are exploratory. Benjamini and Hochberg, 1995 and Benjamini and Yekutieli, 2001 describe multiplicity control for broader testing families, but we do not present a confirmatory discovery list of compression-sensitive layers. The full matrix table is retained, including the six cases where the aggregate direction reverses.

8. A concrete compression-design boundary

The equal-matrix weighting also deserves attention. An attention output matrix and a feed-forward expansion matrix receive one contribution each to the mean log ratio, despite containing different numbers of weights. That choice answers an architectural consistency question rather than a total parameter-error question. A parameter-weighted summary could favor families with larger arrays and would be a different estimand. The processed records retain matrix dimensions and weight counts so that this distinction is visible rather than hidden behind a single model-wide average.

The error calculation is deterministic for the supplied arrays. Re-running it with the same checkpoint and rounding rule reproduces the same values. Statistical uncertainty here therefore does not concern computational randomness or a hidden set of repeated quantization trials. It concerns how one chooses to interpret variation among related projection matrices. Six source layers provide a modest basis for such an interpretation, and no confidence interval can replace the missing evidence from separately trained checkpoints or larger architectures.

A practical extension would evaluate orientation choices against input-dependent projection errors using a calibration corpus. That would require recording the activations, separating calibration from evaluation passages, and preserving a common memory budget across candidate rules. It could then examine whether the orientation preferred by weight error is also preferred by next-token loss. Until those measurements are made, choosing a quantizer solely from the present matrix criterion risks optimizing a proxy whose relationship to the intended language task remains untested.

Another implementation question is whether a favorable orientation is compatible with an efficient execution kernel. Column-oriented scale groups may require a different memory layout or dequantization path from row-oriented groups. Our bit accounting does not include a cost model for that path. The experiment establishes a numerical difference at equal conceptual storage, while a complete compression evaluation would need to report actual packed sizes, conversion overhead, runtime behavior, and functional outputs under a specified implementation.

Documentation follows Gebru et al., 2021 by identifying the checkpoint revision context, selected arrays, rounding convention, group size, and exclusions. Reuse principles in Wilkinson et al., 2016 motivate preserved source hashes and matrix-level results. The evidence is a direct computation on a trained small model, with no claim of new training or a functional large-model benchmark. Its useful conclusion is that equal nominal precision and scale count can still produce materially different reconstruction errors, with orientation sensitivity varying by matrix family.

References

1. Ioannidis, John P. A.. (2005). Why Most Published Research Findings Are False. PLoS Medicine. https://doi.org/10.1371/journal.pmed.0020124.
2. 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
3. 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.
4. 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.
5. 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.
6. Lin, Lawrence I-Kuei. (1989). A Concordance Correlation Coefficient to Evaluate Reproducibility. Biometrics. https://doi.org/10.2307/2532051.
7. Little, Roderick J. A., and Donald B. Rubin. *Statistical Analysis with Missing Data*. Wiley, 2002. https://doi.org/10.1002/9781119013563.
8. 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.
9. Rousseeuw, Peter J.. (1984). Least Median of Squares Regression. Journal of the American Statistical Association. https://doi.org/10.1080/01621459.1984.10477105.
10. 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.
11. Hu, Edward J., et al. “LoRA: Low-Rank Adaptation of Large Language Models.” *International Conference on Learning Representations*, 2022. https://openreview.net/forum?id=nZeVKeeFYf9.
12. 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.
13. 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.
14. Hampel, Frank R.. (1974). The Influence Curve and its Role in Robust Estimation. Journal of the American Statistical Association. https://doi.org/10.1080/01621459.1974.10482962.
15. Koo, Terry K., and Mae Y. Li. "A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research." *Journal of Chiropractic Medicine*, vol. 15, no. 2, 2016, pp. 155-163. https://doi.org/10.1016/j.jcm.2016.02.012.
16. 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.
17. Carroll, Raymond J., David Ruppert, Leonard A. Stefanski, and Ciprian M. Crainiceanu. “Measurement Error in Nonlinear Models.” 2006. https://doi.org/10.1201/9781420010138.
18. Cook, R. Dennis. (1977). Detection of Influential Observation in Linear Regression. Technometrics. https://doi.org/10.1080/00401706.1977.10489493.
19. Ma, Xinyin; Fang, Gongfan; Wang, Xinchao. (2023). LLM-Pruner: On the Structural Pruning of Large Language Models. https://arxiv.org/abs/2305.11627.
20. 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.
21. 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.
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. 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.

24. 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.

25. Rafailov, Rafael; Sharma, Archit; Mitchell, Eric; Ermon, Stefano; Manning, Christopher D.; Finn, Chelsea. (2023). Direct Preference Optimization: Your Language Model is Secretly a Reward Model. arXiv:2305.18290. https://arxiv.org/abs/2305.18290.

26. 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.

27. Huber, Peter J.. (1964). Robust Estimation of a Location Parameter. The Annals of Mathematical Statistics. https://doi.org/10.1214/aoms/1177703732.

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

29. 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.

30. Benjamini, Yoav and Daniel Yekutieli. "The control of the false discovery rate in multiple testing under dependency." *The Annals of Statistics*, vol. 29, no. 4, 2001, pp. 1165-1188. https://doi.org/10.1214/aos/1013699998.