A Influence of token frequency imbalance on unigram and language models

In this section, we provide a detailed explanation of our research question. Natural language exhibits strong contextual dependencies rather than behaving as an i.i.d. process: each token's probability depends on its preceding context. As a result, the entropy rate, $H_\infty = \lim_{t\to\infty} H\big(X_t \mid X_{<t}\big)$ which captures the optimal per-token uncertainty in text, is strictly lower than the unigram Shannon entropy $H_1 = -\sum_w p(w)\,\log p(w)$. Although entropy rate and shannon entropy coincide for truly i.i.d. data, in natural language they can differ by several bits per token.

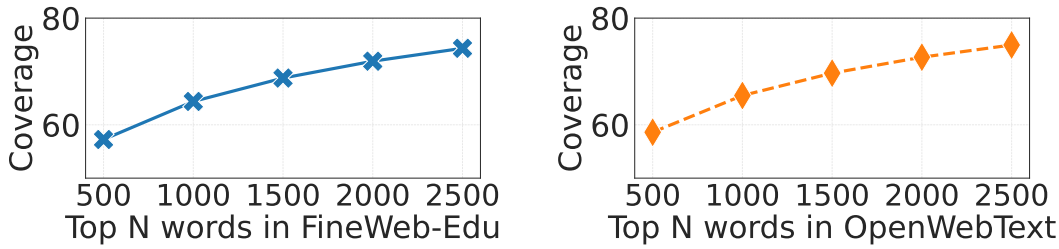
Under a pure unigram model, cross-entropy loss exactly equals Shannon entropy, so modifying the vocabulary size of tokenizer immediately changes the loss. Typically, enlarging the vocabulary segments frequent multi-token patterns into single tokens, driving their individual probabilities up and reducing shannon entropy. But if the vocabulary size grows too large, the new entries tend to be rare tokens, which lengthen the tail and can actually increase the shannon entropy as token frequency imbalance rises [53]. However, experimental results show that expanding a BPE vocabulary to around $80\mathrm{k}$ lowers the Shannon entropy of unigram models, demonstrating that a more skewed token-frequency distribution is advantageous at practical vocabulary scales [37].

Language model loss $\mathcal{L}(\theta)$ minimizes

$$\mathcal{L}(\theta) = H_\infty \;+\; \sum_{x\in V} p(x)\, D_{\mathrm{KL}}\big(P(\cdot \mid x_{<t}) \,\|\, Q_\theta(\cdot \mid x_{<t})\big). \tag{5}$$

where V denotes the tokenizer's vocabulary, $p(x)$ the marginal probability of token x, $P(\cdot \mid x_{<t})$ the true next-token distribution given the full history $x_{<t}$, and $Q_\theta(\cdot \mid x_{<t})$ the model's predicted distribution with parameters θ. When the target label is a one-hot vector, the language model loss can be written as $\mathcal{L}(\theta) = \sum_{x\in V} p(x)\,\big[-\log Q_\theta(x \mid x_{<t})\big]$ so it is a marginal-frequency-weighted average of the model's surprisal $-\log Q_\theta$. By contrast, the Shannon entropy of the unigram model is a weighted average of the self-information $-\log p(x)$ in the dataset. Even though frequent token logits and embedding norms are higher than rare ones (see §4.1 and Appendix D), the loss $-\log Q_\theta$ depends not only on the underlying normalized token frequencies but also on training dynamics as well. We therefore cannot derive a closed-form expression to predict how much the loss on frequent tokens shrinks versus how much the rare token losses grow under $\mathcal{L} = p_{\mathrm{freq}}\, L_{\mathrm{freq}} + p_{\mathrm{rare}}\, L_{\mathrm{rare}}$. This masking effect makes it substantially harder to isolate and measure the true influence of normalized token-frequency imbalance in language models than in the unigram case.

B Coverage of the most frequent N words

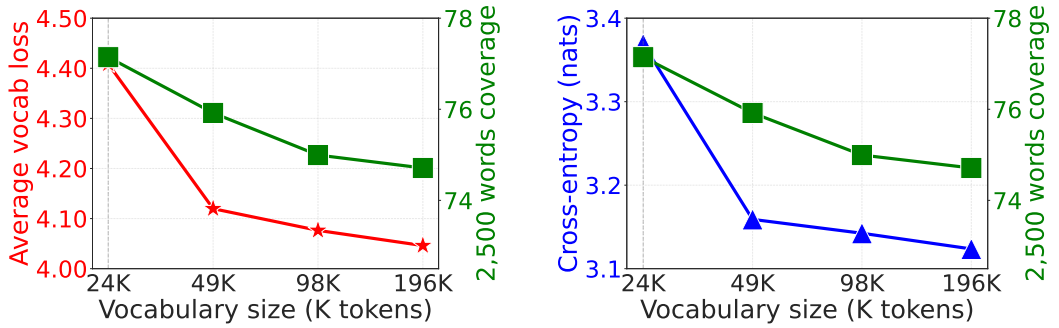


(a) (b)

Figure 6: Figures 6a and 6b illustrate the cumulative coverage of the 2,500 most frequent words in the Fineweb-edu and OpenWebText datasets, respectively.

In this section, we measure the coverage of the frequent words in Fineweb-edu [34] and OpenWebText [16]. Both dataset exhibit a steep rise in cumulative coverage as we include more high-frequency tokens, but with subtly different baselines and slopes. In figure 6a, the most frequent 500 words already cover about 58% of all tokens, climbing to roughly 75% once we take the $2,500$ most frequent words. OpenWebText (Figure 6b) starts marginally higher—around 59% at most frequent 500 words, but follows an almost identical trajectory, reaching about 76% coverage by the frequent $2,500$ words. This pattern underscores how a relatively small core vocabulary captures the vast majority of running text in both corpora, with only modest gains as we move deeper into the long tail.

C OpenWebText experiments results



(a) (b)

Figure 7: Figure 7 shows the results of experiments with OpenWebText dataset [16]. Figure 7a reveals that expanding the vocabulary from 24K to 196K steadily reduces the average per-vocabulary loss of high frequency words. Figure 7b indicates that the most frequent $2,500$ tokens still contribute roughly 75% of the total loss, while the rare words losses grow with vocabulary size, similar to 2b. Figure 7b further shows that the global cross-entropy loss falls by about 0.25 nats as the vocabulary grows, demonstrating that the reduction of loss on frequent words outweighs the inflation of rare-token losses.

To verify that reducing frequent-word loss is not a by-product of dataset quality, we repeat the same experiments in section 3.5 on the OpenWebText dataset. Figure 7a shows that widening the vocabulary from 24K to 196K in OpenWebText progressively reduces the average loss assigned to high-frequency words. Figure 7b indicates that the most frequent $2,500$ tokens still account for about 75% of the total loss, whereas the loss on infrequent tokens grows with vocabulary size, paralleling the pattern seen in Figure 2b. Figure 7b further demonstrates that the global cross-entropy loss falls from 3.37 nats at a 24K vocabulary to 3.12 nats at 196K, implying that the reduction in loss on frequent words more than offsets the increase in rare-token loss, regardless of dataset quality or type.

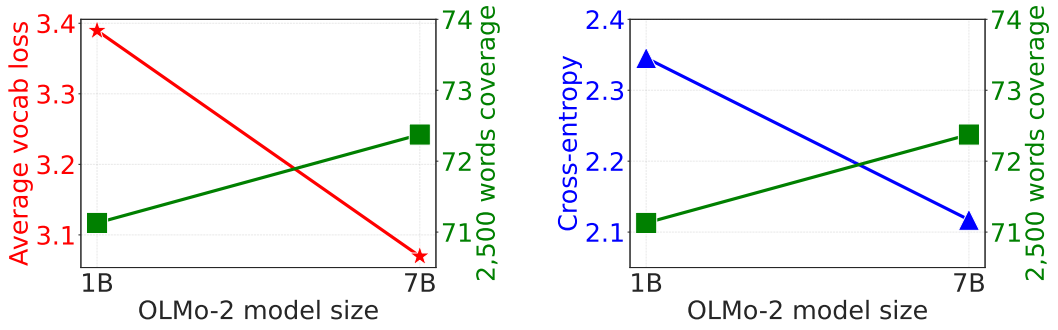
D Frequent and rare token norm in output embedding

In this section, we explain why frequent tokens acquire larger output-embedding norms and logits by deriving and analysing the gradients of the output embedding. Cross-entropy loss with a vocabulary size $|V|$ and hidden dimension size d_{model}, using a one-hot target $t \in \mathbb{R}^{|V|}$, the logit vector is $\ell = hE_{\text{out}}^{\top}$ where $h \in \mathbb{R}^{d_{\text{model}}}$ is the final hidden state and $E_{\text{out}} \in \mathbb{R}^{|V| \times d_{\text{model}}} = [u_1, \ldots, u_{|V|}]$ is the output-embedding matrix. When input and output embeddings are untied, the gradient with respect to each row of E_{out} decomposes into

$$\frac{\partial \mathcal{L}}{\partial E_{\text{out}_t}} = (p_t - 1)\, h, \qquad \frac{\partial \mathcal{L}}{\partial E_{\text{out}_j}} = p_j\, h \;\; (j \neq t), \tag{6}$$

where $p_t = \text{softmax}(\ell)_t$ [5, 30]. Because $p_t \ll 1$ at the start of training, $\left\|\partial \mathcal{L}/\partial E_{\text{out}_t}\right\|_2 \approx \|h\|_2$ while each non-target row scales only with $p_j \|h\|_2$ ($p_j < 1/|V|$). Thus, every time token t appears, its embedding is pushed almost $\|h\|_2$ units along $-h$, whereas each competing row is nudged by a factor of $p_j \ll 1$. As tokens recur in the training data, their token embeddings accumulate gradient updates roughly proportional to their counts, so $\|E_{\text{out}_t}\|_2$ grows in line with token frequency. Since each logit factorizes as $\ell_t = \|h\|_2 \, \|E_{\text{out}_t}\|_2 \cos\theta_t$ and frequent tokens not only acquire the largest norms but also align closely with final hidden-state directions $\cos\theta_t \approx 1$, they end up with disproportionately large logits whereas rare tokens suffer both smaller norms and larger angular deviations [5, 25]. Although ℓ_2 weight decay can slow this norm inflation, they merely dampen the effect rather than remove the underlying frequency norm logit correlation. This frequency-proportional amplification explains the empirical observation that output-embedding norms and softmax logits scale with token frequency in standard language models.

E OLMo-2 result



(a) (b)

Figure 8: Figure 8a illustrates that larger 7B model reduces average per-vocabulary loss from 3.39 nats to 3.07 nats while slightly increasing the proportion of loss covered by frequent words from 71% to 72.5%. Figure 8b further demonstrates that Scaling from 1B to 7B reduces the overall cross-entropy from 2.35 nats to 2.12 nats, confirming the same pattern persists in contemporary large-scale language models.

To identify whether the reduction in frequent words loss with increasing model size holds for contemporary large language models, we perform analogous experiments using the OLMo-2 series [31]. Figure 8a and 8b indicate that the average per vocabulary loss falls from 3.39 nats in the 1B parameter model to 3.07 nats in the 7B variant while slightly increasing the proportion of loss covered by $2,500$ frequent words from 71% to 72.5%. Figure 8b further shows that global cross-entropy loss declines from 2.35 nats for OLMo-2 1B to 2.12 nats for OLMo-2 7B. Notably, OLMo-2 employs a much larger vocabulary (cl100K [32]) than Pythia (50304 tokens [7]) and trains on a larger corpus [31], which helps drive down the average loss on high-frequency words. These results confirm that the same trend holds for modern large-scale language models as well.