Journal of Data Efficiency
Volume 1 · Number 1 · 2026

More Calibration Tokens Change Group Choices, with an Uncertain Loss Benefit

Julian Griffin[1] · Jacob Murphy[2] · Benjamin Lewis[3]

[1] Australia · [2] Netherlands · [3] New Zealand

Original publication: October 24, 2026 | Article 686

Abstract

Increasing the calibration budget sixteenfold changed 1,215 of 9,216 group-level clipping choices, but yielded an uncertain improvement in subsequent text loss. Using Pythia-14M, we estimated input second moments from nested prefixes of 512 and 8,192 tokens and fitted the same diagonal-weighted four-bit rounding rule. On 16,384 evaluation targets elsewhere in the file, the larger budget reduced mean negative log likelihood by 0.003545 nats; the conditional 95% interval ranged from a 0.011496 reduction to a 0.004407 increase. Both rounded models retained substantial loss relative to source weights. The experiment separates visible calibration changes from demonstrated predictive benefit, without establishing that the smaller budget is generally sufficient.

Nested calibration inputs for a fixed four-bit clipping heuristic, followed by actual CPU next-token scoring on a separate local text segment.

More Calibration Tokens Change Group Choices, with an Uncertain Loss Benefit

Nested calibration inputs for a fixed four-bit clipping heuristic, followed by actual CPU next-token scoring on a separate local text segment.

Abstract

Increasing the calibration budget sixteenfold changed 1,215 of 9,216 group-level clipping choices, but yielded an uncertain improvement in subsequent text loss. Using Pythia-14M, we estimated input second moments from nested prefixes of 512 and 8,192 tokens and fitted the same diagonal-weighted four-bit rounding rule. On 16,384 evaluation targets elsewhere in the file, the larger budget reduced mean negative log likelihood by 0.003545 nats; the conditional 95% interval ranged from a 0.011496 reduction to a 0.004407 increase. Both rounded models retained substantial loss relative to source weights. The experiment separates visible calibration changes from demonstrated predictive benefit, without establishing that the smaller budget is generally sufficient.

1. A larger calibration sample must be judged through its fitted policy

Post-training numerical compression often needs representative inputs to estimate which parameters or channels matter. More inputs can alter an estimated importance profile and the resulting precision choices. Those changes alone do not establish a useful improvement in prediction. A calibration procedure can become more stable in one intermediate quantity while an approximation in its objective limits the value of additional observations. The evaluation must therefore follow the fitted policy through to the outcome it is meant to preserve.

Our research question is whether increasing a nested calibration-input budget from 512 to 8,192 tokens improves next-token loss under one fixed four-bit clipping heuristic. The directional proposition favors a negative larger-minus-smaller loss difference on a separate local segment. We also record how often the selected group clipping fractions change and how both policies compare with source weights. The design does not search for an optimal calibration size or claim that a null difference proves the smaller budget sufficient.

Resource requirements in the compression setting of Wang et al., 2026 motivate the question. The executed procedure is a declared diagonal-weighted reconstruction heuristic, rather than an implementation of that framework. Its coefficients are fitted from unlabeled original-model inputs, with no fine-tuning or coding-preference labels. We keep that boundary because conclusions about the number of calibration observations apply to a specific fitting objective and representation. They cannot be transferred automatically to a different compression architecture or learning procedure.

2. The public model, source values, and local runtime

The underlying checkpoint is the final standard-Pile Pythia-14M model documented by Biderman et al., 2023. We load its numeric safetensors state strictly into the maintained GPT-NeoX implementation. Source values stored as float16 are promoted to float32 for both original and reconstructed forward passes. Evaluation mode is retained throughout. The code uses the CPU implementation in Paszke et al., 2019, version 2.6.0+cpu, with Wolf et al., 2019 version 4.57.1, eager attention, disabled caching, and four execution threads.

The model has six transformer layers and hidden width 128. We modify the same four internal projection matrices in every layer, totaling twenty-four matrices and 1,179,648 weights. Input embeddings, output weights, biases, and normalization parameters remain at their source values. Thus the experimental resource is a fixed internal-weight reconstruction problem. We do not report a whole-model low-bit storage size or assume that a small nominal precision for selected matrices describes the much larger unchanged embedding tables.

Numerical guidance in Higham, 2002 and Goldberg, 1991 motivates preserving arithmetic order and source precision. All selected matrix values are finite, group scales are positive, and the clipping search uses the same candidate grid at both budgets. A strict source-state reload precedes each reconstructed policy, preventing cumulative rounding across evaluations. These controls make the numerical intervention identifiable, while leaving cross-machine floating-point differences subject to a numerical reproduction tolerance rather than a universal promise of identical output bytes.

3. Nested text inputs and the separation of local roles

We use the Tiny Shakespeare file identified by Karpathy, 2026, encoded with the captured official tokenizer through Hugging Face contributors, 2025 version 0.22.1. The complete file produces 340,240 token identifiers, and a full decoding round trip reproduces the input exactly. The small calibration budget takes its first 512 input tokens. The larger budget takes the first 8,192, including every input in the smaller set. No special tokens are inserted and no labels are generated for calibration.

Original-model inputs are processed in blocks of 256 tokens, with context reset at each block boundary. The first budget covers two blocks and the second thirty-two. These are contiguous nested prefixes, not independently randomized samples from a deployment population. Increasing the budget also changes the particular text and contexts represented in the estimated moments. The comparison therefore conditions on one realized prefix ordering. It cannot isolate a universal sample-size effect independently of calibration composition or source-domain characteristics.

Evaluation begins much later in the file, at zero-based input offset 170,120. It scores sixty-four blocks with 256 next-token targets each, totaling 16,384 targets. Calibration input positions and evaluation positions do not overlap locally. Both portions may nevertheless have appeared in the checkpoint's original pretraining. The validation issues discussed in Roberts et al., 2017 make it important to call this local role separation rather than certified unseen-language evaluation or independent-domain generalization.

4. Estimating the original model's input second moments

Before modifying any weights, we attach measurement hooks to the twenty-four selected linear modules. Each hook accumulates the squared input value for every feature dimension in float64. Dividing by the number of observed positions yields a diagonal vector of input second moments. Every selected module is verified to have exactly 512 or 8,192 observed positions at the corresponding snapshot. The smaller snapshot is copied before further accumulation, so it is not silently overwritten by the larger budget.

These are uncentered second moments, not a full covariance matrix. They retain featurewise scale information while discarding cross-feature products. Both policies use original-model activations, rather than recalculating downstream inputs after each preceding layer has been quantized. That restriction simplifies the experiment but leaves an approximation between the calibration objective and actual rounded-network behavior. Larger samples can estimate this proxy more faithfully without eliminating the approximation itself.

Calibration-aware methods in Lin et al., 2023, Xiao et al., 2022, Yao et al., 2022, and Shao et al., 2023 give context for using representative inputs in compression. Our method does not reproduce their optimization, transformations, or reported results. The post-training setting in Frantar et al., 2022 likewise supplies a broader motivation, not an implemented benchmark. We state the measured moments and fitting objective directly so that a useful-sounding label cannot imply a stronger algorithm than the calculations actually contain.

5. One fixed group search at both budgets

Each selected matrix is partitioned in row-major order into groups of 128 weights, yielding 9,216 groups. For every group we consider eleven clipping fractions from 0.50 to 1.00 in increments of 0.05. A fraction multiplies the group's maximum absolute weight before division by seven defines the float32 scale. Divided weights are rounded and restricted to signed codes from minus seven through seven. Fractions below one can saturate larger source weights while using a finer scale for the remaining values.

For a candidate fraction, the objective sums squared weight-reconstruction errors weighted by the corresponding input second moments. Each output row uses the same estimated featurewise importance vector for its input dimensions. We choose the candidate with the smallest diagonal-weighted objective for that group; increasing grid order supplies deterministic tie resolution. The objective uses calibration moments only. Evaluation target losses do not choose fractions, revise the grid, or decide which budget enters the report.

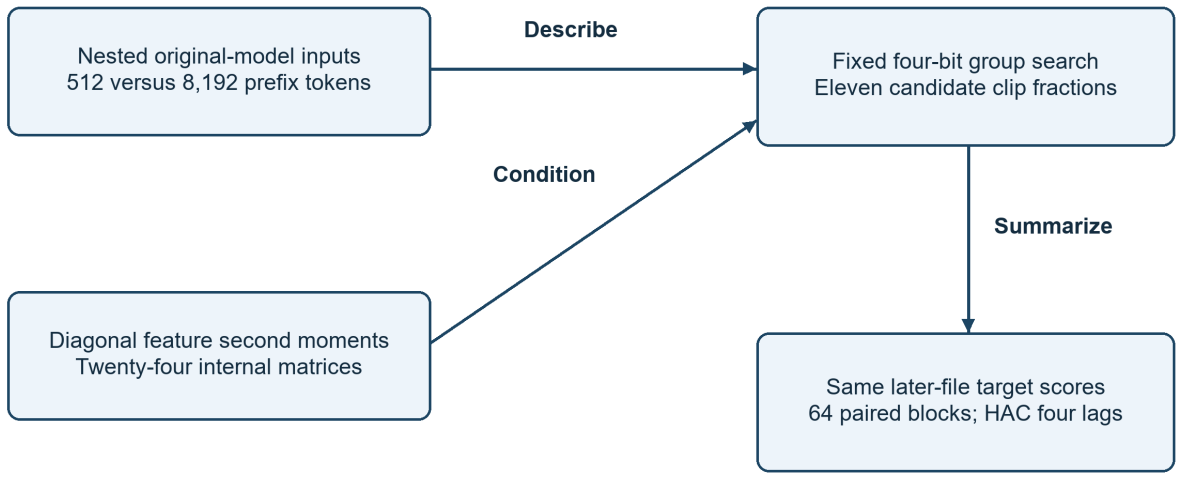


Figure 1. Two nested original-model moment estimates fit the same groupwise clipping rule; reconstructed models are then scored on identical later-file targets.

The eight-bit treatment discussed in Dettmers et al., 2022 and low-rank adaptation in Hu et al., 2022 concern different resource choices. We fit neither an eight-bit comparison nor a trainable low-rank update in this experiment. The source-valued model remains the numerical reference. The diagram contains actual input measurement, discrete group fitting, and forward scoring, with no claim about supervised adaptation, model merging, or the behavior of a target compression method that has not been run.

6. Group changes are observable intermediate outcomes

Moving from the small to the large calibration budget changes 1,215 selected fractions, approximately 13.2 percent of the 9,216 groups. The mean selected fraction changes only from 0.840061 to 0.840473. These summaries answer different questions: a nearly unchanged average can coexist with many different group assignments. The number of changed groups is measured directly from the two fitted policies and does not require defining a successful downstream prediction effect.

The unweighted parameter-space normalized squared errors are 0.008107269 for the smaller-budget policy and 0.008100711 for the larger-budget policy. They pool squared reconstruction error over the affected matrices and divide by their pooled source energy. These values are not the diagonal-weighted calibration objective and are not next-token losses. We do not conclude that one policy must predict better because this parameter-space summary is slightly lower. Actual forward scoring is needed to examine that claim.

Agreement perspectives in Bland and Altman, 1986, Lin, 1989, and Koo and Li, 2016 motivate caution about calling two policies interchangeable merely because their average fractions are close. We do not calculate an agreement coefficient or reliability statistic for clipping assignments. Measurement reasoning in Carroll et al., 2006 also applies to the estimated importance vectors: they summarize observed original-model inputs under a chosen text and context rule. They do not reveal an error-free, domain-invariant measure of parameter importance.

7. Mean loss contrasts on the same evaluation blocks

Every source and reconstructed model scores the same observed next-token identifiers, retaining all 50,304 logits in the probability normalization. There are no generated continuations or task-success labels. We average paired loss differences within the sixty-four blocks, then estimate their means using intercept-only ordinary least squares. The primary contrast is larger-budget minus smaller-budget negative log likelihood. Two accompanying contrasts compare each rounded policy with the source-valued model.

Covariance is HAC with four available-block lags, a finite-sample correction, and a t reference using sixty-three degrees of freedom, following the framework in Newey and West, 1987. The predictive-comparison perspectives in Diebold and Mariano, 1995 and West, 1996 motivate losses paired on identical target contexts. Heteroskedasticity adjustment in White, 1980 alone would not describe serial dependence across neighboring text blocks. All three fits contain only a mean, so their R-squared values are effectively zero and provide no reconstruction-accuracy percentage.

Table 1. Conditional mean next-token loss contrasts in nats per target, with sixty-four paired evaluation blocks.

Model	Term	Estimate (SE)	95% CI	N
M1	Mean outcome/contrast	-0.003545 (0.003979)	[-0.011496, 0.004407]	64
M1: fit	R-squared = 0.0000	-	-	64
M2	Mean outcome/contrast	0.503067 (0.013822)	[0.475447, 0.530687]	64
M2: fit	R-squared = 0.0000	-	-	64
M3	Mean outcome/contrast	0.499522 (0.014916)	[0.469715, 0.529330]	64
M3: fit	R-squared = 0.0000	-	-	64

Model labels: M1 = Larger minus smaller calibration budget; M2 = Smaller calibration versus source; M3 = Larger calibration versus source.

These intervals condition on the fitted calibration policies. They do not include the uncertainty of repeatedly sampling another 512-token prefix, another larger calibration collection, or another pretrained checkpoint. Language within one text can remain dependent beyond four blocks. We retain the declared lag rather than implying that a covariance correction establishes independent contexts. The distinction between evaluation-row variation and calibration-sample variation is central to interpreting a data-requirement statement.

8. More inputs yield a small uncertain predictive change

The source-valued model has mean evaluation loss 4.717160753 nats. The small-budget rounded model has mean loss 5.220227938, while the large-budget model has mean loss 5.216683193. The primary larger-minus-smaller difference is -0.003544745, with standard error 0.003978972 and interval -0.011496085 to 0.004406596. The point estimate favors the larger budget, but the interval includes both an improvement and a modest deterioration under the conditional approximation.

The smaller-budget policy increases loss relative to source by 0.503067184 nats, with interval 0.475447056 to 0.530687313. The larger-budget increase is 0.499522440, with interval 0.469714998 to 0.529329881. Both policies therefore retain substantial numerical loss compared with source values. Sixteen times as many calibration inputs do not come close to eliminating that degradation in this realized comparison. This statement concerns the observed fitted procedure, not the best achievable four-bit representation of the checkpoint.

The directional proposition favoring more calibration receives an uncertain rather than decisive result. We do not translate the interval's inclusion of zero into proof that 512 tokens suffice. An equivalence conclusion would require a justified practical margin and evidence designed to assess it. Nor do 1,215 changed fractions independently confirm better calibration quality. They establish that the fitting result responds to the additional inputs, leaving its predictive value to the separately measured loss contrast.

9. A more precise proxy can still leave the main limitation untouched

The diagonal objective ignores correlations among input dimensions and interactions created when several matrices are reconstructed together. It also relies on source-model inputs rather than the perturbed network's later activations. These approximations provide possible reasons why a much larger input count produces little additional benefit. We have not varied them experimentally, so they remain design limitations and hypotheses for later work, rather than measured explanations of the primary coefficient.

Influence analysis in Cook, 1977 could identify evaluation blocks that contribute most to the budget contrast, but no difficult block is removed here. Model-selection bias described in Varma and Simon, 2006 would arise if the evaluation losses were used to redesign the clipping grid or choose the reported budget. We retain both declared budgets and the original grid. The interpretation supported by Wasserstein and Lazar, 2016 emphasizes the uncertain contrast and persistent source-relative loss rather than assigning a universal sufficient calibration size from one interval.

Reproducibility concerns in Ioannidis, 2005 extend beyond repeating the same prefix. Additional calibration draws, text domains, and model sizes would contribute genuinely new evidence. A different importance approximation or sequential recalibration procedure would also define a new method requiring its own test. None of those operations has been completed by measuring more positions from

this single file. The result remains a focused experiment about how one fitted clipping heuristic responds to a particular nested increase in available inputs.

10. Count inputs, changed decisions, and predictive evidence separately

The completed analysis records three distinct facts: the calibration input count rises sixteenfold, roughly one eighth of group choices change, and the measured predictive improvement remains small and uncertain. Keeping those facts separate prevents an attractive intermediate result from becoming an unsupported data-sufficiency claim. It also shows why a calibration study should retain the source-model comparison: both fitted policies can remain materially degraded even when one is slightly better than the other.

Documentation under Gebru et al., 2021 includes the prefix boundaries, context resets, module hooks, copied moment snapshots, clipping grid, deterministic ties, and absence of certified pretraining separation. Preservation following Wilkinson et al., 2016 links raw model and tokenizer sources with group choices, actual per-target losses, and paired block summaries. These materials support reproduction of the reported budget comparison. Broader advice about calibration size requires further independent evidence and application-specific loss criteria beyond this local small-model experiment.

References

1. 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.
2. Gebru, Timnit, et al. "Datasheets for datasets" *Communications of the ACM*, 2021. https://doi.org/10.1145/3458723.
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. 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.
5. Hugging Face contributors, . (2025). Tokenizers v0.22.1 [Software]. https://github.com/huggingface/tokenizers/releases/tag/v0.22.1.
6. 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.
7. Ioannidis, John P. A.. (2005). Why Most Published Research Findings Are False. PLoS Medicine. https://doi.org/10.1371/journal.pmed.0020124.
8. 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.
9. 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.
10. Cook, R. Dennis. (1977). Detection of Influential Observation in Linear Regression. Technometrics. https://doi.org/10.1080/00401706.1977.10489493.
11. 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.
12. 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.
13. Carroll, Raymond J., David Ruppert, Leonard A. Stefanski, and Ciprian M. Crainiceanu. “Measurement Error in Nonlinear Models.” 2006. https://doi.org/10.1201/9781420010138.

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

15. 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.

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

17. Lin, Lawrence I-Kuei. (1989). A Concordance Correlation Coefficient to Evaluate Reproducibility. Biometrics. https://doi.org/10.2307/2532051.

18. 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.

19. 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.

20. 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.

21. 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.

22. 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.

23. 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.

24. 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.

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

26. 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.

27. 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.

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. 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. 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.