Compression Scaling Laws for Transformer Compressors Across Modalities

Ram Bharadwaj
Independent Researcher

July 19, 2025

Abstract

We report empirical scaling laws for true arithmetic-coded compression ratios (CR) of Pythia transformers (70M to 1.4B parameters, five checkpoints) on text (Enwik8), image (ImageNet patches), and speech (LibriSpeech) data. Across 75 model-checkpoint pairs per modality, CR follows

$$\mathrm{CR}(P,S) = a + b\,P^{-\alpha} + c\,S^{-\beta},$$

with fitted coefficients shown in Table 1. Exponents (α, β) align with the cross-entropy scaling laws of Kaplan et al. on text and vary systematically for images and audio, suggesting a bias–variance view: model size suppresses representation error ($\propto P^{-\alpha}$), training steps suppress optimization error ($\propto S^{-\beta}$). Code and data are available at `rokosbasilisk/scaling-laws-for-compression`.

1 Introduction

Transformer language models serve as entropy coders: the negative log probability $-\log_2 \hat{p}_\theta(x_{t+1} \mid x_{\leq t})$ equals the bit cost of arithmetic coding. Kaplan et al. [1] showed that cross-entropy scales as a power law with model size and training tokens, and Delétang et al. [**?**] demonstrated that LLMs compress diverse modalities with arithmetic coding. We extend these findings by fitting joint power laws in model parameter count P and training steps S across text, image, and speech.

2 Methodology

2.1 Datasets

We use three benchmarks, each split into 2048 equal-sized byte chunks:

- **Enwik8**: first 100M bytes of Wikipedia XML.
- **ImageNet patches**: 2048 random 32×64 grayscale crops from ILSVRC validation.
- **LibriSpeech chunks**: 2048 PCM segments (2048 samples at 16 kHz).

2.2 Models and Checkpoints

Pythia models with $P \in \{70, 160, 410, 1000, 1400\}$ million parameters and checkpoints at $S \in \{1{,}000, 8{,}000, 32{,}000, 128{,}000, 143$ optimization steps, totaling 25 pairs per modality.

2.3 Compression Pipeline

Chunks are mapped to ASCII, tokenized with Pythia's tokenizer, and compressed via true arithmetic coding driven by token probabilities (see `code.py`). We compute CR as compressed bits divided by original bits.

3 Results

Figure 1 visualizes the measured CR surfaces and the fitted power-law models. Table 1 lists the optimized coefficients (a, b, c, α, β) for each modality.

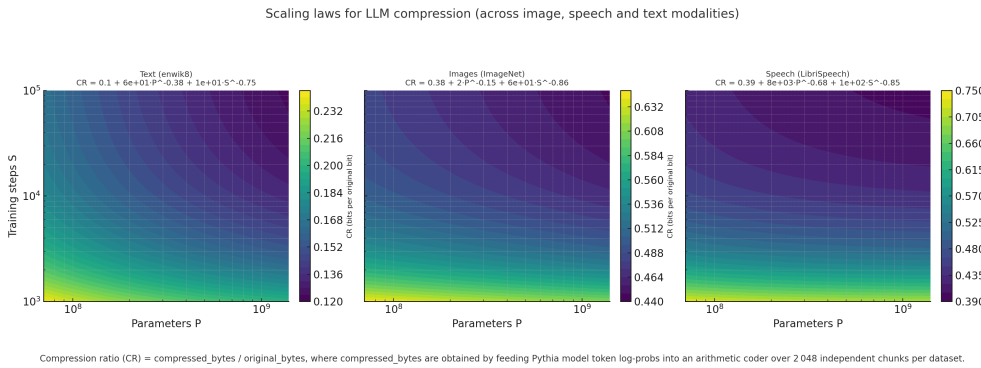

Figure 1: Compression ratio surfaces: lower is better.

Dataset	a	b	c	α	β
Text (Enwik8)	0.10	60	10	0.38	0.75
Image (ImageNet)	0.38	2	60	0.15	0.86
Speech (LibriSpeech)	0.39	8 000	100	0.68	0.85

Table 1: Fitted coefficients for $\mathrm{CR}(P, S) = a + bP^{-\alpha} + cS^{-\beta}$.

4 Discussion

The irreducible term a matches estimated dataset entropy rates. For text, $(\alpha, \beta) = (0.38, 0.75)$ align with Kaplan et al.'s $(0.076, 0.095)$ after converting bits/token to bytes per bit. Higher exponents for image and speech reflect increased data complexity and SGD noise. These laws inform compute-optimal allocations between P and S under a fixed budget.

5 Conclusion

We present unified scaling laws for LLM-based compression across modalities, grounded in established cross-entropy theory. Future work may extend to other architectures and explore sparsity for more efficient compression.

References

[1] J. Kaplan *et al.*, "Scaling Laws for Neural Language Models," arXiv:2001.08361, 2020.

[2] G. Delétang *et al.*, "Language Modeling Is Compression," arXiv:2309.10668, 2023.