Unified Scaling Laws for Compressed Representations

Andrei Panferov [*1] **Alexandra Volkova** [*1] **Ionut-Vlad Modoranu** [1]
Vage Egiazarian [1] **Mher Safaryan** [1] **Dan Alistarh** [1 2]

Abstract

Scaling laws have shaped recent advances in machine learning by predicting model performance based on model size, computation, and data. Concurrently, the rise in computational cost for AI has motivated model compression techniques, notably quantization and sparsification, have become essential for large-scale training and inference. This paper investigates the interplay between scaling laws and compression formats, exploring whether a unified scaling framework can accurately predict model performance when training occurs over various compressed representations, such as sparse, scalar-quantized, or sparse-quantized. We validate a general scaling law formulation and show that it is applicable both individually but also composably across compression types. Our main result is demonstrating that there exists a simple "capacity" metric—based on to fitting random Gaussian data—which can robustly predict parameter efficiency across multiple representations.

1. Introduction

The idea of *predictable scaling* of learning performance with respect to *model, computation and data sizes*, encompassed by *scaling laws* [9], allows to predict the values of these three parameters required to reach a certain model performance. A parallel direction has been *model compression*, which proposes a series of techniques to reduce the computational and memory footprint of model inference and training, via techniques such as *sparsification* [7] and *quantization* [5]. Here, we focus on the interplay between scaling laws and the degree of compression of the representation over which learning occurs. While there is significant emerging work in this direction, e.g. [3; 10; 16; 15], current scaling laws are specialized to single representations (e.g., quantization or sparsity) and/or formats (e.g., integer quantization), and cannot yet address the question of predicting model scaling behavior when training over general compressed representations.

Contributions. This paper is structured two main questions:

Q1: Is there a unified compression scaling law? First, we wish to find a single general law that not only applies to sparse [3] or quantized [10] representations in isolation, but that also provides a good fit for *hybrid formats*, such as sparse-and-quantized weights, or *compound compression*, i.e. sparse weights *and* activations. Through extensive experimentation, we identify this law to be of the form

$$Loss(N, D) \sim A \cdot (N \cdot \rho(R))^{-\alpha} + B \cdot D^{-\beta} + E, \quad (1)$$

where N is the number of model parameters, D is the dataset size, E is the irreducible error, A, B, α and β are constants, and ρ is a parametric function of the representation R.

Crucially, we find that, even for very complex representations—e.g. 3-bit quantization with group size 32 and 1% outliers in full-precision—the parametric function ρ can still predict the scaling of model performance w.r.t. the parameter count N. We call $\rho(R)$ **the representation capacity** of R. Consequently, there is always a "dense equivalent" parameter count $N' = N \cdot \rho(R)$ which would yield the same loss during training.

Q2: Is capacity an "intrinsic" property of the representation? While related forms of the above law have been proposed in prior work [4; 10], we are the first show that capacity is an intrinsic property of the representation, independent of the model and task for which the scaling law is obtained, but relatable to standard information-theoretic measures. Moreover, we establish the applicability of the law across hybrid (e.g. sparse-quantized weights) or composite (e.g. quantized weights-and-activations) representations.

Our main finding is that **capacity is tightly-correlated with the representation's ability to fit random Gaussian data, measured in terms of minimal mean-squared error (MSE)**. Concretely, $\rho(R)$ is a simple parametric function of the MSE of the representation R when fitting random Gaussian data, i.e. $\rho(R) = \tilde{\rho}(MSE(R))$, where instances of the same representation R, e.g. 3 and 4-bit integer quantization, *share the same parametric form* $\tilde{\rho}$. This finding, which we validate across quantized, sparse, quantized-sparse, and even vector-quantized representations, provides a simple

[*]Equal contribution [1]ISTA [2]Red Hat AI. Correspondence to: Dan Alistarh <dan.alistarh@ist.ac.at>.

metric to "rank" different formats implementing the same representation. In addition, this also allows us to determine the "optimal" capacity at a certain bit-width, which is given by theoretical bounds on Gaussian fitting for a given support, which can be easily estimated via Monte Carlo algorithms.

Our second finding is that, except for pathological cases, *capacity factorizes across composite representations*: concretely, the capacity of a 4-bit and 2:4 sparse model is the product between the capacity of the 4-bit dense model, and that of a 2:4-sparse but unquantized model. Factorization allows us to evaluate the capacity of complex representations based on simple ones, and also holds when compressing different model representations.

Practical Implications. The analytical metrics suggested by representation capacity also have non-trivial practical applications. First, the fact that we are able to relate the predictive parameter ρ to intrinsic properties of the underlying representation gives us the ability to *analytically predict the representational power of different compressed numerical formats*. This way, we can accurately compare and predict the efficacy of various formats such as Floating-Point, Integer (INT with and without grouping), or sparse-quantized formats (2:4 + INT) at different compression budgets. Second, this framework inspires an improved approach for sparse training, which we show provides significant improvements (above 20% in some sparsity regimes) in capacity at the same number of parameters.

Overall, our results provide a new lens to view the scaling properties of compressed models, with respect to intrinsic properties of the representation over which training is performed. Thus, we believe that capacity-aware scaling has the potential to become a practical design principle for the next generation of efficient foundation models.

2. Preliminaries

Scaling Laws. We start from the "Chinchilla" scaling law formulation [8] that proposed to model loss scaling as a function of the number of parameters in the model N and the number of data points D the model was trained on, in the form the parametric function:

$$Loss(N, D) = AN^{-\alpha} + BD^{-\beta} + E, \tag{2}$$

where A, B, E, α, and β are the scaling law parameters that can be fit empirically. It is important to note that such scaling laws assume an ideal, well-tuned training setup, and that the parameter may vary slightly depending on architecture, optimizer, and hyper-parameters.

Compressed Representations. For *sparsity*, we assume that a specific fraction, within each parameter group of a certain size G, is set to zero. Sparsity is *unstructured* if the group is the whole tensor, whereas it is semi-structured (N:M) if N parameters out of every M are set to zero. For *quantization*, unless otherwise stated, we assume that parameters are mapped onto a *scalar, symmetric* grid corresponding to the number of bits available for quantization, as is standard [5]. (We will also consider vector quantization in Section 3.1.) For *sparse-quantized* representations, we follow [6] by first applying sparsification, and then quantization, to map continuous parameters onto this format.

Scaling Law Validation. For our scaling law investigations, we pretrained decoder-only Transformers following the Llama architecture [17] for 30M, 50M, 100M and 200M non-embedding parameters. Models were trained on the C4 dataset [14], using the Llama-2 tokenizer [17]. To ensure we operate in a data-rich regime, we use 100 training tokens per model parameter, and train on fixed-length context windows of 512 tokens. We used AdamW [11] with a 0.1 ratio of warm-up epochs with cosine scheduler. This is similar to the setups of Kumar et al. [10]; Frantar et al. [4].

We follow standard quantization-aware training (QAT) methods, combined with various levels of unstructured weight sparsity. For quantization we employ the gradient estimator of Panferov et al. [13], a per-layer uniform quantizer with static scaling factors and gradient masking. Quantization levels range from 1-bit to 8-bit precision. We consider configurations with quantized weights only, activations only, and both simultaneously. For sparsity, we apply unstructured magnitude pruning via top-k thresholding on a per-layer basis. The sparsity mask is recomputed dynamically at each optimization step. For Vector Quantization (VQ), we follow QuEST scalar quantization and apply it to 2- and 4-dimensional HIGGS grids [12].

3. Findings

3.1. Gaussian RMSE Predicts Representation Capacity

Table 1 presents a number of scaling laws that model the same functions via different parametrizations. One can notice, that both the *Sparsity* form of Frantar et al. [3] and the *Quantization* form of Kumar et al. [10] can be reduced to the *Decoupled* form of Frantar et al. [4] in the third row, by imposing additional constraints (e.g. $\text{eff}_P = 1 - e^{-P_w/\gamma_w}$ for quantization). Naturally, the Decoupled form can achieve lower fit error, but it does not provide any information about the interpretation of the capacity term, which we call $\rho(R)$, across different representations R. The *Sparsity* form and the *Quantization* form, on the other hand, feature intertwining and interpretable parameters. For simplicity, we first focus on the *Quantization* form for now.

The Functional Form. Kumar et al. [10] choose the functional form $\rho(P_w) = 1 - e^{-P_w/\gamma_w}$ to model quantization efficiency. By contrast, we propose a different form to model $\rho(R)$:

$$\widetilde{\rho}(GMSE(R)) = L \cdot \tanh(F \cdot \log_{1/4}(GMSE(R)))^C, \tag{3}$$

Table 1. Representation scaling laws (rows) versus the quantities of interest (columns). For all laws, N represents the number of parameters, D is the data, and E is the irreducible error. For the sparsity scaling law of Frantar et al. [2], S is the sparsity and the lowercase parameters are learnable constants. For the precision scaling law of Kumar et al. [10], P_w is the weight precision, and γ_P is a learnable weight sensitivity parameter. For the law of Frantar et al. [4], eff_C is the "effective parameter multiplier," that is explicitly fitted for every instance of compression C. By contrast, our formulation postulates that the parameter efficiency is a simple parametric function of the representation's capacity to fit random Gaussian data ($GMSE(R)$).

Parametrization	**Formulation for** $Loss(N, D)$	**Sparsity fit (Error)**	**Quantization fit (Error)**
Sparsity S Frantar et al. [3]	$\frac{a_S(1-S)^{b_S}+c_S}{N^{b_N}} + \left(\frac{a_D}{D}\right)^{b_D} + E$	$5.7 \cdot 10^{-4}$	N/A
Quantization to P_w Kumar et al. [10]	$A\left[N(1-e^{-P_w/\gamma_w})\right]^{-\alpha} + BD^{-\beta} + E$	N/A	$4.5 \cdot 10^{-3}$
Compression C Frantar et al. [4]	$\frac{A}{(N \cdot \text{eff}_C)^{\alpha}} + \frac{B}{D^{\beta}} + E$	$4.2 \cdot 10^{-4}$	$1.9 \cdot 10^{-3}$
Representation R (OURS)	$\frac{A}{(N \cdot \widetilde{\rho}(GMSE(R)))^{\alpha}} + \frac{B}{D^{\beta}} + E$	$4.7 \cdot 10^{-4}$	$2.1 \cdot 10^{-3}$

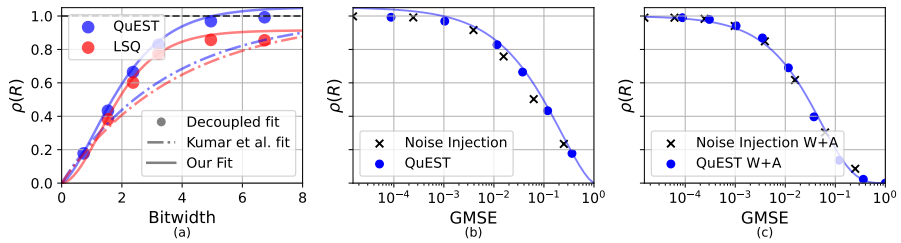


Figure 1. Comparison of ρ fits for scaling law forms from Table 1: **(a, left)** shows quantizations scaling laws, **(b, middle)** and **(c, right)** demonstrate the match between noise injection and QuEST quantization for weight-only and weights+activations quantization.

which depends only on the representation R's Gaussian-MSE fit, denoted by $GMSE(R)$, and on the scalars L, F, and C, detailed below. The $GMSE(R)$ is easily computable for any representation, and allows us to bypass the dependency on representation-specific parametrization, such as bit-width or sparsity. Specifically, we fit the scalar parameters for each compression type, e.g. scalar quantization, and then re-use these parameters while varying $GMSE(R)$ w.r.t. compression parameters, e.g. bit-width. The scalar parameters L, F, and C allow us to accurately model observed effects such as:

- **Imperfect convergence in high-precision:** While modern QAT algorithms such as QuEST reach efficiency $\rho = 1$ for low quantization error, older algorithms such as LSQ (Figure 1 (a)), have an efficiency limit strictly below 1, since for instance its gradient estimator introduces consistent bias. The factor L, defaulting to 1 for saturating representations, allows us to model this imperfection.
- **Various low-precision curvature:** As seen in Figures 1 (b) and (c), different representations behave differently around $GMSE = 1$, with some have noticeably higher curvature ("breakdown"). From Figure 1 (a), one can see how that region disproportionally affects the law of Kumar et al. [10], leading to a very poor fit at higher bitwidths. The factor C, closer to 1 for representations "more linear" around $GMSE = 1$, allows us to more accurately model $\rho(R)$.

Quality of Fit. Table 1 shows that our approach leads similar or better quality-of-fit relative to prior laws, covering both scalar quantization and sparsity, while Figure 1 shows $\rho(R)$ alignment between scaling law forms, compared to Kumar et al. [10], for the QuEST and LSQ quantizers. Again, our approach provides significantly better fit. In Figure 4(a), we show that our method can also provide a good fit for models trained with vector-quantized (VQ) weights, using the projection method of [12], for lattice dimensions 2 and 4. This shows both the versatility of our approach, and the necessity of the L term, since higher-dimensional VQ appears to have clear sub-unit saturation due to higher bias. We provide further examples in Section 3.4.

This result allows for low-cost comparison across compression comparison. Moreover, it facilitates compression hyperparameter tuning and thus predictable model training in a compressed regime.

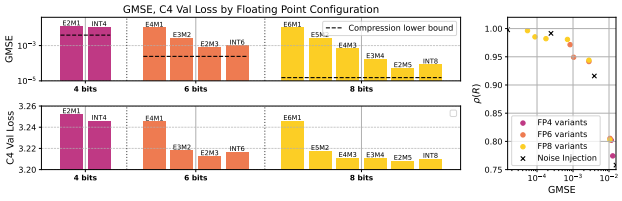


Figure 2. Comparison of floating point and integer data-types in terms of $GMSE$, and C4 Validation Loss when trained using the corresponding formats via QuEST, and the resulting capacity $\rho(R)$. Observe the high correlation between ranking in terms of $GMSE$ (top), and Val. Loss (bottom).

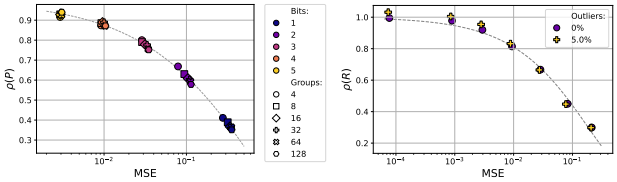


Figure 3. Representation capacity $\rho(R)$ versus MSE for **(a)** group-wise quantization, with markers indicate group counts (color encodes quantization bitwidth), and **(b)** outlier-aware quantization.

3.2. Comparing Compressed Numerical Formats

Practical Formats. The scaling law enables systematic comparison of numerical formats such as INT8, INT4, FP4, or custom low-precision representations, based just on their $GMSE$, which can be determined via fast Monte Carlo algorithms. Figure 2 illustrates this for a number of floating-point and integer data-types. Specifically, we observe a direct correlation between the ranking of $GMSE$ values (top) and the C4 validation loss obtained in actual experiments (bottom). This suggests that our $GMSE$ metric is an accurate predictor of compressed pre-training performance. For instance, it suggests that switching to FP4 (E2M1) will *not* bring gains relative to INT4 training, and that both formats are close to the theoretical lower bound at 4 bits.

3.3. Noise Injection is a Scaling Law Predictor

Next, we ask: what if we plug the *optimal* achievable $GMSE$ for a certain bit-width into the scaling law? Then, the scaling law should allow us to compute a *lower bound* on the achievable parameter efficiency. In turn, we can find out how close existing training techniques or numerical formats, are to the information-theoretic lower bound for that specific representation.

Figure 1 (b) illustrates the "optimality gap" for the QuEST algorithm for scalar weight-only quantization across bit-widths, suggesting that this approach is fairly close to optimal. In Figure 1 (c), we compare the fit between actual runs of this QAT algorithm across bit-widths, and the predicted values via noise injection [1] (plugging in the equivalent $GMSE$) into the scaling law, showing a near-perfect fit.

3.4. Representation Capacity Is Multiplicative Across Compression Types

In prior work, Kumar et al. [10] have claimed that, for their formulation of the law, the representation capacity factorizes independently for quantization of weights and activations. Our experimental findings extend this result, showing that representation capacity, $\rho(R)$, also factorizes naturally across a wide range of compression approaches, whether for the same tensor (sparse-and-quantized weights) or for different state tensors (sparse weights and sparse activations). We fit a scaling law in the 100 toks/param regime, and show that representation capacity factorizes for:

1. **Sparse weights and activations:** For sparsity, independently applied to weight and activations,
$$\rho(R_{s_w,s_a}) = \rho(R_{s_w}) \cdot \rho(R_{s_a}). \tag{4}$$
We summarize the fitted values of $\rho(R)$ levels in a matrix M (Figure 4(b)), where each entry corresponds to the fitted efficiency for a model trained with a specific sparsity configuration. Remarkably, the matrix can be accurately approximated by a rank-1 outer product of the first column $M_{0,:}$ (weight-only) and the top row $M:,0$ (activations-only) elements, i.e. $M \approx M_{0,:} \otimes M_{0,:}$. The resulting parameter efficiencies closely match the product of efficiencies obtained for runs with weight-only and activations-only configurations.

2. **Sparse and quantized weights:** Given a weight sparsity level s combined with q-bit quantization, we claim that the representation capacity can be represented as the product: $\rho(R_{q,s}) = \rho(R_q) \cdot \rho(R_s)$. We report the results for different sparsity levels and bit width in Figure 8. Similarly, the matrix $\rho(R)$ factorizes into the outer product of marginal vectors for quantization-only and sparsity-only representation.

3. **Sparse and quantized weights, and quantized activations:** Finally, we observe that factorization extends to quantization of activations as well. In supplementary experiments, we apply quantization to activation tensors alongside with weight sparsity and quantization. Our results indicate that the representation capacity with weight sparsity s_w and quantization bitwidth q_w, and activation sparsity q_a follows $\rho(R_{q_w,s_w,q_a}) = \rho(R_{q_w}) \cdot \rho(R_{s_w}) \cdot \rho(R_{q_a})$.

The Impact of Parameter Grouping and Outlier Preservation. A related question regarding formats is whether more complex approaches, such as group-wise quantization, or outlier preservation in higher precision, can disrupt the scaling law. We examine this in Figure 3, which it shows that preserving no outliers (0 %) lies on the Pareto-optimal boundary: higher outlier ratios achieve a worse trade-off between the MSE and the representation capacity $\rho(R)$. This suggests that, for pre-training it is more effective to allocate

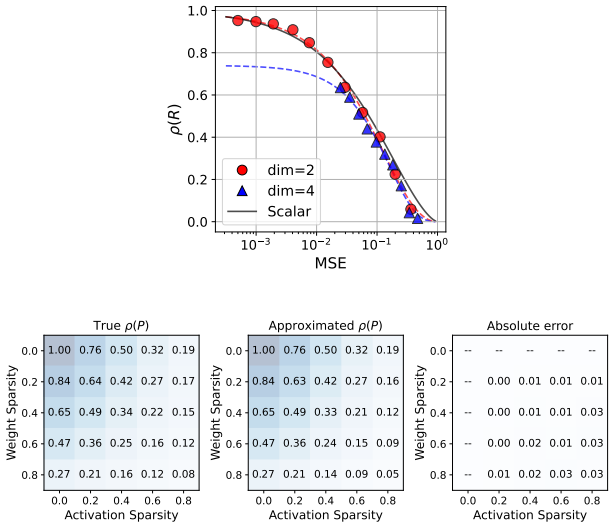


Figure 4. **(a)** Scaling law for 2- and 4-dimensional vector quantization. **(b)** Representation capacity across weight and activation sparsity levels: baseline, factorized prediction, and relative errors. Note the low errors for the factorized predictions, with slight increases at the larger sparsity levels.

bits to encoding the values distribution rather than outlier preservation or careful grouping. This further demonstrates that the proposed RMSE dependency is a general property and remains valid even under diverse structured compression techniques.

Compositionality. An immediate practical application of the multiplicative behavior of the law (Section 3.4) is the ability to estimate the model's performance in advance for arbitrary compression configuration. Given the individual efficiencies of different compression methods, such as quantization or sparsity, applied to weights or activations, one can predict the combined effect without spending additional compute for training.

4. Discussion and Limitations

Our study introduces *representation capacity*—roughly defined as a simple monotone transform of the Gaussian MSE—as a unified metric when training compressed models across various representations. Capacity enables format comparisons without retraining or exhaustive grid searches, so that future hardware designers can expose any format whose capacity ρ dominates the Pareto frontier, confident that software will exploit it optimally. Moreover, our law *factorizes*, further simplifying the search for the "optimal" training format.

A few caveats remain. First, in line with prior work in this area, our experiments are limited to decoder-only Llama-style architectures trained on C4 in the data-rich regime (100 toks/param); we plan to extend this at larger scale. Second, the law may need specific fits for ultra-low precision (e.g. 2-bit or ternary formats) and for vector-quantization codebooks below 8 entries, suggesting second-order effects may need to be taken into account. Third, while our theoretical evidence uses standard assumptions, it could be extended to more complex representation types.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

[1] Baskin, C., Liss, N., Schwartz, E., Zheltonozhskii, E., Giryes, R., Bronstein, A. M., and Mendelson, A. Uniq: Uniform noise injection for non-uniform quantization of neural networks. *ACM Transactions on Computer Systems (TOCS)*, 37(1-4):1–15, 2021.

[2] Frantar, E., Ashkboos, S., Hoefler, T., and Alistarh, D. Gptq: Accurate post-training quantization for generative pre-trained transformers. *arXiv preprint arXiv:2210.17323*, 2022.

[3] Frantar, E., Ruiz, C. R., Houlsby, N., Alistarh, D., and Evci, U. Scaling laws for sparsely-connected foundation models. In *International Conference on Learning Representations*, 2024.

[4] Frantar, E., Evci, U., Park, W., Houlsby, N., and Alistarh, D. Compression scaling laws:unifying sparsity and quantization, 2025. URL https://arxiv.org/abs/2502.16440.

[5] Gholami, A., Kim, S., Dong, Z., Yao, Z., Mahoney, M. W., and Keutzer, K. A survey of quantization methods for efficient neural network inference. In *Low-power computer vision*, pp. 291–326. Chapman and Hall/CRC, 2022.

[6] Harma, S. B., Chakraborty, A., Kostenok, E., Mishin, D., Ha, D., Falsafi, B., Jaggi, M., Liu, M., Oh, Y., Subramanian, S., and Yazdanbakhsh, A. Effective interplay between sparsity and quantization: From theory to practice. In *International Conference on Learning Representations*, 2025. arXiv:2405.20935.

[7] Hoefler, T., Alistarh, D., Ben-Nun, T., Dryden, N., and Peste, A. Sparsity in deep learning: Pruning and growth for efficient inference and training in neural networks. *Journal of Machine Learning Research*, 22 (241):1–124, 2021.

[8] Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., de Las Casas, D., Hen-

dricks, L. A., Welbl, J., Clark, A., Hennigan, T., Noland, E., Millican, K., van den Driessche, G., Damoc, B., Guy, A., Osindero, S., Simonyan, K., Elsen, E., Vinyals, O., Rae, J. W., and Sifre, L. Training compute-optimal large language models. In *Advances in Neural Information Processing Systems*, 2024.

[9] Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[10] Kumar, T., Ankner, Z., Spector, B. F., Bordelon, B., Muennighoff, N., Paul, M., Pehlevan, C., Ré, C., and Raghunathan, A. Scaling laws for precision. *arXiv preprint arXiv:2411.04330*, 2024.

[11] Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

[12] Malinovskii, V., Panferov, A., Ilin, I., Guo, H., Richtárik, P., and Alistarh, D. HIGGS: Pushing the limits of large language model quantization via the linearity theorem. In Chiruzzo, L., Ritter, A., and Wang, L. (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 10857–10886, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. URL https://aclanthology.org/2025.naacl-long.543/.

[13] Panferov, A., Chen, J., Tabesh, S., Castro, R. L., Nikdan, M., and Alistarh, D. Quest: Stable training of llms with 1-bit weights and activations. *arXiv preprint arXiv:2502.05003*, 2025.

[14] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pp. 13934–13944. PMLR, 2020. URL https://arxiv.org/abs/1910.10683. T5 and C4 Dataset.

[15] Sardana, N., Portes, J., Doubov, S., and Frankle, J. Beyond chinchilla-optimal: Accounting for inference in language model scaling laws. In *International Conference on Machine Learning*, 2024.

[16] Sun, X., Li, S., Xie, R., Han, W., Wu, K., Yang, Z., Li, Y., Wang, A., Li, S., Xue, J., Cheng, Y., Tao, Y., Kang, Z., Xu, C., Wang, D., and Jiang, J. Scaling laws for floating-point quantization training. *arXiv preprint arXiv:2501.02423*, Jan 2025.

[17] Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P. S., Lachaux, M.-A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E. M., Subramanian, R., Tan, X. E., Tang, B., Taylor, R., Williams, A., Kuan, J. X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., and Scialom, T. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

A. Appendix Roadmap

This appendix provides supporting material organized as follows:

- **Experimental Setup** (Appendix B): Model architectures, hyperparameters, and training configurations.
- **Factorization of Representation Capacity** (Appendix C): Detailed analysis showing how representation capacity matrices can be factorized for various compression techniques including quantization, sparsity, and their combinations.
- **Ablation Studies on Law Formulation** (Appendix D): Investigation of different noise distributions (Gaussian, Logistic, Student's t, Laplace) and functional forms (tanh, logistic) for the scaling law formulation.
- **Scaling Laws for Vector Quantization** (Appendix E): Implementation details and algorithms for vector quantization approaches, including forward and backward pass descriptions for HIGGS-based training.
- **Breaking the Scaling Law** (Appendix F): Demonstration of how training-time overparameterization with learnable block diagonal matrices can exceed FP16 performance ($\rho > 1$).

B. Experimental setup

Hyperparameters. Table 2 summarizes the architectural and training hyperparameters for each model size.

Model size	# Layers	# Heads	# Embeddings	Learning rate
30 M	6	5	640	$1.2 \cdot 10^{-3}$
50 M	7	6	768	$1.2 \cdot 10^{-3}$
100 M	8	8	1024	$6 \cdot 10^{-4}$
200 M	10	10	1280	$3 \cdot 10^{-4}$

Table 2. Key training hyperparameters for each model size.

We use 8x80GB H100 machines for efficient training, and training one model takes on average 1 hour. To produce the full set of results we ran in total approximately 250 such training runs for various compression configurations.

C. Factorization of Representation Capacity

Figures 5-8 show factorization of the representation capacity matrix for various in-training compression techniques:

1. Quantized weights and activations (Fig. 5).
2. Sparsity + QuEST quantizer (Fig. 6).
3. Joint sparse & quantized weights + activations (Fig. 7), for all combinations (s_a, q_a, q_b) for sparsity $s_a \in [0.25, 0.5, 0.75]$ and bit widths $q_a, q_b \in [2, 4, 6]$.
4. Sparsity + uniform quantizer with maximum absolute value as a scale (Fig. 8).

From the factorized representation-capacity matrices we observe the following:

1. The element-wise error of the fitted coefficients ρ (from our scaling law) is of order 10^{-3}–10^{-2}.
2. A rank-1 row-column outer product accurately approximates the matrix, confirming the multiplicative property of representation capacity ρ in various scenarios.
3. Approximation error remains of the order 10^{-2}, except for the cases of *extreme* 2-bit quantization, where $\rho \lesssim 0.1$. We explain this gap due to the poorer performance of the optimizer in these extreme compression regimes, which is not taken into account currently by our model (as it uses the same coefficients for both 16 and 2 bits).

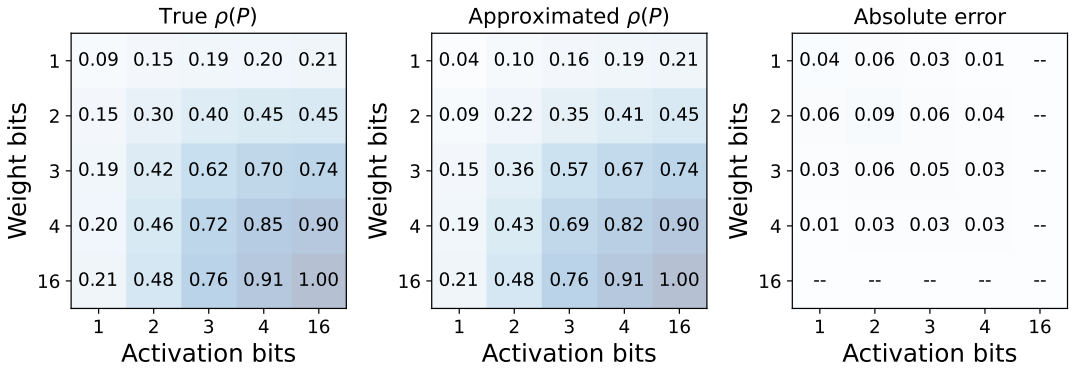


Figure 5. Representation capacity coefficients for independent quantization of weights and activations. Element-wise ρ fitting error is not greater than $5 \cdot 10^{-3}$.

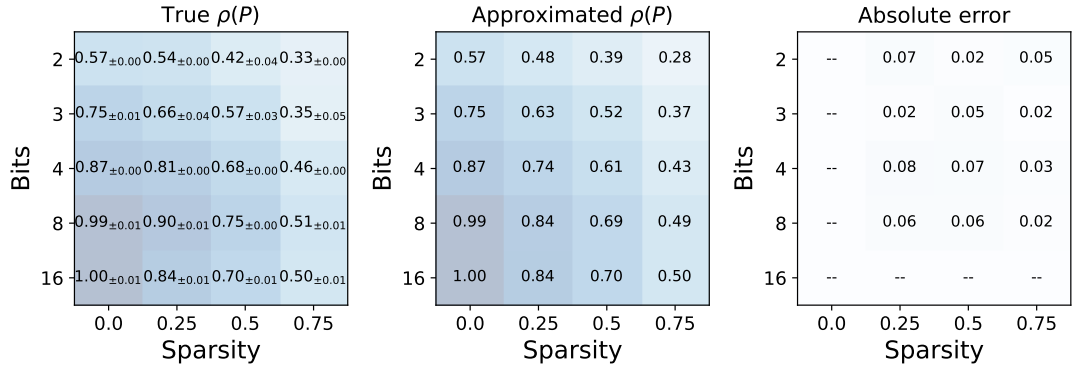


Figure 6. Representation capacity coefficients with fit errors in case of sparsity combined with the QuEST quantization.

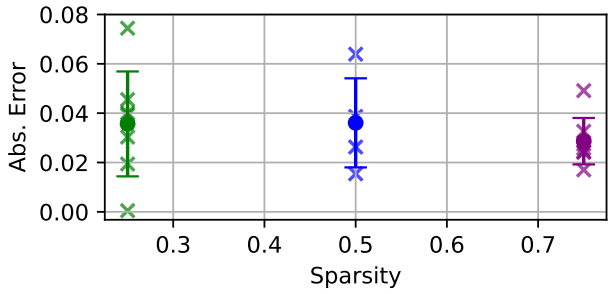


Figure 7. Representation capacity fit errors for sparse+quantized weights and quantized activations. Error bars denote ± 1 standard deviation from the mean.

D. Ablation studies on Law Formulation

D.1. Evaluating RMSE across Different Distributions

We investigate how the choice of noise distribution used in our law formulation from Sec. 3.1 affects the predicted representation capacity. In Figure 9a we plot the mapping $\rho(MSE)$ for different bit widths using Logistic, Student's t, and

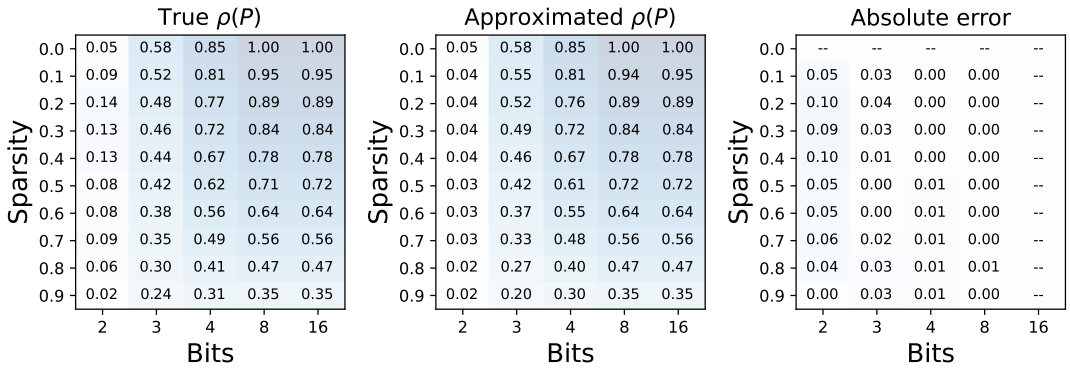
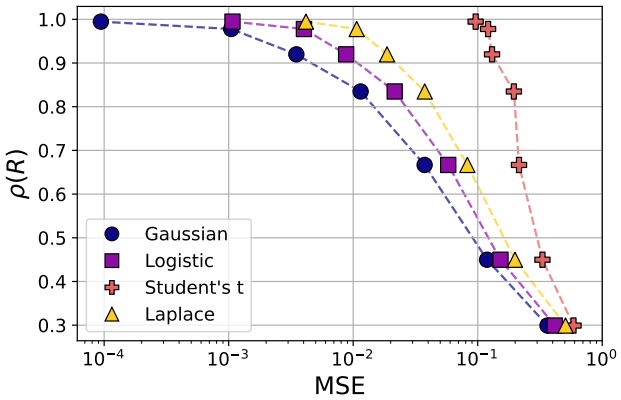
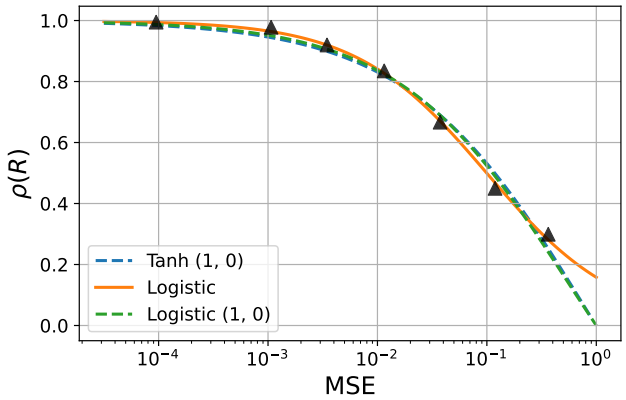


Figure 8. Representation capacity coefficients matrix for sparsity applied with uniform quantization. Element-wise ρ fitting error is not greater than $2 \cdot 10^{-3}$.



(a) Effect of input noise distribution on the mapping $\rho(MSE)$.



(b) Different functions used to fit $\rho(MSE)$

Laplace noise distributions. Each distribution is rescaled to have zero mean and unit variance.

We observe that, no matter which noise distribution we choose, the mapping $\rho(MSE)$ always remains strictly monotonically decreasing. In principle, one could use heavy-tailed distributions (for example, Student-t or Laplace) to give more weight to extreme outlier errors. However, this leads to a smaller range of MSE values. By contrast, assuming Gaussian noise—which we propose—produces the widest spread of MSE, which in turn allows for a better fit for the scaling law. In short, although monotonicity is preserved under various distributions, the Gaussian MSE delivers the best overall representation capacity prediction, so we adopt it as our default formulation.

Throughout this work, unless specified otherwise, MSE is computed over standard Gaussian input.

D.2. Functional form of the Law

The behavior of $\rho(GMSE)$ observed in our experiments can be captured by fitting multiple smooth, monotonically decreasing functions, with no more than 2 additional parameters. In principle, a wide range of such functions can be used to model this relationship, depending on the desired fit properties.

For lower overall fitting error, we found it beneficial to constrain the function to satisfy boundary conditions $f(0) = 1$ and $f(\infty) = 0$. For instance, the logistic form $\frac{1}{1+\exp(a \cdot \log(\text{MSE}+b)} = \frac{1}{1+B \cdot \text{MSE}^A}$ provides a good empirical fit, as shown in Figure 9b for weight quantization across 1-8 bit widths.

In cases where it is important to constrain MSE below 1, one may instead prefer the condition $f(1) = 0$. Although this typically results in a worse overall fit, it enforces the correct behavior in the high-error region $MSE \lesssim 1$, which is critical for stable predictions in the extreme compression cases. The corresponding fits, including those constrained at $f(1) = 0$, are

summarized in Table 3 and visualized in Figure 9b.

	Functional form	Fitting error (MSE)
Tanh	$\tanh(F \cdot \log_{1/4} \text{MSE})^C$	$1 \cdot 10^{-3}$
Logistic	$(1 + B \cdot \text{MSE}^A)^{-1}$	$1 \cdot 10^{-4}$
Logistic (1, 0)	$\frac{1 - \text{MSE}^A}{1 + B \cdot \text{MSE}^A}$	$1 \cdot 10^{-3}$

Table 3. Functional form choices and associated fitting error.

The choice of functional form reflects the trade-off between global fit quality and targeted accuracy for larger MSE values. Throughout this work, we adopt the constraint $f(1) = 0$ and functional form of hyperbolic tangent. As for the exact functional form, under the stated constraints, we find that the the specific choice between tanh and logistic sigmoid has little effect on overall fit quality.

E. Scaling Laws for Vector Quantization

In this section, we provide detailed information about the Vector Quantization approach used to produce the results in Figure 4(a). Algorithms 1 and 2 describe the forward and backward passes over a linear layer actively quantized with HIGGS for row-major weights. As was described earlier, our method is combines ideas from Panferov et al. [13] for the gradient estimator, and Malinovskii et al. [12] for the lattice representation. We use the trust estimation method that zeros out gradients for any point lying outside a hypersphere of radius R: $\|x\|_2^2 > R^2$. Our experiments were conducted on 30M and 50M models using the same set of hyperparameters as in Sec. 2.

Algorithm 1 `VQ Training Forward`

Input: Input activations $\mathbf{x}$, row-major weight $\mathbf{w}$
$\mathbf{w}_h = \text{HT}(\mathbf{w})$
$\hat{\mathbf{w}}_h = \text{proj}_{grid}\, \mathbf{w}_h$
$\mathbf{y} = \mathbf{x}\hat{\mathbf{w}}_h^T$
Return: $\mathbf{y}$, $\mathbf{x}$, $\hat{\mathbf{w}}_h$, $M_{grid}(\mathbf{w}_h; \hat{\mathbf{w}}_h)$

Algorithm 2 `VQ Training Backward`

Input: $\frac{\partial L}{\partial \mathbf{y}}$, $\mathbf{x}$, $\hat{\mathbf{w}}_h$, $M_{grid}(\mathbf{w}_h; \hat{\mathbf{w}}_h)$
$\frac{\partial L}{\partial \mathbf{x}} = \frac{\partial L}{\partial \mathbf{y}} \hat{\mathbf{w}}_h$
$\frac{\partial L}{\partial \hat{\mathbf{w}}_h} = \mathbf{x}^T \frac{\partial L}{\partial \mathbf{y}}$
$\frac{\partial L}{\partial \mathbf{w}} = \text{IHT}\left(M_{grid}(\mathbf{w}_h; \hat{\mathbf{w}}_h) \odot \frac{\partial L}{\partial \hat{\mathbf{w}}_h}\right)$
Return: $\frac{\partial L}{\partial \mathbf{x}}$, $\frac{\partial L}{\partial \mathbf{w}}$

F. "Breaking" the Scaling Law by Training-Time Overparametrization

One observation stemming from our law is that it is possible to overparameterize the model during training while keeping the number of inference-time parameters the same. This can be achieved by multiplying the weight matrix W by a learnable block diagonal matrix R, which is then "folded" into the model at inference time. The forward pass takes the form of $X(RW)^T$ during training and XW^T at inference. While R is omitted during evaluation, maintaining the original model size, additional parameters add flexibility during training and improve the representation quality.

For each weight matrix, we initialize our rotation matrix with a block-diagonal Hadamard, with the block sizes equal to 128, and learn it alongside W using our loss function. We train the 30M and 50M models using the same experimental setup as in Sec. 2 with rotation matrices, calculate the effective representation capacity, and compare it to baseline.

We observed that it results in a representation capacity $\rho = (1.07 \pm 0.04)$, indicating that model trained with such overparameterization outperforms the bf16 baseline ($\rho = 1$), even though their inference costs are identical.