Population Variation and Equity
Volume 1 · Number 1 · 2026

Vocabulary Averages and Actual Lookups: Reweighting an Embedding Quantization Audit

Luke Sullivan[1] · Gabriel Brooks[2] · Matthew Fisher[3]

[1] Ireland · [2] Denmark · [3] Norway

Original publication: November 3, 2026 | Article 638

Abstract

The same quantized embedding matrix can receive different audit scores depending on whether token types or token occurrences receive equal weight. We reconstructed 50,275 ordinary Pythia-14M input embedding rows using symmetric four-bit rounding. Mean row-normalized error was 0.013830 across vocabulary entries but 0.022794 when weighted by 340,240 token occurrences in Tiny Shakespeare. Five frequently encountered entries accounted for 24.4 percent of lookups and 40.9 percent of the weighted error. A simple seen-versus-unseen indicator showed little adjusted separation, while frequency had a small association within encountered types. The audit measures stored-vector distortion; it does not measure generation quality or establish which tokens deserve additional precision.

Direct numerical reconstruction of a public Pythia-14M input embedding matrix, with occurrence counts from Tiny Shakespeare.

Vocabulary Averages and Actual Lookups: Reweighting an Embedding Quantization Audit

Direct numerical reconstruction of a public Pythia-14M input embedding matrix, with occurrence counts from Tiny Shakespeare.

Abstract

The same quantized embedding matrix can receive different audit scores depending on whether token types or token occurrences receive equal weight. We reconstructed 50,275 ordinary Pythia-14M input embedding rows using symmetric four-bit rounding. Mean row-normalized error was 0.013830 across vocabulary entries but 0.022794 when weighted by 340,240 token occurrences in Tiny Shakespeare. Five frequently encountered entries accounted for 24.4 percent of lookups and 40.9 percent of the weighted error. A simple seen-versus-unseen indicator showed little adjusted separation, while frequency had a small association within encountered types. The audit measures stored-vector distortion; it does not measure generation quality or establish which tokens deserve additional precision.

1. A vocabulary table and a stream of lookups weight different objects

An input embedding matrix contains one learned vector for each available token identifier. A matrix audit can give every row equal importance, irrespective of how often that row is selected in a particular document. An application instead performs lookups according to its input stream. Those two weighting schemes can yield different summaries without any change to the compressed vectors. The distinction is particularly relevant when a few formatting or punctuation entries appear much more often than most lexical pieces.

This article asks whether a simple weight-reconstruction score changes materially when vocabulary rows are weighted by occurrences in a specified public text. A related subgroup question asks whether entries encountered in that text have systematically different distortion from entries not encountered, after accounting for encoded piece length and a leading-space indicator. These are descriptive questions about one checkpoint and one text collection. They do not identify pretraining token frequencies or infer the performance of an interactive language-model application.

The compression setting of Wang et al., 2026 motivates examining the object over which a distortion summary is averaged. We neither implement that compression procedure nor evaluate its accuracy, throughput, or architecture. Our executed operation is symmetric rounding of a single input embedding matrix. Its value is an inspectable example of how the evaluation population changes a summary. A claim about the target system would require its own compressed model and task outputs, which are absent from this study.

2. Embedding and tokenization choices define the audit population

The learned-vector perspectives represented by Mikolov et al., 2013 and Pennington et al., 2014 give background for treating embedding rows as meaningful model parameters rather than arbitrary storage cells. Their training objectives and released vectors are not used here. Our weights come from the final public Pythia-14M checkpoint associated with Biderman et al., 2023. We read its numeric safetensors file directly and select only the input embedding tensor. Output embeddings, attention projections, normalization parameters, and subsequent activations do not enter the calculation.

The selected tensor has 50,304 rows and 128 columns. The official tokenizer maps 50,277 identifiers, including its registered added tokens. Twenty-seven matrix rows therefore lack a corresponding vocabulary identifier. We exclude those padded rows and the two explicitly special identifiers, leaving 50,275 ordinary rows. This boundary prevents unused storage padding from silently changing an equal-vocabulary average. Ordinary added space tokens remain eligible because the tokenizer treats them as nonspecial pieces that can actually occur in text.

Subword segmentation, as motivated in Sennrich et al., 2015, makes frequency a property of token pieces rather than necessarily whole words. The captured tokenizer specifies byte-level preprocessing and a byte-pair model. A vocabulary piece can encode a space prefix, punctuation, a line break, part of a word, or a complete word. We therefore avoid describing unseen pieces as rare words or encountered pieces as common words. Their status is defined only by the actual encoding of the chosen text.

3. A reproducible text stream, with no generated tokens

The corpus is the public Tiny Shakespeare input file preserved in Karpathy, 2026. Its UTF-8 representation contains 1,115,394 bytes. We process the complete file as one sequence, using the captured official tokenizer JSON through Hugging Face contributors, 2025, version 0.22.1. No beginning marker, end marker, padding, or truncation is inserted. This operation yields 340,240 token occurrences distributed across 11,225 ordinary token types. The remaining 39,050 analyzed types have zero occurrences in this specific file.

As a compatibility check, decoding the complete encoded identifier sequence with special-token removal disabled reproduces the original text exactly. Neither excluded special identifier occurs in the encoded stream. These checks support the relationship between occurrence counts and the preserved input. They do not establish that Tiny Shakespeare is a representative deployment workload, a held-out sample from pretraining, or an independent evaluation corpus for language-model accuracy. No continuation is generated and no probability of the observed text is calculated.

For the regression controls we measure the length of the stored vocabulary-piece string and indicate whether it begins with the encoded space marker or a literal space. The first measure is called encoded piece length, not word length or a universal count of human-readable characters. Byte-level alphabets and added tokens require that distinction. Counts refer to the full-file tokenization, so they also depend on the file's existing line breaks and punctuation rather than a normalized prose-only version.

4. Equal row policies with explicitly rounded scales

For each ordinary embedding row, the four-bit policy divides its largest absolute weight by seven and rounds that scale to float32 before reconstruction. Every coordinate is divided by the stored scale, rounded to the nearest integer, restricted to the range minus seven through seven, and multiplied back by the scale. The eight-bit reference uses minus 127 through 127 in the same way. The stored source weights are float16; converting them to higher precision for arithmetic does not create a more accurate underlying reference model.

Each row's distortion is its mean squared reconstruction error divided by its mean squared source weight. All selected rows have positive source energy and positive measured distortion under both policies. This normalization prevents a large-magnitude row from dominating solely because its coordinates have a different scale. It also defines a particular endpoint: averaging relative row distortions is different from dividing a pooled occurrence-weighted error by pooled embedding energy. We report the former and do not rename it a global activation error.

The quantization context in Frantar et al., 2022, Lin et al., 2023, Xiao et al., 2022, Dettmers et al., 2022, and Dettmers et al., 2023 underscores that sophisticated compression can use information beyond an isolated row's maximum. None of those algorithms supplies a fitted compressed model to this audit. Pruning approaches in Sun et al., 2023 and Ma et al., 2023 address another form of parameter reduction. Their inclusion situates the resource problem without implying that our simple scalar policy approximates their task performance or reproduces their calibration procedures.

5. The diagram separates weights, counts, and task outcomes

The nominal four-bit convention uses fifteen signed integer levels and leaves one code unused; the eight-bit convention uses 255 levels. With one float32 scale per 128 coordinates, the arithmetic accounting is 4.25 and 8.25 bits per weight respectively. Those figures exclude a packed file format's other overhead. We construct reconstructed arrays for error measurement and do not claim to have produced or benchmarked an optimized compressed inference artifact.

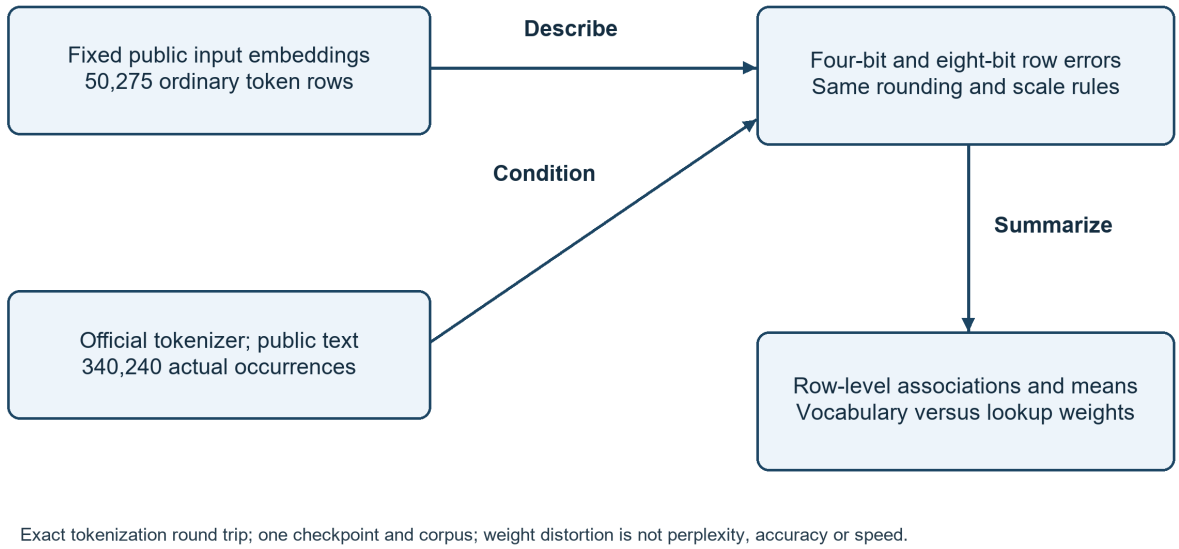


Figure 1. Fixed source embedding rows are reconstructed independently of text; actual token counts then define occurrence weighting and descriptive subgroup regressions.

Numerical considerations in Higham, 2002 and Goldberg, 1991 motivate preserving the scale precision and rounding order. Hardware reasoning in Williams et al., 2009 also explains why a nominal bit reduction cannot by itself supply a speedup estimate. We record no memory-bandwidth measurement, kernel timing, or device configuration. The figure ends at vector-error summaries because adding an accuracy or latency endpoint would introduce an unperformed experiment.

6. Two regression populations answer distinct questions

The first ordinary least-squares model uses all 50,275 analyzed vocabulary rows. Its outcome is the natural logarithm of four-bit normalized distortion. Predictors are encountered status, log encoded piece length, and the leading-space indicator. The second model is restricted to the 11,225 encountered types and replaces status with log of one plus occurrence count, while also controlling for log row root-mean-square magnitude and the two piece features. Both design matrices are verified to have full column rank.

We report HC3 covariance estimates following the heteroskedasticity-consistent approach of White, 1980, and normal-reference intervals. Each token type contributes one regression row, regardless of how often it occurs. This is essential: duplicating a row once per occurrence would create an illusion of hundreds of thousands of independent weight observations. Corpus weighting is used for an aggregate lookup-oriented summary, while the regressions remain explicit row-level descriptive associations.

Table 1. Associations with log four-bit row-normalized distortion in the full ordinary vocabulary and the encountered subset.

Model	Term	Estimate (SE)	95% CI	N
M1	Encountered in corpus	0.0034 (0.0028)	[-0.0020, 0.0088]	50275
M1	Log encoded piece length	-0.0076 (0.0026)	[-0.0127, -0.0026]	50275
M1	Leading-space indicator	0.0013 (0.0026)	[-0.0039, 0.0064]	50275
M1: fit	R-squared = 0.0002	-	-	50275
M2	Log(1 + occurrence count)	0.0078 (0.0024)	[0.0030, 0.0126]	11225
M2	Log row RMS magnitude	-0.1010 (0.0400)	[-0.1794, -0.0226]	11225
M2	Log encoded piece length	-0.0071 (0.0073)	[-0.0213, 0.0071]	11225
M2	Leading-space indicator	0.0024 (0.0063)	[-0.0099, 0.0147]	11225
M2: fit	R-squared = 0.0035	-	-	11225

Model labels: M1 = Seen versus unseen token rows with piece controls; M2 = Frequency association within encountered token rows.

The dependence principles in Cameron and Miller, 2015 limit inferential interpretation. Vocabulary vectors were learned together in a single model, and the available rows do not create independent model or training-run replications. We do not invent a clustering scheme from arbitrary identifier ranges. The reported HC3 intervals provide a conditional descriptive approximation across rows. They cannot certify generalization to other checkpoints or corpora. Cross-domain validation concerns in Roberts et al., 2017 remain relevant even with a large number of available vector rows.

7. A weak binary distinction coexists with a large weighting change

Equal weighting across the ordinary vocabulary gives mean four-bit normalized distortion 0.013829944. Within the encountered subset the corresponding unweighted average is 0.013891868, compared with 0.013812144 among unencountered types. In the first regression, encountered status has coefficient 0.003427 with standard error 0.002753 and interval -0.001970 to 0.008823. The model R-squared is only 0.000243. Being encountered at least once therefore offers little separation in average log distortion under these controls.

Occurrence weighting produces a different result: mean four-bit normalized distortion becomes 0.022793555, about 64.8 percent above the equal-vocabulary mean. The eight-bit summaries are 0.000042018 and 0.000070197 under vocabulary and occurrence weighting respectively. These are different averages of the same reconstructed rows. Neither contrast is a before-and-after change in a

generated answer. The arithmetic shows that the composition of lookups matters to this audit even though a simple binary subgroup label is weakly informative.

Within encountered types, the log-count coefficient is 0.007775, with standard error 0.002444 and interval 0.002985 to 0.012565. Log row magnitude has coefficient -0.100978 and interval -0.179351 to -0.022605. The second R-squared is 0.003463. These small explanatory fractions matter: an association distinguishable from zero is not a useful predictive model for most row-level variation. Frequency should not be presented as a near-complete explanation of quantization difficulty or as a causal property of the learned vector.

8. Formatting entries contribute disproportionately in this corpus

The five most frequent ordinary identifiers account for 24.4 percent of token occurrences and 40.9 percent of the occurrence-weighted four-bit row loss. They represent the encoded line break, comma, colon, period, and a space-prefixed article. Selection of these five uses count and identifier ordering rather than reconstruction error. Their loss shares are then measured on the reconstructed rows. This accounting helps explain why treating the corpus as unformatted lexical content would misdescribe the actual lookup distribution.

The line-break entry alone occurs 39,996 times, or about 11.8 percent of the stream, and supplies approximately 22.1 percent of the weighted loss. Its row-normalized four-bit error is 0.042768. The result is not evidence that line breaks deserve high precision in every application. It shows that document format and token-vector geometry interact in this one workload. Removing line breaks would change both the source text and its segmentation, requiring a newly specified audit rather than a cosmetic reinterpretation of this calculation.

Influence concerns associated with Cook, 1977 motivate making this concentration visible. We retain these frequent entries rather than excluding them as inconvenient outliers. Measurement reasoning in Carroll et al., 2006 also cautions that the chosen normalized error is only one observable summary. Its prominence cannot establish that the same entries dominate downstream loss, where positional context, later layers, and output sensitivity could change the relationship completely.

9. This audit does not select an accuracy-preserving compression policy

Calibration work represented by Guo et al., 2017 and scoring-rule principles in Gneiting and Raftery, 2007 concern model predictions and their evaluation. We calculate neither a predicted probability nor a proper predictive score. Even a sizeable change in lookup-weighted vector distortion need not produce a proportionate change in perplexity or task accuracy. Activation-aware evaluation would require forwarding the actual text through the original and reconstructed models. We have not performed that computation and do not substitute the embedding audit for it.

Model-selection concerns in Varma and Simon, 2006 would arise if these corpus-specific results were used to choose precision assignments and the same text then served as final evaluation evidence. Multiplicity guidance associated with Benjamini and Hochberg, 1995 is also relevant to a future search over many row features or subgroups. The present report does not optimize a mixed-precision policy or announce a preferred allocation. Its two specifications and descriptive concentration account are retained as an exploratory analysis, with the limits emphasized under Wasserstein and Lazar, 2016 rather than hidden behind small standard errors.

The reproducibility questions raised in Ioannidis, 2005 are answered only partially by exact source preservation. The weights and counts can be recomputed, but one small checkpoint cannot establish

behavior across scales, languages, or application domains. A useful next experiment would prespecify additional corpora and measure actual model loss before using occurrence weights for policy selection. Such work could validate or overturn the relevance of this proxy; neither outcome is already implied by the current numbers.

10. Specify the weighting population before comparing audit scores

The main lesson is operationally modest. A researcher reporting an embedding distortion average should say whether it covers storage rows, mapped ordinary vocabulary entries, encountered types, or actual occurrences. Those choices produced materially different scores in this public example. The fact that encountered status alone was weakly associated with distortion makes the lesson more specific: counting a token once and accounting for how often it is used are not interchangeable summaries.

Records motivated by Gebru et al., 2021 include the special identifiers, unmapped padding rows, exact corpus, tokenizer version, decoding round trip, and normalized-error definition. Preservation following Wilkinson et al., 2016 keeps the numeric checkpoint source, tokenizer, text snapshot, per-row errors, and concentration table available together. These materials support reproduction of the vector audit. Their scientific interpretation remains restricted to that audit until downstream task evidence establishes whether the observed weighting difference has practical consequences.

References

1. 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.
2. 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.
3. 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.
4. 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.
5. 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.
6. Hugging Face contributors, . (2025). Tokenizers v0.22.1 [Software]. https://github.com/huggingface/tokenizers/releases/tag/v0.22.1.
7. 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.
8. Ma, Xinyin; Fang, Gongfan; Wang, Xinchao. (2023). LLM-Pruner: On the Structural Pruning of Large Language Models. https://arxiv.org/abs/2305.11627.
9. 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.
10. 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.
11. 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.
12. 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.
13. 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.

14. 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.

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

16. Ioannidis, John P. A.. (2005). Why Most Published Research Findings Are False. PLoS Medicine. https://doi.org/10.1371/journal.pmed.0020124.

17. Mikolov, Tomas; Chen, Kai; Corrado, Greg; Dean, Jeffrey. (2013). Efficient Estimation of Word Representations in Vector Space. https://arxiv.org/abs/1301.3781.

18. 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.

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

20. 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.

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. 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.

23. Sennrich, Rico; Haddow, Barry; Birch, Alexandra. (2015). Neural Machine Translation of Rare Words with Subword Units. https://arxiv.org/abs/1508.07909.

24. 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.

25. 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.

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

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

28. 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.

29. 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.

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