Linguistic Collapse:
Neural Collapse in (Large) Language Models

Robert Wu
University of Toronto, Vector Institute
rupert@cs.toronto.edu

Vardan Papyan
University of Toronto, Vector Institute
vardan.papyan@utoronto.ca

Abstract

Neural collapse ($\mathcal{NC}$) is a phenomenon observed in classification tasks where top-layer representations collapse into their class means, which become equinorm, equiangular and aligned with the classifiers. These behaviours — associated with generalization and robustness — would manifest under specific conditions: models are trained towards zero loss, with noise-free labels belonging to balanced classes, which do not outnumber the model's hidden dimension. Recent studies have explored $\mathcal{NC}$ in the absence of one or more of these conditions to extend and capitalize on the associated benefits of ideal geometries. Language modelling presents a curious frontier, as *training by token prediction* constitutes a classification task where none of the conditions exist: the vocabulary is imbalanced and exceeds the embedding dimension; different tokens might correspond to similar contextual embeddings; and large language models (LLMs) in particular are typically only trained for a few epochs. This paper empirically investigates the impact of scaling the architectures and training of causal language models (CLMs) on their progression towards $\mathcal{NC}$. We find that $\mathcal{NC}$ properties that develop with scale (and regularization) are linked to generalization. Moreover, there is evidence of some relationship between $\mathcal{NC}$ and generalization independent of scale. Our work thereby underscores the generality of $\mathcal{NC}$ as it extends to the novel and more challenging setting of language modelling. Downstream, we seek to inspire further research on the phenomenon to deepen our understanding of LLMs — and neural networks at large — and improve existing architectures based on $\mathcal{NC}$-related properties. Our code is hosted on GitHub: https://github.com/rhubarbwu/linguistic-collapse.
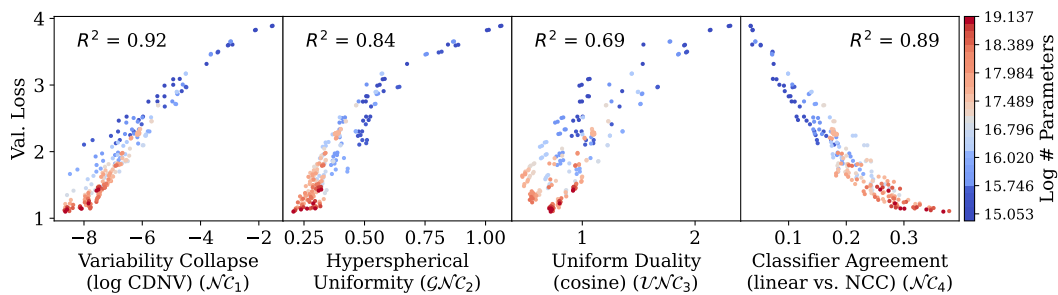
Figure 1: Simultaneous development of the four *neural collapse* ($\mathcal{NC}$) [1] properties in 230 causal language models trained on TinyStories [2], alongside improvement in generalization (i.e. validation performance). Left to right: $\mathcal{NC}_1$) within-class (representation) variability collapse; $\mathcal{GNC}_2$) hyperspherical uniformity of class means; $\mathcal{UNC}_3$) uniform duality between class means and corresponding classifiers; and $\mathcal{NC}_4$) agreement between token (maximum a prior) classifiers and implicit nearest-class centre classifiers. Coloured by model size and annotated with coefficient of determination (R^2).

1 Introduction

1.1 Neural Collapse

A learning phenomenon known as *neural collapse* ($\mathcal{NC}$) emerges during the terminal phase of training (TPT) deep neural networks with cross-entropy (CE) loss for classification tasks.[1] It was originally characterized as the co-occurrence of the following properties in a model's top-layer (also known as last-layer) representations (also known as features or embeddings) and linear classifier weights:

($\mathcal{NC}_1$) **Within-class variability collapse:** Top-layer representations collapse to their class means.

($\mathcal{NC}_2$) **Convergence to a simplex ETF:** Class means tend towards equinorm and equiangular vectors when centred about the global average. The resulting geometry — known as a simplex *equiangular tight frame* (ETF) — maximizes pairwise angles and distances.

($\mathcal{NC}_3$) **Convergence to self-duality:** Linear classifier weight vectors converge to their corresponding top-layer class mean vectors, and thus also form a simplex ETF.

($\mathcal{NC}_4$) **Nearest decision rule:** Linear classifiers approximate a nearest-class centre (NCC) classifier: top-layer representations predict the class with the closest mean (implied by $\mathcal{NC}$1-3).

These behaviours, often associated with improved generalization and robustness [3–5] (among other benefits, such as those discussed in §1.4), traditionally manifest under the following conditions:

Rq1) **Few classes:** The number of classes is upper-bounded by the embedding dimension plus one: $C \leq d + 1$; this is required to construct a perfect simplex ETF.

Rq2) **Balanced classes:** The number of samples is equal across classes: $N_c = N_{c'}, \forall c \neq c'$.

Rq3) **Noise-free labels:** Identical (or very similar) embeddings should belong to the same class.

Rq4) **Sufficient training (TPT):** The model is trained past zero error, towards zero loss.

Absent these conditions, one does not typically anticipate $\mathcal{NC}$. Since then, follow-up studies have extended beyond and proposed techniques of quantifying or achieving $\mathcal{NC}$ (discussed in Section 2).

1.2 (Large) Language Models

$\mathcal{NC}$ is a phenomenon observed specifically in classification tasks. While not traditionally thought of as classifiers, language models — including large language models (LLMs) — learn to model aleatoric uncertainty, which can be viewed as stochastic token prediction [6]. For instance, masked language models (MLMs) such as BERT [7] predict one or several masked tokens within an input sequence based on the surrounding *context*. Likewise, autoregressive or causal language models (CLMs) such as GPT [8] perform next-token prediction (NTP) in a sequence given the *context* of previous tokens. Most of these models are essentially *(pre-)trained* by token classification on their vocabularies. This parallel — also drawn by [9] — raises a few natural questions:

1. Does the pre-training stage of a language model exhibit $\mathcal{NC}$?

2. How do scaling and other training configurations influence $\mathcal{NC}$ in (L)LMs?

3. To what extent is $\mathcal{NC}$ in (L)LMs correlated with their generalization abilities?

4. Do such correlations, between $\mathcal{NC}$ and improved generalization, persist independent of the (potential confounders of) model size and training?

To address these, we first examine the specific settings of training CLMs as they are opposed ($\neg$) to the traditional prerequisites (R1-4, §1.1) for $\mathcal{NC}$ to manifest.

¬Rq1) **Many classes:** The unique tokens found in language modelling vocabularies are vast, usually numbering in the tens of thousands and far exceeding the hidden dimension [10].

¬Rq2) **Imbalanced classes:** The distribution of tokens in natural language is typically very imbalanced [11, 12], as is the case in TinyStories [2], the dataset we use (Appendix Figure 4). It has an average of 16K samples per class but a standard deviation of over 32K.

¬Rq3) **Ambiguous contexts:** There may exist very similar or even identical contexts followed by different tokens in the natural language data [13]. For instance, over half of the sequences in TinyStories [2] lead with "`Once upon a time`", but only three-quarters of these follow with a comma ("`Once upon a time,`"). In other words, there is almost certainly some *ambiguity* to contend with in our context embeddings.

¬Rq4) **Undertraining:** Most practical language models (including LLMs) are not trained for more than a few epochs [14, 15]. Further optimization typically renders diminishing returns in improving evaluation performance [16] long before any possible TPT.

1.3 Contributions

We train a suite of Transformer-based [17] CLMs[1] across a grid of model widths, depths, and training epochs on the TinyStories dataset [2] to assess the degrees to which $\mathcal{NC}$ properties develop and how they relate to generalization performance. Our findings (summarized in Figure 1) reveal:

- **Emergence of $\mathcal{NC}$ with scale:** As we scale model size and training, the properties of $\mathcal{NC}$ emerge; within-class variability ($\mathcal{NC}_1$) and interference are reduced while hyperpsherical uniformity ($\mathcal{GNC}_2$), uniform duality ($\mathcal{UNC}_3$), and classifier agreement ($\mathcal{NC}_4$) improve.

- **Progression towards hyperspherical uniformity:** Class means, while unable to form a simplex ETF ($\mathcal{NC}_2$), nonetheless tend towards uniform dispersion on a hypersphere, a geometry theorized by [18] and formalized by [19] as *hyperspherical uniformity* ($\mathcal{GNC}2$).

- **Tendency towards uniform duality:** Direct alignment (self-duality) between class means and classifiers ($\mathcal{NC}_3$) does not appear to develop with scale. Instead, the variability of (mis)alignment across classes decreases with width and training, suggesting its minimization – which we term *uniform duality* ($\mathcal{UNC}_3$) – may be more cohesive with $\mathcal{NC}$.

- **Correlation between $\mathcal{NC}$ and generalization:** The developments of $\mathcal{NC}$ properties are correlated with improved validation performance. We show these correlations to persist even when fixing the (potential confounders of) model architecture and training by simply varying the random seeds for initialization and data shuffling. This suggests that $\mathcal{NC}$ is not simply a side-effect of training, but possibly a factor of generalization in its own right.

1.4 Significance

Recently, methods building on $\mathcal{NC}$ have found use in deep learning at large. We highlight areas such as federated learning [20], graph neural networks [21], incremental/continual learning [22–24], meta-learning [24, 25], out-of-distribution detection [26–28], privacy [29, 30], learning theory [31–36] transfer learning [3, 5, 37–40], and even LLM prompting [41]. With our results, we aim to extend insights from such contributions and other related works to the autoregressive language modelling domain and ultimately assist researchers in improving and interpreting their (L)LMs.

2 Related Works

$\mathcal{NC}$ was initially observed in image classification tasks such as on CIFAR-10/100 [42] and ImageNet [43]. Since then, the phenomenon has been further studied both theoretically and empirically [5, 18, 35, 44–66], with several works venturing into settings without some of the traditional prerequisites (¬Rq1-4, §1.2) and proposing adaptations of the analysis framework or optimization procedures:

A large number of classes (¬Rq1) $\mathcal{NC}$ established the simplex ETF as an optimal configuration. However, a perfect simplex ETF ($\mathcal{NC}_2$) requires that the number of classes C not exceed $d + 1$ where d is the embedding dimension. This requirement that d be sufficiently large is impractical when the classes number beyond the thousands.[2] For instance, GPT-2 [68] and Llama 3.1 [69] have vocabularies of around 50K and 128K tokens, respectively.

In such a scenario, one might still expect class means to be uniformly distributed on a d-dimensional hypersphere [18]. [19] formalize this as *hyperspherical uniformity* within a *generalized neural*

---

[1]Our models range from 3.4M (small) to 205M (large), so we inclusively use "CLM" instead of "LLM".

[2]Following [67], one might describe such a setting ($C > d + 1$) as a model "in superposition".

collapse ($\mathcal{GNC}$) framework, which [9] then empirically demonstrate. These latter two works mention language modelling as applicable for $\mathcal{GNC}$; [9] even framed token prediction as a classification task, just as we do. We remark however that both earlier works *simulate* a large number of classes by drastically shrinking the embedding dimension. In contrast, we study realistic NTP, using the full class space (vocabulary) with an imbalanced token distribution and ambiguous samples.

Class imbalance (¬Rq2) $\mathcal{NC}$ traditionally assumed that classes were sample-balanced. Since then, follow-up works have investigated the effect of class imbalance on the prevalence of $\mathcal{NC}$. [47] studied the *layer-peeled model* (LPM) and discovered that *minority collapse* occurs when classes are imbalanced across two equally-sized groups of classes; a threshold for minority collapse was later characterized by [35]. [52] showed that $\mathcal{NC}$ still occurs in such an LPM when the classifier is initialized as a fixed ETF. [70] introduced *simplex-encoded-label interpolation* (SELI) but noted that more severe imbalance worsens even this geometry. Recently, feature regularization has been employed to induce $\mathcal{NC}$ and improve generalization in class-imbalanced settings [58, 61, 62].

Multi-label classification (¬Rq3) In some problems, one might encounter mixed or multi-label samples, be they natural or due to noise or augmentation [71, 72]. $\mathcal{NC}$ was also recently studied for such data by [73], who observed that multi-label class means arrive at the average of their labels' respective means. They also devise an augmented CE loss function to accommodate such samples.

Likewise, most of our ambiguous samples could be considered multi- or soft-label: identical (or very similar) context samples with different hard token labels (¬Rq3). Under popular CLM pre-training (teacher-forcing with CE loss), this effectively precludes the prospect of achieving zero classification error and potentially introduces irreducible noise.

A recent study showed that memorization of noisy labels likely leads to degradation of $\mathcal{NC}$ [60].

Early stages of training (¬Rq4) [53] studied $\mathcal{NC}$ in small ResNet [74] models that had not yet converged (similar to most LLMs). They show that within-class variability drops and "plateaus" ($\mathcal{NC}_1$) earlier than other $\mathcal{NC}$ metrics, a result that we also observe (§4.1, Figures 6, 7).

Natural language processing (NLP) An earlier study reported that the ratio of within-class to between-class covariances of word embeddings increases with model depth [75, 76], seemingly at odds with $\mathcal{NC}_1$. It can, however, be distinguished from literature studying layer-wise $\mathcal{NC}$ in that it does not centre the mean representation vectors (i.e. subtract the global mean).

[19] fine-tuned BERT [7] using their *hyperspherical uniformity gap* (HUG) objective on binary classification tasks from the GLUE benchmark [77]. [78] conducted a tangentially-related investigation of convolutional neural networks for few-class semi-supervised clustering in which they identify $\mathcal{NC}$. Our work is distinct from these in several ways: a) our class space far exceeds our embedding dimension ($C \gg d + 1$) because we classify on the full token vocabulary; b) we analyze embeddings at a token-level granularity rather than at the sequence level; and c) our NTP task is causal (context-dependent) as opposed to the per-sample independence of their few-category classification.

A more related work studied *feature collapse* in individual word representations [79], but the authors note that their analysis is limited to shallow NLP modelling on more rigid and tabular text data.

3 Methodology

Below we describe the training setup for our CLMs (§3.1), procedures[3] for collecting top-layer embeddings (§3.2, 3.7), and measurements of $\mathcal{NC}$ and generalization (§3.4, 3.5, 3.6, 3.7, 3.8).

3.1 Dataset and Language Models

TinyStories [2] is a synthetic[4] dataset generated by GPT-3.5 and GPT-4 using around 1500 English words a child might use. NTP is performed by sampling from the token vocabulary $\mathbb{V} = [\![1, 29233]\!]$,

---

[3]Leveraging our generic $\mathcal{NC}$ package: https://github.com/rhubarbwu/neural-collapse.

[4]TinyStories was developed and evaluated as a faithful emulation of large language modelling, so we chose it for experimentation to train hundreds of CLMs and analyze embeddings at a low cost. See Appendix A.

which for our purposes can therefore be framed as classification among $C = 29{,}233$ classes.[5] Following the GPT-style preprocessing regime [8], raw sequences are packed into S chunks of size T, providing $N = S(T-1)$ token samples for training.[6] Details are listed in Appendix A.

We use 30 CLM architectures based on GPT Neo [80], configured similarly to [2]. They vary in width (embedding dimension) $d \in \{64, 128, 256, 512, 768, 1024\}$ and depth (number of self-attention layers) $L \in \{1, 2, 4, 8, 12\}$. Our models were trained by teacher-forcing[7] using CE loss. For each architecture, we trained multiple models for 1, 3, and 10 epochs ablated over weight decay factors $\beta = 0.0005$ [51] and $\beta = 0.1$ [81]. Further details are listed in Appendices B, C.

3.2 Context Embeddings

Base CLMs perform next-token prediction: given a sequence of tokens $\boldsymbol{x}_{1:t} \in \mathbb{V}^t$, a top-layer context embedding $\boldsymbol{h}(\boldsymbol{x}_{1:t}) \in \mathbb{R}^d$ is used to predict the next token $x'_{t+1} \in \mathbb{V}$ where $1 \leq t \leq T$. A classifier for class c with weights $\boldsymbol{w}_c$ and bias[8] b_c would make maximum a prior (MAP) predictions as

$$x'_{t+1} := \underset{c \in \mathbb{V}}{\operatorname{argmax}} \langle \boldsymbol{w}_c, \boldsymbol{h}(\boldsymbol{x}_{1:t}) \rangle + b_c. \tag{1}$$

Class embedding means   For each token class c, we are interested in accumulating the mean embedding $\boldsymbol{\mu}_c \in \mathbb{R}^d$ across sequences s and their contexts $\boldsymbol{x}_{1:t}^{(s)}$, where the next token $x_{t+1}^{(s)}$ is ground-truth ($t < T$) and equal to c:

$$\boldsymbol{\mu}_c := \frac{1}{N_c} \sum_{s=1}^{S} \sum_{t=1}^{T-1} \boldsymbol{h}\left(\boldsymbol{x}_{1:t}^{(s)}\right) \mathbb{I}\left(x_{t+1}^{(s)} = c\right), \tag{2}$$

where N_c is the number of samples of class c and $\mathbb{I}$ is the (binary) indicator function. We also utilize their unweighted[9] average $\bar{\boldsymbol{\mu}} := \mathbb{E}_c[\boldsymbol{\mu}_c]$, and subsequently the globally-centred means $\hat{\boldsymbol{\mu}}_c = \frac{\boldsymbol{\mu}_c - \bar{\boldsymbol{\mu}}}{\|\boldsymbol{\mu}_c - \bar{\boldsymbol{\mu}}\|_2}$.

Class embedding variances   In a second pass, we accumulate the sample variance norms:[10]

$$\sigma_c^2 := \frac{1}{N_c} \sum_{s=1}^{S} \sum_{t=1}^{T-1} \left\| \boldsymbol{h}\left(\boldsymbol{x}_{1:t}^{(s)}\right) - \boldsymbol{\mu}_c \right\|_2^2 \mathbb{I}\left(x_{t+1}^{(s)} = c\right). \tag{3}$$

3.3 Homogeneity and Variability

For some collapse measures (such as $(\mathcal{G})\mathcal{NC}_2$ and $\mathcal{NC}_3$), we are primarily interested in the *variation* rather than the average of pairwise relationships. To that end, we also include in our analysis the *coefficient of variation* (CoV) of several measurements, which is the ratio of their standard deviations to their means: $\sigma(\cdot)/\mu(\cdot)$. We can interpret this as a normalized measure of variability.

3.4 Signal-to-Noise Ratio — $\mathcal{NC}1$

The ability to disambiguate between classes depends on the ratio of within-class to between-class variabilities. Building upon foundational works [85, 86], $\mathcal{NC}$ originally measured variability through an inverse *signal-to-no ratio* (SNR), whose minimization constitutes *within-class variability collapse* ($\mathcal{NC}_1$). We instead employ a *class-distance normalized variance* (CDNV) similar to [3]:

$$\hat{\sigma}_{c,c'} := \frac{1}{\|\boldsymbol{\mu}_c - \boldsymbol{\mu}_{c'}\|_2^k} \cdot \frac{\sigma_c^2 + \sigma_{c'}^2}{2\|\boldsymbol{\mu}_c - \boldsymbol{\mu}_{c'}\|_2^2}, \quad \forall c \neq c'. \tag{4}$$

Our metric differs in that we divide by an additional power $\|\boldsymbol{\mu}_c - \boldsymbol{\mu}_{c'}\|_2^k$ of the mean distance norm. This further downweights the CDNV within well-separated class pairs in favour of emphasizing more

---

[5]Although the GPT-Neo [80] tokenizer has over 50K tokens, only a subset vocabulary appears in TinyStories.

[6]We cannot use the first ground-truth nor the last predicted token in any chunk.

[7]Parallelized training using sequences' ground-truth labels for context as opposed to predicted tokens.

[8]Similar to many causal LLMs [80–84], our classifiers do not include additive biases, so $b_c = 0$.

[9]Different from most literature where classes were balanced and weighting is already equal.

[10]This sample variance is computed across all unnormalized dimensional entries.

mutually noisy pairs. We found this augmented CDNV with $k = 2$ to be especially useful in our setting of many imbalanced and variably confusable token classes.

These pairwise measures constitute the off-diagonal[11] entries of a symmetric matrix in $\mathbb{R}^{C \times C}$, whose average we use as an inverse SNR. Within-class variability collapse is then re-characterized by the minimization of this quantity: $\hat{\sigma}_{c,c'} \to 0, \forall c \neq c'$. This alternative convergence is empirically faithful to $\mathcal{NC}_1$ but more robust and numerically stable [3].

3.5 Geometric Structures — $(\mathcal{G})\mathcal{NC}2$

The separability of our representations also depends on the geometric structures found in our embeddings. [1] characterize $\mathcal{NC}_2$ as convergence to a *simplex equiangular tight frame* (ETF) [87, 88].

Equinormness Such a near-orthonormal configuration firstly implies that class means are equinorm:
$$\log \|\boldsymbol{\mu}_c - \bar{\boldsymbol{\mu}}\|_2 - \log \|\boldsymbol{\mu}_{c'} - \bar{\boldsymbol{\mu}}\|_2 \to 0, \quad \forall c \neq c'. \tag{5}$$
We measure CoV in the *logarithms* of class mean norms to assess the degree of "equinormness".

Equiangularity $\mathcal{NC}_2$ also entails that class means are equiangular about their centre $\bar{\mu}$: pairwise distances and angles between their class means should be maximized and similar. Following [1], we measure *interference* (sometimes known as similarity or *coherence* [89, 90]). Its minimization,
$$\langle \hat{\boldsymbol{\mu}}_c, \hat{\boldsymbol{\mu}}_{c'} \rangle \to \frac{-1}{C-1}, \quad \forall c \neq c', \tag{6}$$
together with equinormness (Equation 5) constitute convergence to a simplex ETF. Although this geometry is not ultimately attainable since there are too many classes ($C > d + 1$), it can still be meaningful to measure a model's tendency towards one. As with CDNV noise (Equation 4), pairwise interferences form off-diagonal[12] entries in a symmetric matrix in $\mathbb{R}^{C \times C}$. The minimization of CoV in interferences therefore expresses the degree of "equiangularity".

Hyperspherical uniformity A relaxation from the ETF is *hyperspherical uniformity* ($\mathcal{GNC}_2$), with equinorm (Eq. 5) means $\boldsymbol{\mu}_c$ uniformly distributed on the d-dimensional hypersphere [18, 19]. We likewise gauge the angular uniformity with pairwise interactions through some distance kernel K:[13]
$$\sum_{c \neq c'} K(\hat{\boldsymbol{\mu}}_c, \hat{\boldsymbol{\mu}}_{c'}) \to \min_{\hat{\boldsymbol{\mu}}_1, \ldots, \hat{\boldsymbol{\mu}}_C} \sum_{c \neq c'} K(\hat{\boldsymbol{\mu}}_c, \hat{\boldsymbol{\mu}}_{c'}). \tag{7}$$

3.6 Alignment Between Classifiers and Class Embedding Means — $(\mathcal{U})\mathcal{NC}3$

The linear classifiers $\{\boldsymbol{w}_c\}_{c=1}^C$ lie in a dual vector space to that of the class means $\{\boldsymbol{\mu}_c\}_{c=1}^C$. While convergence to self-duality ($\mathcal{NC}_3$) was initially measured as distances [1] between class means and classifiers (Equation 11), we follow [5] to inspect class-wise cosine similarities:[14]
$$\left\langle \frac{\boldsymbol{w}_c}{\|\boldsymbol{w}_c\|_2}, \hat{\boldsymbol{\mu}}_c \right\rangle \to 1, \quad \forall c \in \mathbb{V}. \tag{8}$$

For intuition analogous to that for equinormness and equiangularity (§3.5), we also measure *uniform duality* ($\mathcal{UNC}_3$) as the minimization of the CoV of these similarities (Appendices N, O).

3.7 Agreement of the Classifiers — $\mathcal{NC}4$

Finally, $\mathcal{NC}_4$ is described as the simplification (or approximation) of the linear classifier's MAP prediction behaviour (Equation 1) to that of the implicit *nearest-class centre* (NCC) classifier:
$$\underset{c \in \mathbb{V}}{\arg\max} \langle \boldsymbol{w}_c, \boldsymbol{h} \rangle + b_c \to \underset{c \in \mathbb{V}}{\arg\min} \|\boldsymbol{h} - \boldsymbol{\mu}_c\|_2, \quad \forall \boldsymbol{h}. \tag{9}$$

---

[11]The main diagonal of CDNVs is undefined (due to the zero denominator) and ignored.

[12]The main diagonal of interferences is equal to $\mathbf{1}$ (maximal coherence or self-similarity).

[13]Following [19], we employ the logarithmic inverse distance kernel $K_{\log}(\boldsymbol{a}, \boldsymbol{b}) = \log \|\boldsymbol{a} - \boldsymbol{b}\|_2^{-1}$ for its ability to emphasize gaps between small distances while scaling down the effect of larger distances.

[14]Dot-product would be confounded by norms and therefore inappropriate for similarity up to rescaling.

We calculate[15] agreement as the proportion of validation samples on which the classifiers agree: [16]

$$\frac{1}{N_{\text{val}}} \sum_{s=1}^{S_{\text{val}}} \sum_{t=1}^{T_{\text{val}}-1} \mathbb{I}\left(x'_{t+1}{}^{(s)} = \underset{c\in\mathcal{V}}{\text{argmin}} \left\| \boldsymbol{h}\left(x_{1:t}^{(s)}\right) - \boldsymbol{\mu_c} \right\|_2 \right). \tag{10}$$

3.8   Probing a Relationship Between $\mathcal{NC}$ and Generalization

To isolate the effect of $\mathcal{NC}$ on generalization independent of model scaling and training (if it exists), we selected a two-layer 768-wide architecture of which to train twenty more instances with weight decay $\beta = 0.0005$, each with a different data shuffling seed. We then followed the remainder of the pipeline described above to collect and analyze embeddings with respect to $\mathcal{NC}$. Finally, we performed a permutation test with 10^4 trials to determine the statistical significance of any correlation between $\mathcal{NC}$ and generalization that remains when we hold constant all factors but shuffling seeds.

4   Experimental Results

In this section, we present the results from our empirical study on scaling and generalization:

($\mathcal{NC}_1$) Within-class variability is reduced across model scale (more so by width than depth) and training (up to 6 epochs), and is tightly correlated with validation performance.

($\mathcal{NC}_2$) Equinormness/equiangularity shows some improvement with scale, training, and performance. Hyperspherical uniformity ($\mathcal{GNC}_2$) also improves but more clearly and consistently.

($\mathcal{NC}_3$) Our models fail to achieve self-duality: class means do not align with classifiers. However, uniform duality ($\mathcal{UNC}3$) is correlated with model width, training, and performance.

($\mathcal{NC}_4$) Larger or more trained models exhibit closer agreement between their linear and implicit NCC classifiers. Agreement is also associated with validation performance.

4.1   Within-Class Variability Collapse — $\mathcal{NC}1$

Scaling our models dramatically reduces normalized variance, which is further aided by more training epochs and stronger weight decay (Appendix Figs. 6, 7). These noise reductions tightly associate with generalization (Fig. 1, left, "$\mathcal{NC}_1$"). The relationship is most apparent at scale.

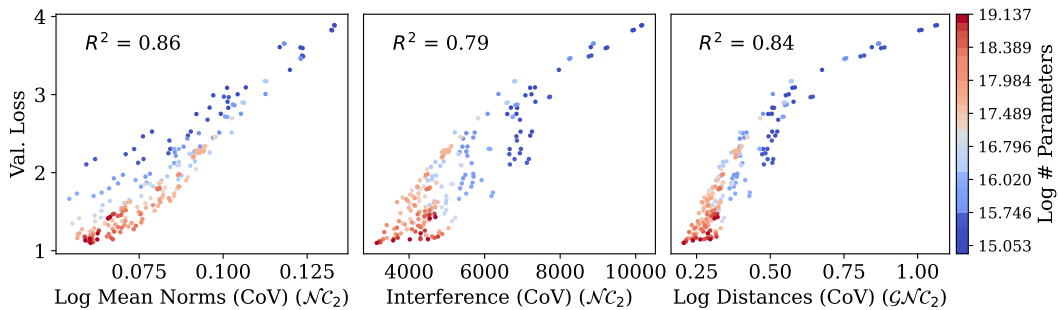### 4.2   Geometric Structures — $(\mathcal{G})\mathcal{NC}2$

Figure 2: Validation loss is correlated with all three measurements: **(left)** equinormness ($\mathcal{NC}2$) expressed as variation in logarithmic norms; **(centre)** equiangularity ($\mathcal{NC}2$) as variation in interference; **(right)** hyperspherical uniformity ($\mathcal{GNC}2$) as variation in logarithmic pairwise distances.

Equinormness   Logarithmic class mean norms grow with model width and training (Appendix Fig. 8), and subtly with depth (Appendix Fig. 9). Meanwhile, the variation of these (logarithmic) norms consistently decreases (improving equinormness) with scale (Appendix Figs. 10, 11). Both trends correlate with improved generalization (Appendix Fig. 20).

---

[15]In practice, we use an equivalent decomposition (Eq. 12).

[16]Note that agreement is not equivalent to accuracy.

Equiangularity   Scaling model dimensions reduces average interference (Appendix Figs. 12, 13) down to an empirical plateau of approximately 10^{-3}, which is more apparent in less-trained models. However, the variation of interference rises and persists when scaling (Appendix Figs. 14, 15), suggesting that interference becomes more concentrated between some pairs of classes. These results could be due to various factors, including but not limited to unfriendly conditions of language modelling (§1.2) or the impossibility of a perfect simplex ETF (§3.5).

Appendix Figure 21 shows only a rough performance correlation with average interference (i.e. coherence) and a more definite — albeit still noisy — one with the variation of interference (i.e. equiangularity). In other words, the limited trends we observed toward a simplex ETF suggest that the association of $\mathcal{NC}2$ with generalization begins to break down when $C > d + 1$.

Hyperspherical uniformity   Logarithmic distances drop more gradually and consistently with scale (Appendix Figs. 16, 17), implying this quantity is more robust or may not face the same geometric barriers seen in conventional interference (Appendix Figs. 14, 15). Variation of these logarithmic distances is also consistently reduced with scale (Appendix Fig. 18, 19).

And finally, generalization has much stronger correlations with logarithmic distances than it has with regular interference (Fig. 2), validating the applicability of $\mathcal{GNC}$ [19] when $C > d + 1$.

4.3   Classifier (Mis)alignment and Duality — $(\mathcal{U})\mathcal{NC}3$

Model width (d) is correlated faintly with the average similarity between class means and their respective classifiers (Appendix Fig. 23), but strongly with variation in similarity (Appendix Fig. 25). The relationships to generalization follow the same pattern (Fig. 3), suggesting that uniform duality ($\mathcal{UNC}_3$) *might* serve as a better $\mathcal{NC}$ property than self-duality ($\mathcal{NC}_3$) overall; we discuss this in §5.1.

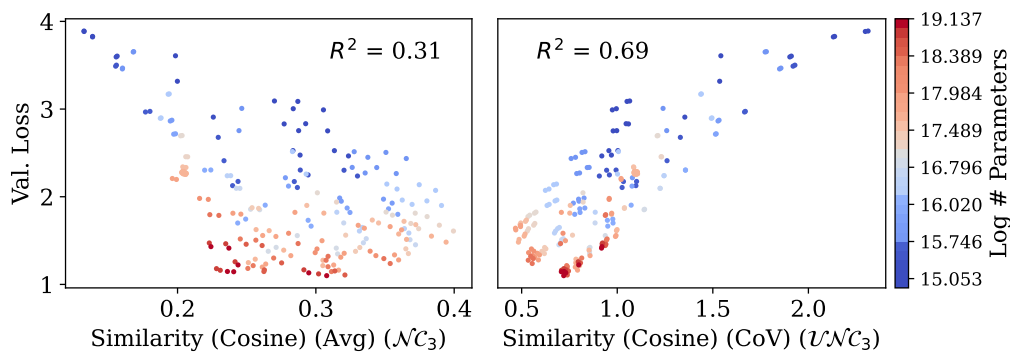
Figure 3: Validation loss shows a negligible relationship with self-duality ($\mathcal{NC}_3$, left) and some correlation with uniform duality ($\mathcal{UNC}_3$, right). In other words, $\mathcal{UNC}_3$ develops with scale and correlates with generalization much better than $\mathcal{NC}_3$.

4.4   Classifier Agreement — $\mathcal{NC}4$

The linear and NCC classifiers agree on far more examples than random chance, and model scaling encourages this agreement (Appendix Figs. 29, 30). Increasing width for certain depths happens to plateau or even regress the agreement rate, but this limitation is overcome with further training. And finally, agreement is a strong indicator of generalization (Fig. 1, right, $\mathcal{NC}_4$).

5   Analysis

We find that $\mathcal{NC}$ is generally promoted by model size and training and correlated with generalization (validation performance). We also discern some of this correlation independent of scale (§5.1).

Table 1: Permutation test of $\mathcal{NC}$ measurements with respect to validation loss. Twenty-one (21) identical two-layer 768-wide models were trained with different data shuffling seeds and permuted with 10^4 trials. The p-values below 0.05 (bolded) show those properties to be statistically significant.

Property	Measurement	R^2 Correlation ($\uparrow$)	p-value ($\downarrow$)
$\mathcal{NC}_1$	Within-Class Variability Collapse	0.192011	**0.0485**
$\mathcal{NC}_2$	Equinormness	0.026174	0.4870
$\mathcal{NC}_2$	Equiangularity	0.218574	**0.0317**
$\mathcal{GNC}_2$	Hyperspherical Uniformity	0.487935	**0.0002**
$\mathcal{NC}_3$	Self-Duality	0.322210	**0.0063**
$\mathcal{UNC}_3$	Uniform Duality	0.000036	0.9784
$\mathcal{NC}_4$	Classifier Agreement	0.490278	**0.0001**

5.1 Neural Collapse's Relationship with Generalization

Table 1 presents the correlation scores of $\mathcal{NC}$ metrics with generalization alongside their associated p-values from the permutation tests described in §3.8. The majority of the correlations are statistically significant ($p < 0.05$) independent of scale, affirming that $\mathcal{NC}$ is correlated with generalization.

5.2 Duality of Duality

The dichotomy of self-duality ($\mathcal{NC}_3$) and uniform duality ($\mathcal{UNC}_3$) is rather conspicuous. Our main experiments find that $\mathcal{NC}_3$ does not consistently develop with scale while $\mathcal{UNC}_3$ does (Fig. 3). However, within a fixed scale, the opposite is true, implying that $\mathcal{UNC}_3$ may be confounded by model capacity while $\mathcal{NC}_3$ is a subtle and fine-grained indicator of generalization.

5.3 The Effect of Weight Regularization

Our models trained with either weight decay factor exhibited very similar patterns in the emergence of $\mathcal{NC}$ (or lack thereof), but the more aggressive factor $\beta = 0.1$ resulted in stronger development of $\mathcal{NC}$ properties than with $\beta = 0.0005$ (Appendices E, F, G, H, I, J, K, M, N, P). These findings empirically affirm $\beta = 0.1$ weight decay as sensible for CLM pre-training [81], and concur with [51] on the pivotal role that appropriate regularization plays in the emergence of $\mathcal{NC}$.

6 Limitations

Neural collapse While to the best of our knowledge, no previous work has studied realistic stochastic token prediction, it is possible that the quantities that we measure are not perfectly suited for $\mathcal{NC}$ in language modelling. As we described in §1.1, the $\mathcal{NC}$ framework does not translate neatly to the language modelling space due to many adverse conditions, so full convergence to $\mathcal{NC}$ in the TPT was highly improbable. This paper leaves much room for future work to better adapt $\mathcal{NC}$ for next-token prediction, which we discuss further in Section 7.

Language modelling Our work focused on autoregressive pre-training in its most basic form. We did not conduct experiments into encoder, multi-modal, or instruction-tuned models. Post-training techniques such as supervised fine-tuning, reinforcement learning with human feedback [91] or direct preference optimization [92] are also out-of-scope. This paper uses validation CE loss as the sole indicator of performance, leaving out any downstream task evaluations.

Confounder of model scale The models that we used in our permutation test (§5.1, Table 1) are only of a single small architecture trained for one epoch with relatively weak weight regularization ($\beta = 0.0005$). Therefore, our experimental results on scale-independent links between $\mathcal{NC}$ and generalization may not necessarily translate to larger models. Further investigation on (foundation) LLMs orders of magnitude larger than our CLMs trained with modern NLP methods would provide more robust insight into any direct correlations.

7 Discussion

Layer/depth-wise neural collapse   Past works have established that properties resembling $\mathcal{NC}$ evolve as a function of model depth [4, 36, 53, 56, 59, 60, 93–104]. Layer-wise $\mathcal{NC}$ — sometimes dubbed *deep neural collapse* ($\mathcal{DNC}$) [53, 59] — and related phenomena at intermediate layers remain an interesting subtopic. We leave their study and induction in CLMs (like [105]) as future work.

Learning to collapse   Given the evidence for the development of $\mathcal{NC}$ and associated generalization under various loss functions [19, 49, 60, 100, 106–108] in other domains, NLP researchers may still benefit from analyzing, adapting or training towards $\mathcal{NC}$. As alluded to earlier, the simplex ETF and even the CE loss may not be truly optimal for this problem setting, so we anticipate future works to both construct more amenable geometries with better-suited objectives and then capitalize on their benefits downstream. As discussed in §2, there is an abundance of literature in $\mathcal{NC}$, some of which could potentially adapt $\mathcal{NC}$ to be useful for NLP; we hope to inspire more such investigations.

Interpretability   At a high level, the number and density of clusters for a token can reflect its learned meanings and uses. This would be particularly useful as LLMs adapt to ever-evolving language and further expansion into non-English domains. Our formulae (Section 3) and results (Section 4) expose the pairwise token class interactions in noise, interference, classifier duality, and classifier agreement in the top-level features. Similarly to works in other domains [21, 55, 72, 109], these $\mathcal{NC}$ metrics can serve as a form of low-level interpretability to aid understanding certain behaviours of (L)LMs. Between tokens, one can often discern how related or interchangeable words are based on their pair-wise interactions, or how antithetical or unrelated they are based on orthogonality. For example, we present a rudimentary analysis of homonyms and English first names in Appendix Q.

Fairness   Foundation models are ubiquitous for their comprehensive capabilities and adaptability. As previous work discussed class imbalance [35, 58, 61, 62], our work may extend these strategies to measure and perhaps promote fairness in foundation LLMs, some of which are designed to be multilingual or multicultural. For example, [110] contemporarily explores the use of $\mathcal{UNC}3$ to mitigate biases in BERT-based [7] models.

While $\mathcal{NC}$ itself would not lead to unfairness, its potential interpretability may, in theory, enable an (adversarial) agent to measure and optimize for (un)fairness as they manipulate an LLM.

LLM Evaluations   Researchers in NLP and multimodal settings are ultimately interested in measuring model performance on practical tasks; notable benchmarks include GLUE [77], MMLU [111], and BIG-bench [112]. However, several contemporaries [113–115] have demonstrated that models' downstream capabilities are roughly correlated with their abilities to effectively compress their pre-training data. Based on their findings, our application of the $\mathcal{NC}$ framework to the pre-training stage of CLMs against validation CE loss should be an appropriate first step in this intersection. Looking forward, we anticipate exciting analysis for language modelling tasks or benchmarks, especially on creativity and retrieval for natural language understanding and generation.

Conversely, some form of $\mathcal{NC}$ could be an alternative evaluation. Although it would be prohibitively expensive to measure $\mathcal{NC}$ on the vast and sometimes obscure pre-training data of most frontier production LLMs, doing so on a small set of in-distribution data (i.e. test set) would be realistic.

8 Conclusion

In this paper, we apply the *neural collapse* ($\mathcal{NC}$) framework to the next-token prediction problem, where models are undertrained and next-tokens are variably drawn from numerous and imbalanced token classes. We leverage canonical and more recent metrics to demonstrate that $\mathcal{NC}$ emerges as we scale the size and training of hundreds of causal language models. Our results show a correlation between $\mathcal{NC}$ and generalization, a relationship that persists even when the model scale is fixed.

In the short term, our work presents rudimentary techniques to analyze and interpret token-level properties of (L)LMs. We anticipate future work to suitably adapt $\mathcal{NC}$ (and related frameworks) to the still-fresh frontier of autoregressive language modelling. Researchers could then effectively capitalize on previous learnings from $\mathcal{NC}$ to better understand and improve the pre/post-training processes of increasingly complex and large language (multimodal) models.

Acknowledgements

References

[1] Vardan Papyan, X. Y. Han, and David L. Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40):24652–24663, sep 2020. doi: 10.1073/pnas.2015509117. URL `https://www.pnas.org/doi/full/10.1073/pnas.2015509117`.

[2] Ronen Eldan and Yuanzhi Li. Tinystories: How small can language models be and still speak coherent english?, 2023. URL `https://arxiv.org/abs/2305.07759`.

[3] Tomer Galanti, András György, and Marcus Hutter. On the role of neural collapse in transfer learning, 2022. URL `https://arxiv.org/abs/2112.15121v2`.

[4] Tomer Galanti, Liane Galanti, and Ido Ben-Shaul. On the implicit bias towards minimal depth of deep neural networks, 2022. URL `https://arxiv.org/abs/2202.09028v9`.

[5] Vignesh Kothapalli. Neural collapse: A review on modelling principles and generalization. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL `https://openreview.net/forum?id=QTXocpAP9p`.

[6] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445922. URL `https://doi.org/10.1145/3442188.3445922`.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. URL `https://arxiv.org/abs/1810.04805`.

[8] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

[9] Jiachen Jiang, Jinxin Zhou, Peng Wang, Qing Qu, Dustin Mixon, Chong You, and Zhihui Zhu. Generalized neural collapse for a large number of classes, 2023. URL `https://arxiv.org/abs/2310.05351`.

[10] Zhilin Yang, Zihang Dai, Ruslan Salakhutdinov, and William W. Cohen. Breaking the softmax bottleneck: A high-rank rnn language model, 2018. URL `https://arxiv.org/abs/1711.03953`.

[11] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948. doi: 10.1002/j.1538-7305.1948.tb01338.x.

[12] P Sargant Florence. Human behaviour and the principle of least effort. *The Economic Journal*, 60(240):808–810, 1950.

[13] Daniel Jurafsky and James H Martin. Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition, 2009.

[14] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models, 2020. URL `https://arxiv.org/abs/2001.08361v1`.

[15] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models, 2022. URL `https://arxiv.org/abs/2203.15556v1`.

[16] Niklas Muennighoff, Alexander M. Rush, Boaz Barak, Teven Le Scao, Aleksandra Piktus, Nouamane Tazi, Sampo Pyysalo, Thomas Wolf, and Colin Raffel. Scaling data-constrained language models, 2023. URL `https://arxiv.org/abs/2305.16264v4`.

[17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[18] Jianfeng Lu and Stefan Steinerberger. Neural collapse with cross-entropy loss, 2021. URL `https://arxiv.org/abs/2012.08465v2`.

[19] Weiyang Liu, Longhui Yu, Adrian Weller, and Bernhard Schölkopf. Generalizing and decoupling neural collapse via hyperspherical uniformity gap, 2023. URL `https://arxiv.org/abs/2303.06484v2`.

[20] Zexi Li, Xinyi Shang, Rui He, Tao Lin, and Chao Wu. No fear of classifier biases: Neural collapse inspired federated learning with synthetic and fixed classifier. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5319–5329, 2023.

[21] Vignesh Kothapalli, Tom Tirer, and Joan Bruna. A neural collapse perspective on feature evolution in graph neural networks. *Advances in Neural Information Processing Systems*, 36, 2024.

[22] Yibo Yang, Haobo Yuan, Xiangtai Li, Jianlong Wu, Lefei Zhang, Zhouchen Lin, Philip Torr, Dacheng Tao, and Bernard Ghanem. Neural collapse terminus: A unified solution for class incremental learning and its variants, 2023. URL `https://arxiv.org/abs/2308.01746v1`.

[23] Qinhao Zhou, Xiang Xiang, and Jing Ma. Hierarchical task-incremental learning with feature-space initialization inspired by neural collapse. *Neural Processing Letters*, 55(8):10811–10827, 2023.

[24] Hang Ran, Weijun Li, Lusi Li, Songsong Tian, Xin Ning, and Prayag Tiwari. Learning optimal inter-class margin adaptively for few-shot class-incremental learning via neural collapse-based meta-learning. *Information Processing & Management*, 61(3):103664, 2024. ISSN 0306-4573. doi: https://doi.org/10.1016/j.ipm.2024.103664. URL `https://www.sciencedirect.com/science/article/pii/S0306457324000244`.

[25] Saaketh Medepalli and Naren Doraiswamy. On the role of neural collapse in meta learning models for few-shot learning, 2023. URL `https://arxiv.org/abs/2310.00451v2`.

[26] Jarrod Haas, William Yolland, and Bernhard Rabus. Linking neural collapse and l2 normalization with improved out-of-distribution detection in deep neural networks, 2023. URL `https://arxiv.org/abs/2209.08378v3`.

[27] Mouïn Ben Ammar, Nacim Belkhir, Sebastian Popescu, Antoine Manzanera, and Gianni Franchi. Neco: Neural collapse based out-of-distribution detection, 2024. URL `https://arxiv.org/abs/2310.06823`.

[28] Jiawei Zhang, Yufan Chen, Cheng Jin, Lei Zhu, and Yuantao Gu. Epa: Neural collapse inspired robust out-of-distribution detector, 2024. URL `https://arxiv.org/abs/2401.01710v1`.

[29] Donghao Li, Yang Cao, and Yuan Yao. Neuromixgdp: A neural collapse-inspired random mixup for private data release. In *Conference on Parsimony and Learning*, pages 480–514. PMLR, 2024.

[30] Chendi Wang, Yuqing Zhu, Weijie J. Su, and Yu-Xiang Wang. Neural collapse meets differential privacy: Curious behaviors of noisygd with near-perfect representation learning, 2024. URL https://arxiv.org/abs/2405.08920v2.

[31] Tolga Ergen, Arda Sahiner, Batu Ozturkler, John Pauly, Morteza Mardani, and Mert Pilanci. Demystifying batch normalization in relu networks: Equivalent convex optimization models and implicit regularization, 2022. URL https://arxiv.org/abs/2103.01499v3.

[32] Tolga Ergen and Mert Pilanci. Revealing the structure of deep neural networks via convex duality. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 3004–3014. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/ergen21b.html.

[33] Mariia Seleznova, Dana Weitzner, Raja Giryes, Gitta Kutyniok, and Hung-Hsu Chou. Neural (tangent kernel) collapse. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 16240–16270. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/3477ca0ce484aa2fa42c1361ab601c25-Paper-Conference.pdf.

[34] Matus Telgarsky. Feature selection with gradient descent on two-layer networks in low-rotation regimes, 2022. URL https://arxiv.org/abs/2208.02789v1.

[35] Wanli Hong and Shuyang Ling. Neural collapse for unconstrained feature model under cross-entropy loss with imbalanced data, 2023. URL https://arxiv.org/abs/2309.09725v2.

[36] Gerard Ben Arous, Reza Gheissari, Jiaoyang Huang, and Aukosh Jagannath. High-dimensional sgd aligns with emerging outlier eigenspaces, 2023. URL https://arxiv.org/abs/2310.03010v1.

[37] Like Hui, Mikhail Belkin, and Preetum Nakkiran. Limitations of neural collapse for understanding generalization in deep learning, 2022. URL https://arxiv.org/abs/2202.08384v1.

[38] Tomer Galanti, András György, and Marcus Hutter. Improved generalization bounds for transfer learning via neural collapse. In *First Workshop on Pre-training: Perspectives, Pitfalls, and Paths Forward at ICML 2022*, 2022. URL https://openreview.net/forum?id=VrK7pKwOhT_.

[39] Zijian Wang, Yadan Luo, Liang Zheng, Zi Huang, and Mahsa Baktashmotlagh. How far pre-trained models are from neural collapse on the target dataset informs their transferability. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5549–5558, 2023.

[40] Xiao Li, Sheng Liu, Jinxin Zhou, Xinyu Lu, Carlos Fernandez-Granda, Zhihui Zhu, and Qing Qu. Understanding and improving transfer learning of deep models via neural collapse. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=o8r84MzTQB.

[41] Didi Zhu, Zexi Li, Min Zhang, Junkun Yuan, Yunfeng Shao, Jiashuo Liu, Kun Kuang, Yinchuan Li, and Chao Wu. Understanding prompt tuning for v-l models through the lens of neural collapse, 2023. URL https://arxiv.org/abs/2306.15955v3.

[42] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical Report 0, University of Toronto, Toronto, Ontario, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

[43] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[44] Dustin G. Mixon, Hans Parshall, and Jianzong Pi. Neural collapse with unconstrained features, 2020. URL https://arxiv.org/abs/2011.11619v1.

[45] Tomaso Poggio and Qianli Liao. Explicit regularization and implicit bias in deep network classifiers trained with the square loss, 2020. URL https://arxiv.org/abs/2101.00072v1.

[46] Weinan E and Stephan Wojtowytsch. On the emergence of simplex symmetry in the final and penultimate layers of neural network classifiers, 2021. URL https://arxiv.org/abs/2012.05420v3.

[47] Cong Fang, Hangfeng He, Qi Long, and Weijie J Su. Exploring deep neural networks via layer-peeled model: Minority collapse in imbalanced training. *Proceedings of the National Academy of Sciences*, 118(43):e2103091118, 2021.

[48] Zhihui Zhu, Tianyu Ding, Jinxin Zhou, Xiao Li, Chong You, Jeremias Sulam, and Qing Qu. A geometric analysis of neural collapse with unconstrained features, 2021. URL https://arxiv.org/abs/2105.02375v1.

[49] X.Y. Han, Vardan Papyan, and David L. Donoho. Neural collapse under MSE loss: Proximity to and dynamics on the central path. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=w1UbdvWH_R3.

[50] Can Yaras, Peng Wang, Zhihui Zhu, Laura Balzano, and Qing Qu. Neural collapse with normalized features: A geometric analysis over the riemannian manifold. *Advances in neural information processing systems*, 35:11547–11560, 2022.

[51] Akshay Rangamani and Andrzej Banburski-Fahey. Neural collapse in deep homogeneous classifiers and the role of weight decay. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4243–4247, 2022. doi: 10.1109/ICASSP43922.2022.9746778.

[52] Yibo Yang, Shixiang Chen, Xiangtai Li, Liang Xie, Zhouchen Lin, and Dacheng Tao. Inducing neural collapse in imbalanced learning: Do we really need a learnable classifier at the end of deep neural network?, 2022. URL https://arxiv.org/abs/2203.09081v3.

[53] Tom Tirer and Joan Bruna. Extended unconstrained features model for exploring deep neural collapse. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 21478–21505. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/tirer22a.html.

[54] Peng Wang, Huikang Liu, Can Yaras, Laura Balzano, and Qing Qu. Linear convergence analysis of neural collapse with unconstrained features. In *OPT 2022: Optimization for Machine Learning (NeurIPS 2022 Workshop)*, 2022.

[55] Jinxin Zhou, Chong You, Xiao Li, Kangning Liu, Sheng Liu, Qing Qu, and Zhihui Zhu. Are all losses created equal: A neural collapse perspective, 2022. URL https://arxiv.org/abs/2210.02192v2.

[56] Akshay Rangamani, Marius Lindegaard, Tomer Galanti, and Tomaso A Poggio. Feature learning in deep classifiers through intermediate neural collapse. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28729–28745. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/rangamani23a.html.

[57] Tom Tirer, Haoxiang Huang, and Jonathan Niles-Weed. Perturbation analysis of neural collapse. In *International Conference on Machine Learning*, pages 34301–34329. PMLR, 2023.

[58] Hien Dang, Tho Tran, Stanley Osher, Hung Tran-The, Nhat Ho, and Tan Nguyen. Neural collapse in deep linear networks: From balanced to imbalanced data, 2023. URL https://arxiv.org/abs/2301.00437v5.

[59] Peter Súkeník, Marco Mondelli, and Christoph H Lampert. Deep neural collapse is provably optimal for the deep unconstrained features model. *Advances in Neural Information Processing Systems*, 36, 2024.

[60] Duc Anh Nguyen, Ron Levie, Julian Lienen, Eyke Hüllermeier, and Gitta Kutyniok. Memorization-dilation: Modeling neural collapse under noise. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=cJWxqmmDL2b`.

[61] Zhisheng Zhong, Jiequan Cui, Yibo Yang, Xiaoyang Wu, Xiaojuan Qi, Xiangyu Zhang, and Jiaya Jia. Understanding imbalanced semantic segmentation through neural collapse. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19550–19560, 2023.

[62] Xuantong Liu, Jianfeng Zhang, Tianyang Hu, He Cao, Yuan Yao, and Lujia Pan. Inducing neural collapse in deep long-tailed learning. In *International Conference on Artificial Intelligence and Statistics*, pages 11534–11544. PMLR, 2023.

[63] Peifeng Gao, Qianqian Xu, Peisong Wen, Huiyang Shao, Zhiyong Yang, and Qingming Huang. A study of neural collapse phenomenon: Grassmannian frame, symmetry and generalization, 2023. URL `https://arxiv.org/abs/2304.08914v2`.

[64] Mufan Bill Li, Mihai Nica, and Daniel M. Roy. The neural covariance sde: Shaped infinite depth-and-width networks at initialization, 2023. URL `https://arxiv.org/abs/2206.02768v3`.

[65] Zhanxuan Hu, Yichen Wang, Hailong Ning, Yonghang Tai, and Feiping Nie. Neural collapse inspired semi-supervised learning with fixed classifier. *Information Sciences*, 667:120469, 2024.

[66] Gao Peifeng, Qianqian Xu, Yibo Yang, Peisong Wen, Huiyang Shao, Zhiyong Yang, Bernard Ghanem, and Qingming Huang. Towards demystifying the generalization behaviors when neural collapse emerges, 2024. URL `https://openreview.net/forum?id=XVv4S6LnMk`.

[67] Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposition. *Transformer Circuits Thread*, 2022. URL `https://transformer-circuits.pub/2022/toy_model/index.html`.

[68] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

[69] A. Dubey et al. (101 additional authors). The llama 3 herd of models. 2024. URL `https://arxiv.org/abs/2407.21783`. All authors were affiliated with Meta.

[70] Christos Thrampoulidis, Ganesh Ramachandra Kini, Vala Vakilian, and Tina Behnia. Imbalance trouble: Revisiting neural-collapse geometry. *Advances in Neural Information Processing Systems*, 35:27225–27238, 2022.

[71] Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. Learning with noisy labels. *Advances in neural information processing systems*, 26, 2013.

[72] Quinn Fisher, Haoming Meng, and Vardan Papyan. Pushing boundaries: Mixup's influence on neural collapse, 2024. URL `https://arxiv.org/abs/2402.06171v1`.

[73] Pengyu Li, Xiao Li, Yutong Wang, and Qing Qu. Neural collapse in multi-label learning with pick-all-label loss, 2024. URL `https://arxiv.org/abs/2310.15903v4`.

[74] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[75] David Mimno and Laure Thompson. The strange geometry of skip-gram with negative sampling. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel, editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2873–2878, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1308. URL `https://aclanthology.org/D17-1308`.

[76] Kawin Ethayarajh. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings, 2019. URL https://arxiv.org/abs/1909.00512.

[77] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding, 2019. URL https://arxiv.org/abs/1804.07461v3.

[78] Jia Hui Feng, Edmund M-K Lai, and Weihua Li. A study of neural collapse for text classification. In *International Conference on Deep Learning Theory and Applications*, pages 126–142. Springer, 2023.

[79] Thomas Laurent, James H. von Brecht, and Xavier Bresson. Feature collapse, 2023. URL https://arxiv.org/abs/2305.16162v1.

[80] Sid Black, Leo Gao, Phil Wang, Connor Leahy, and Stella Biderman. GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow, March 2021. URL https://doi.org/10.5281/zenodo.5297715. If you use this software, please cite it using these metadata.

[81] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[82] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. Opt: Open pre-trained transformer language models, 2022. URL https://arxiv.org/abs/2205.01068v4.

[83] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023. URL https://arxiv.org/abs/2302.13971v1.

[84] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. URL https://arxiv.org/abs/2310.06825v1.

[85] Ronald A Fisher. The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2):179–188, 1936.

[86] C Radhakrishna Rao. The utilization of multiple measurements in problems of biological classification. *Journal of the Royal Statistical Society. Series B (Methodological)*, 10(2):159–203, 1948.

[87] Thomas Strohmer and Robert W Heath Jr. Grassmannian frames with applications to coding and communication. *Applied and computational harmonic analysis*, 14(3):257–275, 2003.

[88] Shayne FD Waldron. *An introduction to finite tight frames*. Springer, 2018.

[89] David L Donoho, Michael Elad, and Vladimir N Temlyakov. Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Transactions on information theory*, 52(1):6–18, 2005.

[90] Joel A Tropp. Just relax: Convex programming methods for identifying sparse signals in noise. *IEEE transactions on information theory*, 52(3):1030–1051, 2006.

[91] Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences, 2020. URL https://arxiv.org/abs/1909.08593.

[92] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 53728–53741. Curran Associates, Inc., 2023. URL `https://proceedings.neurips.cc/paper_files/paper/2023/file/a85b405ed65c6477a4fe8302b5e06ce7-Paper-Conference.pdf`.

[93] Vardan Papyan. Traces of class/cross-class structure pervade deep learning spectra, 2020. URL `https://arxiv.org/abs/2008.11865v1`.

[94] Christopher R. Hoyt and Art B. Owen. Probing neural networks with t-sne, class-specific projections and a guided tour, 2021. URL `https://arxiv.org/abs/2107.12547v1`.

[95] John Zarka, Florentin Guth, and Stéphane Mallat. Separation and concentration in deep networks, 2021. URL `https://arxiv.org/abs/2012.10424v2`.

[96] Ido Ben-Shaul and Shai Dekel. Nearest class-center simplification through intermediate layers. In Alexander Cloninger, Timothy Doster, Tegan Emerson, Manohar Kaul, Ira Ktena, Henry Kvinge, Nina Miolane, Bastian Rieck, Sarah Tymochko, and Guy Wolf, editors, *Proceedings of Topological, Algebraic, and Geometric Learning Workshops 2022*, volume 196 of *Proceedings of Machine Learning Research*, pages 37–47. PMLR, 25 Feb–22 Jul 2022. URL `https://proceedings.mlr.press/v196/ben-shaul22a.html`.

[97] Hangfeng He and Weijie J. Su. A law of data separation in deep learning, 2022. URL `https://arxiv.org/abs/2210.17020v2`.

[98] Liam Parker, Emre Onal, Anton Stengel, and Jake Intrater. Neural collapse in the intermediate hidden layers of classification neural networks, 2023. URL `https://arxiv.org/abs/2308.02760v1`.

[99] Wojciech Masarczyk, Mateusz Ostaszewski, Ehsan Imani, Razvan Pascanu, Piotr Mił oś, and Tomasz Trzcinski. The tunnel effect: Building data representations in deep neural networks. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 76772–76805. Curran Associates, Inc., 2023. URL `https://proceedings.neurips.cc/paper_files/paper/2023/file/f249db9ab5975586f36df46f8958c008-Paper-Conference.pdf`.

[100] Daniel Beaglehole, Peter Súkeník, Marco Mondelli, and Mikhail Belkin. Average gradient outer product as a mechanism for deep neural collapse, 2024. URL `https://arxiv.org/abs/2402.13728v2`.

[101] Peng Wang, Xiao Li, Can Yaras, Zhihui Zhu, Laura Balzano, Wei Hu, and Qing Qu. Understanding deep representation learning via layerwise feature compression and discrimination, 2024. URL `https://arxiv.org/abs/2311.02960v2`.

[102] Sicong Wang, Kuo Gai, and Shihua Zhang. Progressive feedforward collapse of resnet training, 2024. URL `https://arxiv.org/abs/2405.00985v1`.

[103] Connall Garrod and Jonathan P. Keating. Unifying low dimensional observations in deep learning through the deep linear unconstrained feature model, 2024. URL `https://arxiv.org/abs/2404.06106v1`.

[104] Emanuele Zangrando, Piero Deidda, Simone Brugiapaglia, Nicola Guglielmi, and Francesco Tudisco. Neural rank collapse: Weight decay and small within-class variability yield low-rank bias, 2024. URL `https://arxiv.org/abs/2402.03991v1`.

[105] Jiachen Jiang, Jinxin Zhou, and Zhihui Zhu. On layer-wise representation similarity: Application for multi-exit models with a single classifier, 2024. URL `https://arxiv.org/abs/2406.14479`.

[106] Mengjia Xu, Akshay Rangamani, Qianli Liao, Tomer Galanti, and Tomaso Poggio. Dynamics in deep classifiers trained with the square loss: Normalization, low rank, neural collapse, and generalization bounds. *Research*, 6:0024, 2023.

[107] Tong Liang and Jim Davis. Inducing neural collapse to a fixed hierarchy-aware frame for reducing mistake severity. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1443–1452, 2023.

[108] Guglielmo Bonifazi, Iason Chalas, Gian Hess, and Jakub Łucki. Can we understand plasticity through neural collapse?, 2024. URL `https://arxiv.org/abs/2404.02719v1`.

[109] Li Guo, Keith Ross, Zifan Zhao, George Andriopoulos, Shuyang Ling, Yufeng Xu, and Zixuan Dong. Cross entropy versus label smoothing: A neural collapse perspective, 2024. URL `https://arxiv.org/abs/2402.03979v2`.

[110] Jingxuan Xu, Wuyang Chen, Linyi Li, Yao Zhao, and Yunchao Wei. Collapsed language models promote fairness, 2024. URL `https://arxiv.org/abs/2410.04472`.

[111] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

[112] BIG bench authors. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL `https://openreview.net/forum?id=uyTL5Bvosj`.

[113] Zhengxiao Du, Aohan Zeng, Yuxiao Dong, and Jie Tang. Understanding emergent abilities of language models from the loss perspective, 2024. URL `https://arxiv.org/abs/2403.15796`.

[114] Yuzhen Huang, Jinghan Zhang, Zifei Shan, and Junxian He. Compression represents intelligence linearly. In *First Conference on Language Modeling*, 2024. URL `https://openreview.net/forum?id=SHMj84U5SH`.

[115] Mingjia Yin, Chuhan Wu, Yufei Wang, Hao Wang, Wei Guo, Yasheng Wang, Yong Liu, Ruiming Tang, Defu Lian, and Enhong Chen. Entropy law: The story behind data compression and llm performance, 2024.

[116] Stephen Merity, Ilya Sutskever, Simon Kornblith, and Nikhil Goyal. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*, 2016.

[117] Yao Zhu, Zhen Yu, Chao Zhang, Yijia Wu, et al. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. *arXiv preprint arXiv:1506.05829*, 2015.

[118] Common Crawl. Common crawl, 2023. URL `https://commoncrawl.org`. Accessed: 2023-10-30.

[119] Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. The pile: An 800gb dataset of diverse text for language modeling, 2020. URL `https://arxiv.org/abs/2101.00027`.

[120] Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, and Ross Anderson. The curse of recursion: Training on generated data makes models forget, 2024. URL `https://arxiv.org/abs/2305.17493`.

[121] Matthias Gerstgrasser, Rylan Schaeffer, Apratim Dey, Rafael Rafailov, Henry Sleight, John Hughes, Tomasz Korbak, Rajashree Agrawal, Dhruv Pai, Andrey Gromov, Daniel A. Roberts, Diyi Yang, David L. Donoho, and Sanmi Koyejo. Is model collapse inevitable? breaking the curse of recursion by accumulating real and synthetic data, 2024. URL `https://arxiv.org/abs/2404.01413`.

[122] Neil Burgess, Jelena Milanovic, Nigel Stephens, Konstantinos Monachopoulos, and David Mansell. Bfloat16 processing for neural networks. In *2019 IEEE 26th Symposium on Computer Arithmetic (ARITH)*, pages 88–91, 2019. doi: 10.1109/ARITH.2019.00022.

A Dataset

TinyStories is a dataset of short children's stories generated by GPT-3.5 and GPT-4 [2], released with the CDLA-Sharing-1.0 licence. We trained and evaluated models on their first version, as described:

- The 2,141,709 stories are split into 2,119,719 train and 21,990 validation stories.
- Their experimental setup [2] called for the GPT-2 [68] tokenizer, of which only a subset vocabulary $\mathbb{V} = [\![1, 29233]\!]$ appears in TinyStories.
- Following the GPT-style teacher-forcing regime for training/evaluation [8], raw sequences (stories) from the train set are packed (by two preprocessing workers) into 229,367 (S) chunks of 2048 (T) tokens each. This setup provides 469,514,249 (N) ground-truth[17] token samples for training.
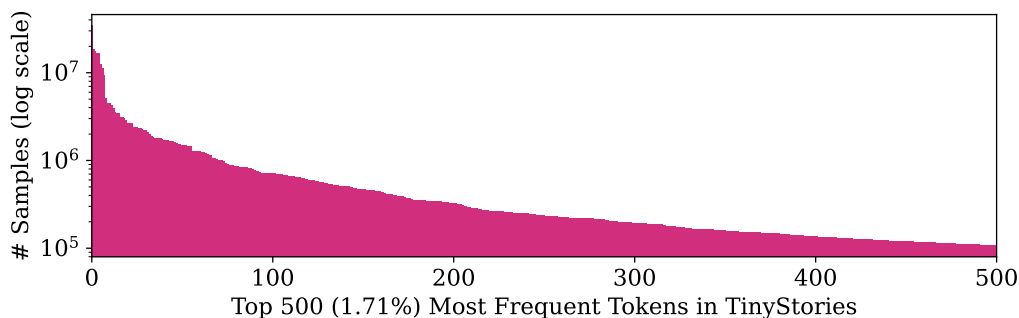
Figure 4: The 500 most frequent classes from TinyStories [2] exhibit significant sample imbalance. Despite the synthetic nature of TinyStories, such a distribution is typical of natural language [11, 12].

A.1 Alternative (Real) Datasets

The study of $\mathcal{NC}$ in causal language modelling at the token level would be unreasonably expensive, so the motivation to use a small dataset is clear. However, most commonly used text datasets such as WikiText [116], BookCorpus [117], CommonCrawl [118], or most subsets from the Pile [119] are much too complex and broad to be effectively compressed by CLMs of the scale that we work with.

WikiText-2 and WikiText-103 present significant drawbacks for our experiments. Both datasets contain a considerable amount of low-quality data that does not concentrate on essential linguistic structures such as grammar, vocabulary, facts, and reasoning. WikiText-2 has a similar empirical vocabulary to TinyStories under the GPT-Neo [80] tokenizer (27K vs. 29K) but only has around 34K rows of training data compared to 2.1M in TinyStories. Our small-scale NC experiment on WikiText-2 revealed that the models were very brittle and prone to overfitting. On the other hand, WikiText-103 is comparably sized to TinyStories but utilizes around 44K unique tokens. Our CLMs trained on WikiText-103 struggled to produce coherent sentences, likely due to the excessive breadth and information, as noted by the authors of TinyStories. Beyond these two, we were unable to find any real datasets that both followed established scaling laws [14, 15] for CLMs at our scale and are simple enough to suit the analysis of $\mathcal{NC}$.

A.2 On the Use of TinyStories

According to its authors, TinyStories [2] is explicitly designed to preserve the essential elements of natural language, such as grammar, vocabulary, facts, and reasoning, while being smaller and more refined in terms of its breadth and diversity. Unlike large corpora that can overwhelm small language models (SLMs) due to their excessive breadth and diversity, TinyStories offers a concentrated dataset that hones in on core linguistic structures and reasoning capabilities. This is evident in its small vocabulary, consisting of approximately 1500 words that a child would use, and in its 29K empirical vocabulary under the GPT-Neo tokenizer.

---

[17] $N = S(T - 1)$ as we cannot use the first ground-truth nor the last predicted token in any chunk.

Despite its concentrated nature, TinyStories enables models trained on it to produce grammatically correct, factual, and reasonable stories. Additionally, these models can be finetuned on specific instructions found in the TinyStories-Instruct dataset. The authors of TinyStories also demonstrate that their models can creatively produce stories dissimilar enough to their training data, indicating a balanced capability for generalization and creativity.

One particular advantage of TinyStories is the small vocabulary relative to total training tokens, rendering a reasonable number of classes with higher average token counts. This is relevant because the possibility of $\mathcal{NC}$ and a CLM's ability to compress language data into distinct geometries depend partially on the ratios between embedding dimension, vocabulary size, and average token frequency. Conveniently, frequency analysis of the overall dataset produced a distribution (Figure 4) similar to real human language, so TinyStories should provide a good balance for an initial study of this $\mathcal{NC}$.

Additionally, TinyStories has a more regular structure as GPT-3.5/4 was instructed to produce children's stories with certain themes and forms with a conservative vocabulary. We believe this would reduce the amount of clustering noise from the breadth of information and structures in real general data, and allow our smaller CLMs to exhibit some clear trends toward $\mathcal{NC}$.

Furthermore, TinyStories was created using GPT 3.5/4, advanced language models with significantly larger architectures trained on orders of magnitude more tokens; this should help minimize the effect of the synthetic nature of the generated dataset. We also considered a possible effect of model collapse as a result of training on synthetic data [120] and follow-up work [121] suggest that a single iteration of data generation (as generated TinyStories) has a very negligible model collapse.

B Model Architectural Details

Table 2: Sample architectural configuration for a 12-layer 1024-dimensional causal language model (CLM) based on [2] and GPT-Neo [80]. Shallower models have configurations with `attention_layers` and `attention_types` truncated. Narrower models adjust `hidden_size` to their width (d). All other configuration values are the same across models.

SETTING	VALUE
activation_function	gelu_new
architectures	GPTNeoForCausalLM
attention_dropout	0
attention_layers	global, local, global, local, ...
attention_types	[[global, local], 6]
bos_token_id	50256
embed_dropout	0
eos_token_id	50256
gradient_checkpointing	false
hidden_size	1024
initializer_range	0.02
intermediate_size	null
layer_norm_epsilon	1e-05
max_position_embeddings	2048
model_type	gpt_neo
num_heads	16
num_layers	12
resid_dropout	0
summary_activation	null
summary_first_dropout	0.1
summary_proj_to_labels	true
summary_type	cls_index
summary_use_proj	true
torch_dtype	float32
transformers_version	4.28.1
use_cache	true
vocab_size	50257
window_size	256

C Optimization

The training was performed using an adaptation of an open-source causal language modelling script from Huggingface: `https://github.com/huggingface/transformers/blob/main/examples/pytorch/language-modeling/run_clm.py`

- Each model was trained on a single NVIDIA A100 (40GB) GPU for up to 8 hours per epoch.
- Learning rates were set by a linear schedule based on the number of steps with no warm-up.
- Training was performed in `bfloat16` [122] mixed precision.
- The results presented in this work are from two sets of models trained with weight decay $\beta = 0.0005$ [51] and $\beta = 0.1$ [81]. A previous set of models was trained without weight decay and the results are very similar to $\beta = 0.0005$.

Table 3: Batch sizes used to train models on a single NVIDIA A100 (40GB) GPU. Width (d) and depth (L) correspond to `hidden_size` and length of `attention_layers`, respectively, in Table 2.

DEPTH (L) ↓ WIDTH (d) →	64	128	256	512	768	1024
1-layer	16	16	16	16	16	16
2-layer	16	16	16	16	16	8
4-layer	8	8	8	8	8	8
8-layer	8	8	8	4	4	4
12-layer	4	4	4	4	4	4
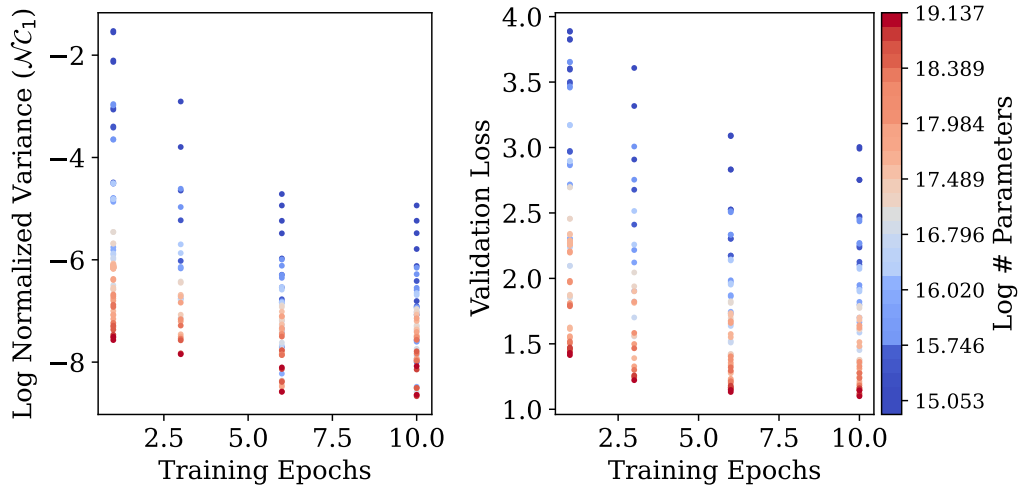
Figure 5: Average (logarithmic) class-distance normalized variance (CDNV, $\mathcal{NC}_1$) (left) and validation (cross-entropy) loss (right) with respect to training epochs.

D Embeddings Collection & $\mathcal{NC}$ Analysis

Codes for (post-)training analysis are hosted on GitHub:

- Main code `https://github.com/rhubarbwu/linguistic-collapse`
- Auxillary package: `https://github.com/rhubarbwu/neural-collapse`

One pass over the train set for embeddings collection can take up to 6 hours on a single NVIDIA A100 (40GB) GPU. Analysis of a single metric for a given model takes less than 5 minutes.

E   Within-Class Variability Collapse with Scale — $\mathcal{NC}_1$
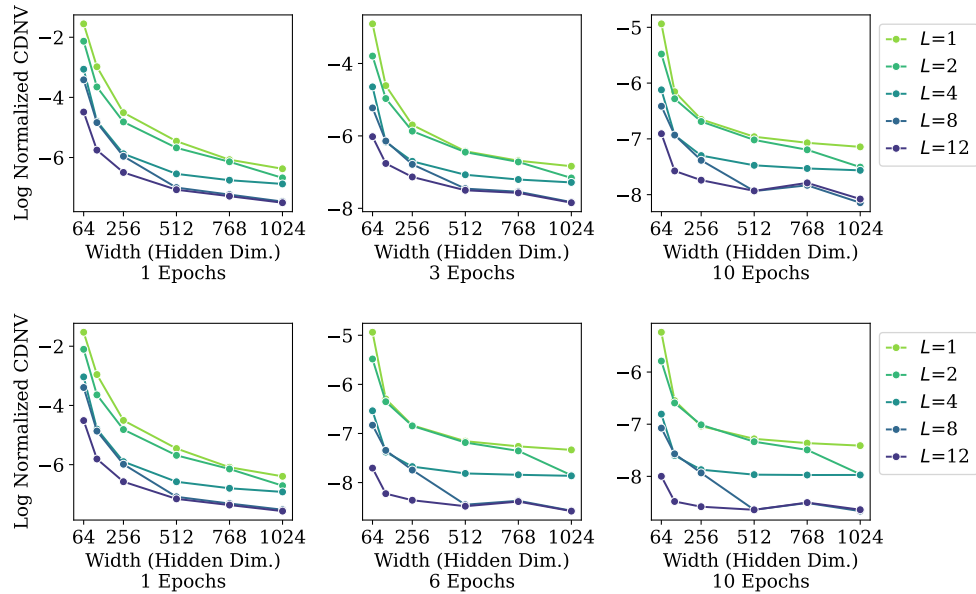
Figure 6: Average (logarithmic) class-distance normalized variance (CDNV) is reduced ($\mathcal{NC}_1$) when scaling width (d) and across training for 1 (left) through 10 (right) epochs with weight decays $\beta = 0.0005$ (top) and $\beta = 0.1$ (bottom).
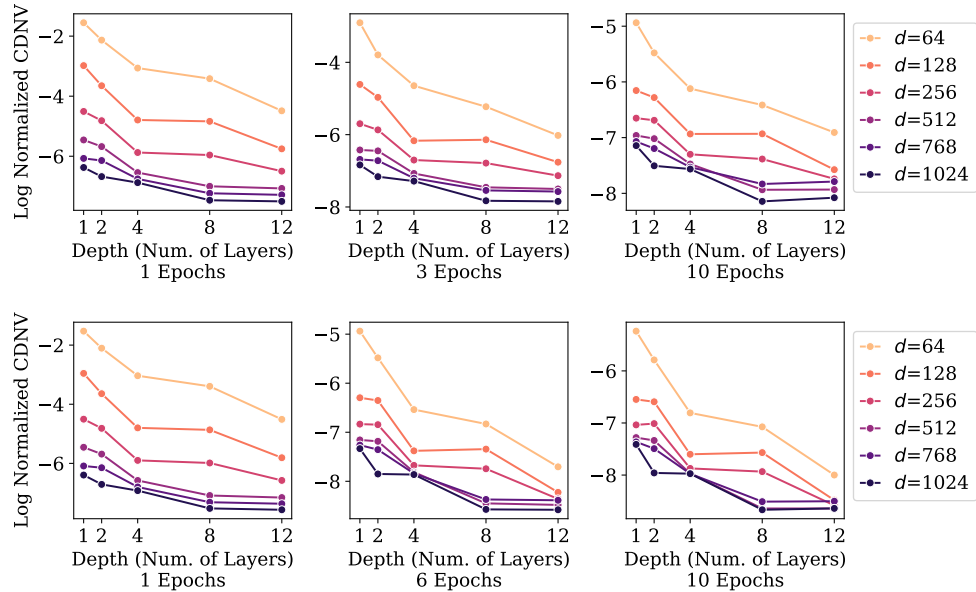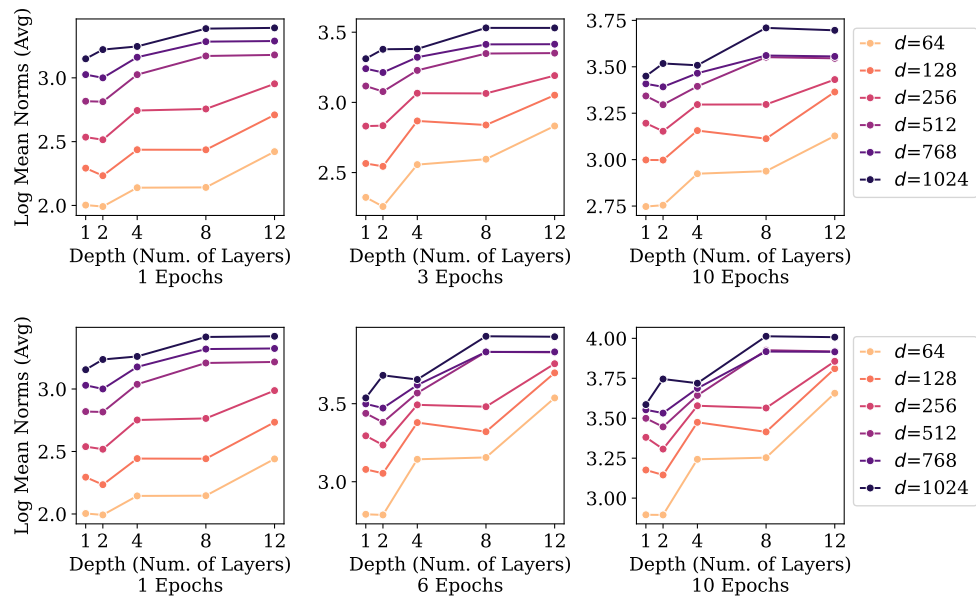

Figure 7: Average (logarithmic) class-distance normalized variance (CDNV) is reduced ($\mathcal{NC}_1$) when scaling depth (L) and across training for 1 (left) through 10 (right) epochs with weight decays $\beta = 0.0005$ (top) and $\beta = 0.1$ (bottom).

F   Mean Norms Growth with Scale — (Related to $\mathcal{NC}_2$)
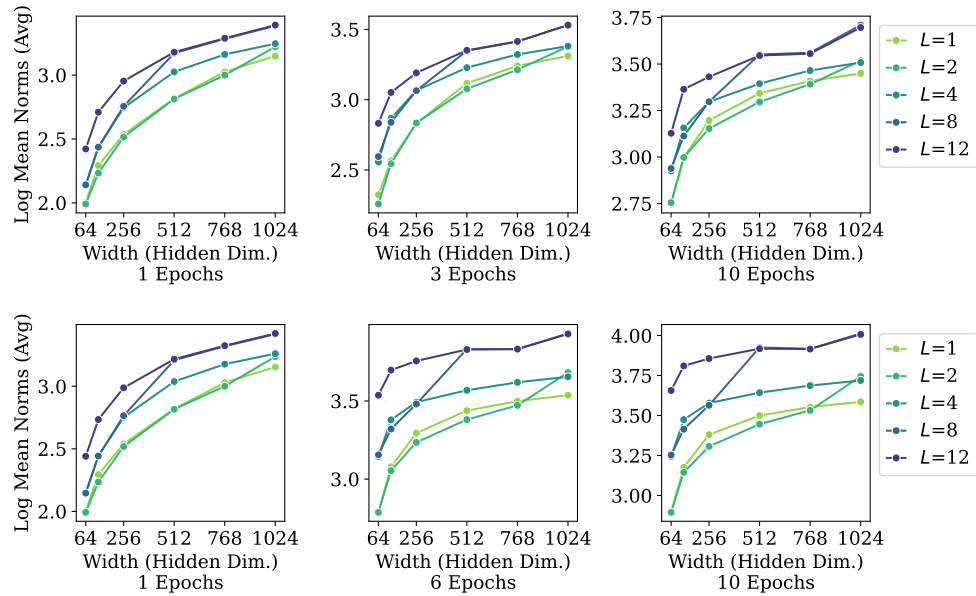
Figure 8: Logarithmic class mean norms grow when scaling width (d) and across training for 1 (left) through 10 (right) epochs with weight decays $\beta = 0.0005$ (top) and $\beta = 0.1$ (bottom).

Figure 9: Logarithmic class mean norms grow when scaling depth (L) and across training for 1 (left) through 10 (right) epochs with weight decays $\beta = 0.0005$ (top) and $\beta = 0.1$ (bottom).

G Equinormness with Scale — $(\mathcal{G})\mathcal{NC}_2$
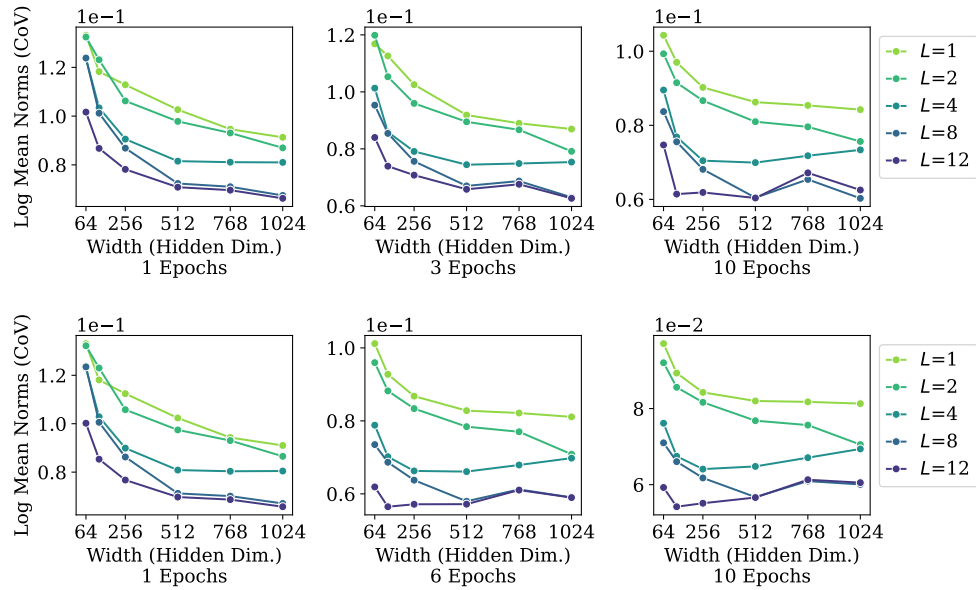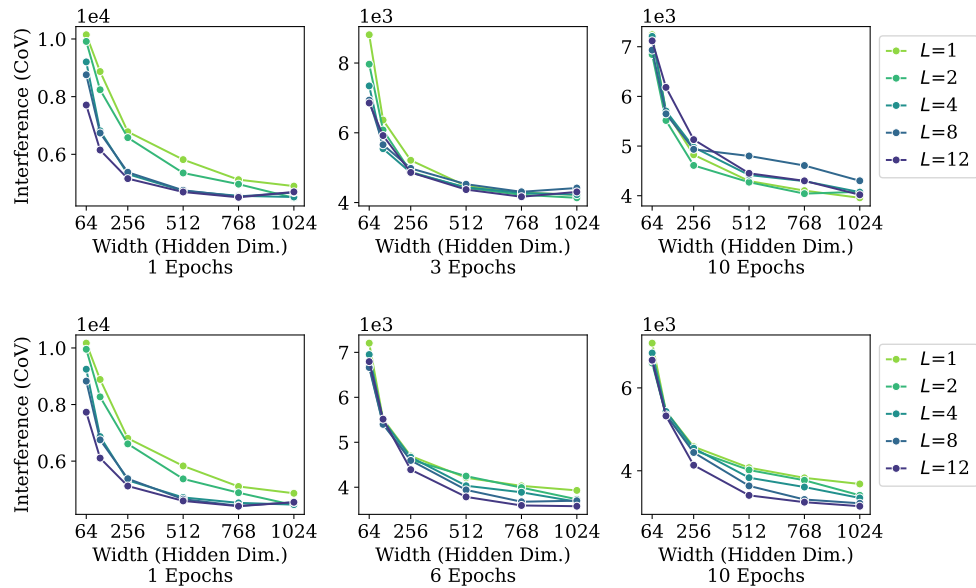

Figure 10: Variation in (logarithmic) norms decreases ($\mathcal{NC}_2$) when scaling width (d) in models trained for 1 (left) through 10 (right) with weight decays $\beta = 0.0005$ (top) and $\beta = 0.1$ (bottom). Note that the degree of equinormness eventually plateaus for sufficiently deep and trained models.
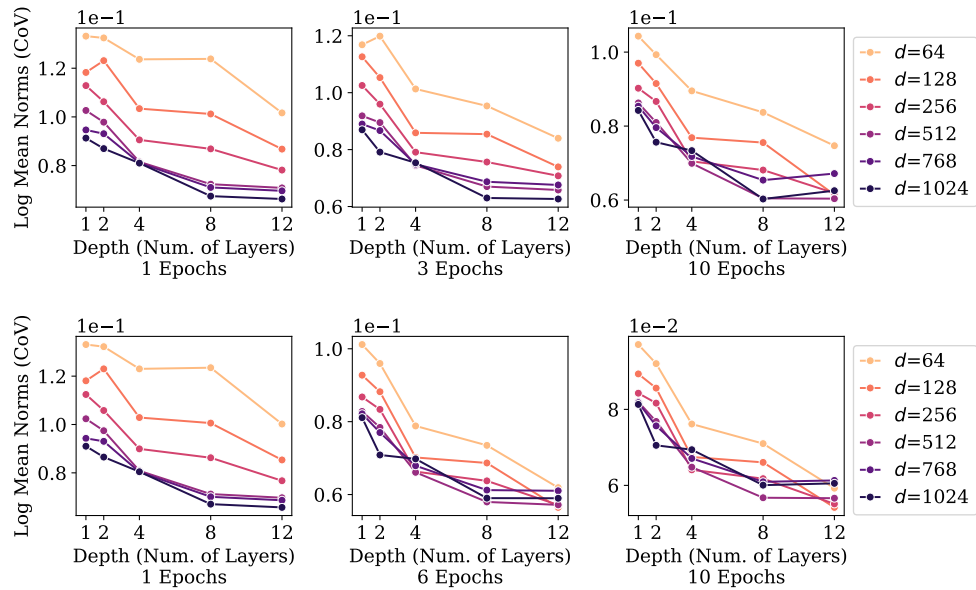
Figure 11: Variation in (logarithmic) norms decreases ($\mathcal{NC}_2$) when scaling depth (L) in models trained for 1 (left) through 10 (right) with weight decays $\beta = 0.0005$ (top) and $\beta = 0.1$ (bottom).

H   Interference with Scale — (Related to $\mathcal{NC}_2$)
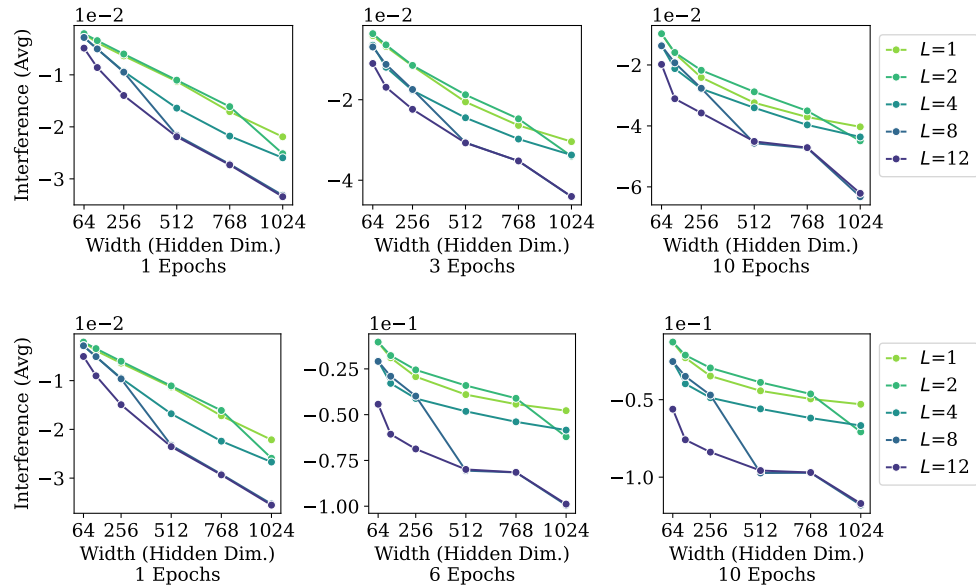
Figure 12: Average interference decreases (to some extent) when scaling width (d) in models trained for 1 (left) through 10 (right) with weight decays $\beta = 0.0005$ (top) and $\beta = 0.1$ (bottom).
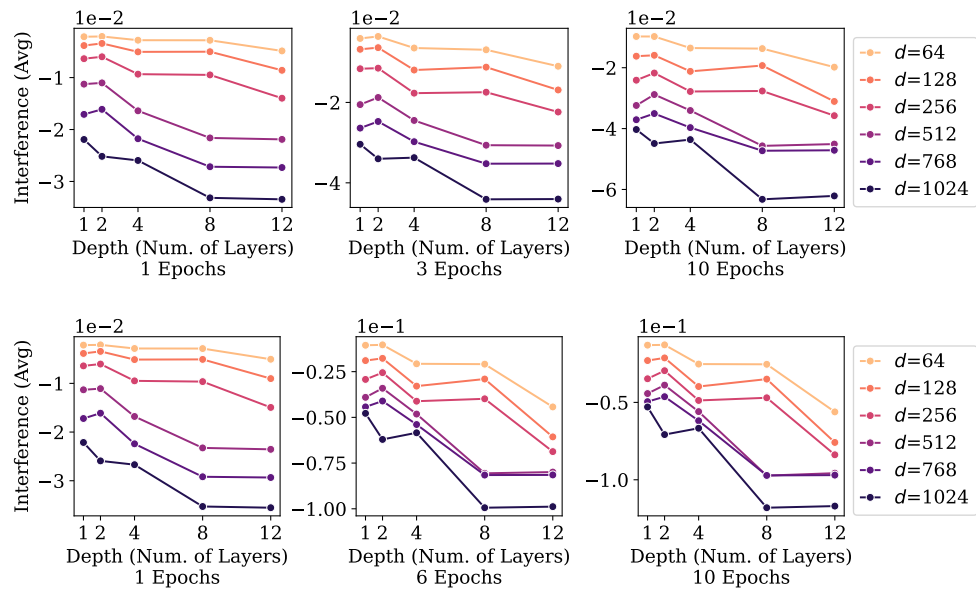
Figure 13: Average interference decreases (to some extent) when scaling depth (L) in models trained for 1 (left) through 10 (right) with weight decays $\beta = 0.0005$ (top) and $\beta = 0.1$ (bottom).

I Equiangularity with Scale — (Against $\mathcal{NC}_2$)

Figure 14: Variation in interference roughly increases when scaling width (d) in models trained for 1 (left) through 10 (right) with weig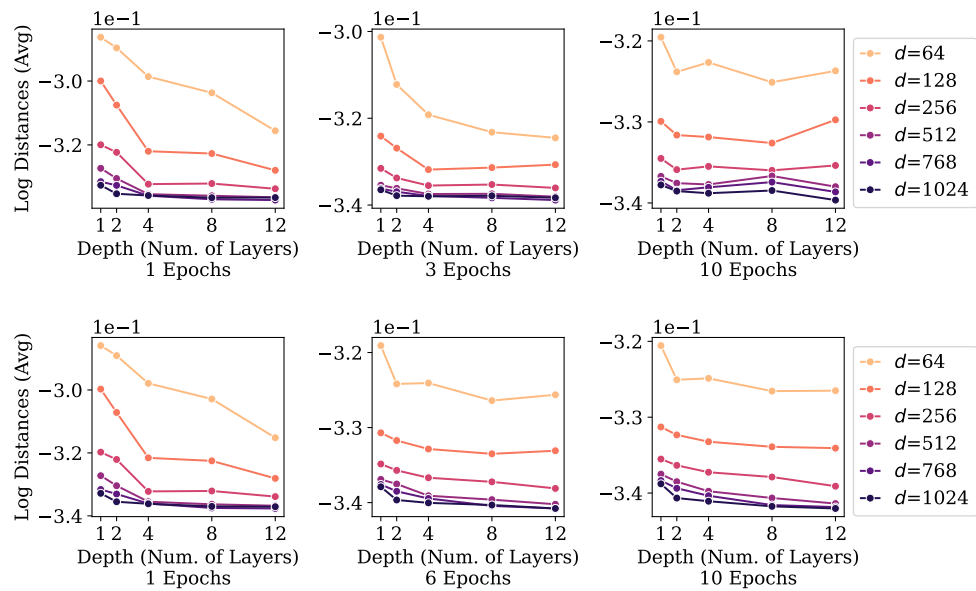ht decays $\beta = 0.0005$ (top) and $\beta = 0.1$ (bottom). Note this trend is against equiangularity, affirming the traditional $\mathcal{NC}2$ to be less useful than $\mathcal{GNC}2$ [19].
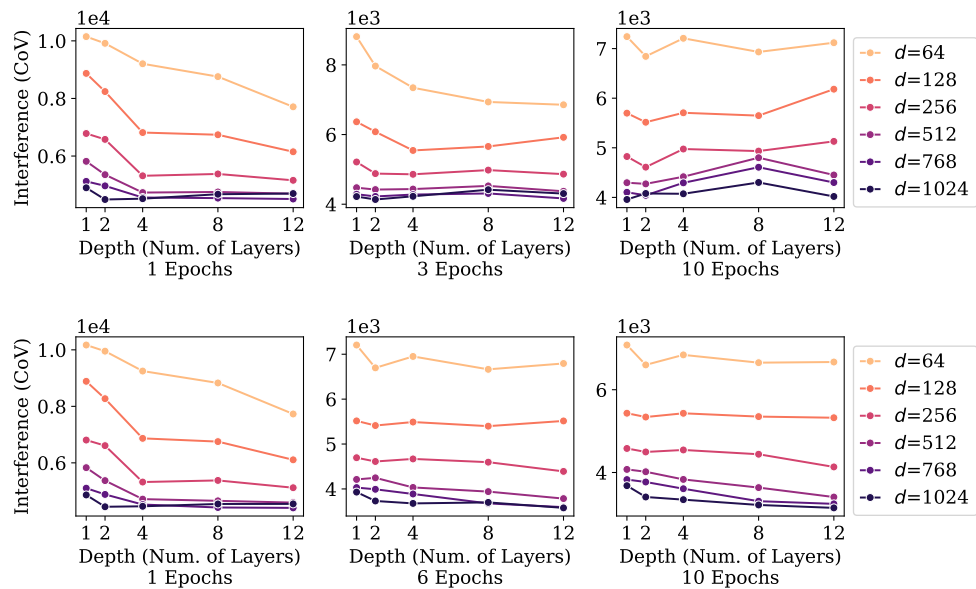
Figure 15: Variation in interference increases when scaling depth (L) in models trained for 1 (left) through 10 (right) with weight decays $\beta = 0.0005$ (top) and $\beta = 0.1$ (bottom). Note this trend is against equiangularity, affirming the traditional $\mathcal{NC}_2$ to be less useful than $\mathcal{GNC}_2$ [19].

J Logarithmic Distances with Scale — (Related to $\mathcal{GNC}_2$)
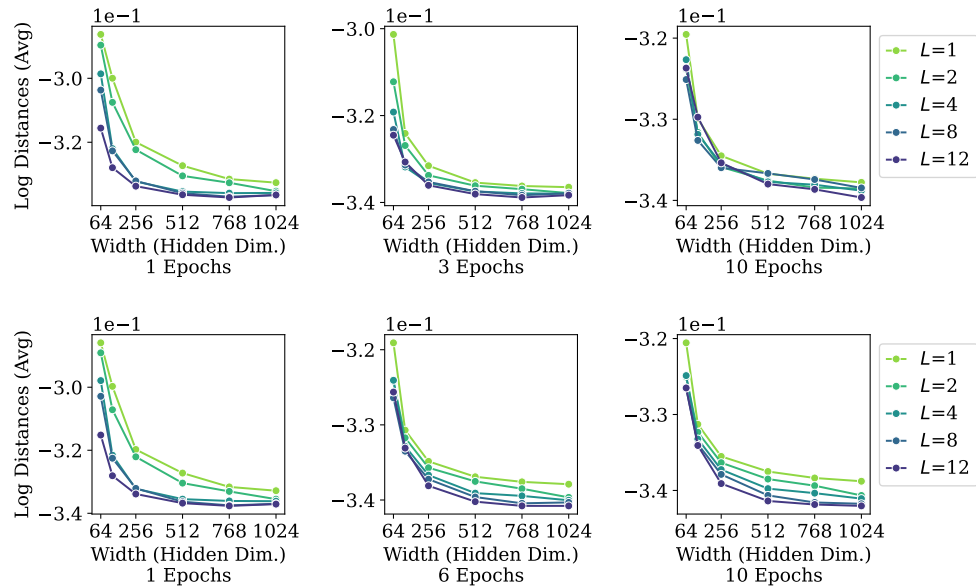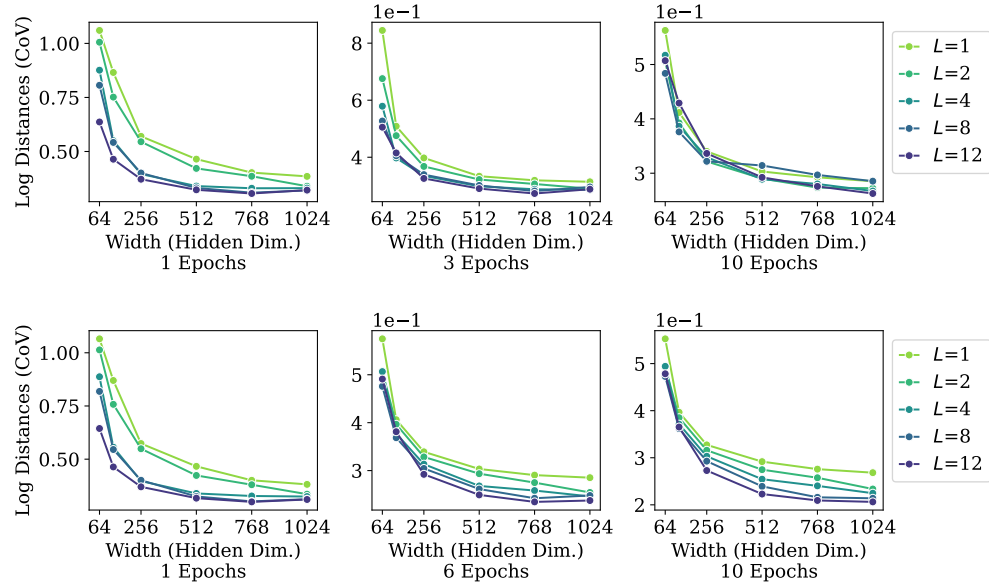

Figure 16: Average logarithmic distance decreases when scaling width (d) in models trained for 1 (left) through 10 (right) with weight decays $\beta = 0.0005$ (top) and $\beta = 0.1$ (bottom).

Figure 17: Average logarithmic distance decreases when scaling depth (L) in models trained for 1 (left) through 10 (right) with weight decays $\beta = 0.0005$ (top) and $\beta = 0.1$ (bottom).

K Hyperspherical Uniformity with Scale — $\mathcal{GNC}_2$

Figure 18: Variation in logarithmic distances decreases when scaling width (d) in models trained for 1 (left) through 10 (right) with weight decays $\beta = 0.0005$ (top) and $\beta = 0.1$ (bottom). This consistent trend towards hyperspherical uniformity affirms that $\mathcal{GNC}2$ [19] is more useful than $\mathcal{NC}2$.
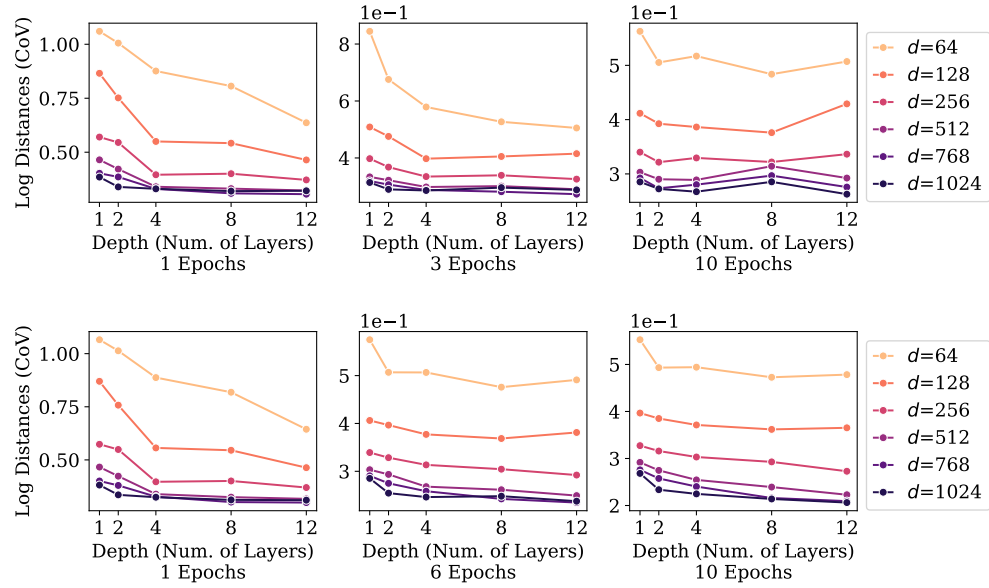

Figure 19: Variation in logarithmic distances decreases when scaling depth (L) in models trained for 1 (left) through 10 (right) with weight decays $\beta = 0.0005$ (top) and $\beta = 0.1$ (bottom). This consistent trend towards hyperspherical uniformity affirms that $\mathcal{GNC}2$ [19] is more useful than $\mathcal{NC}2$.

L Correlations of $(\mathcal{G})\mathcal{NC}_2$ with Generalization Performance
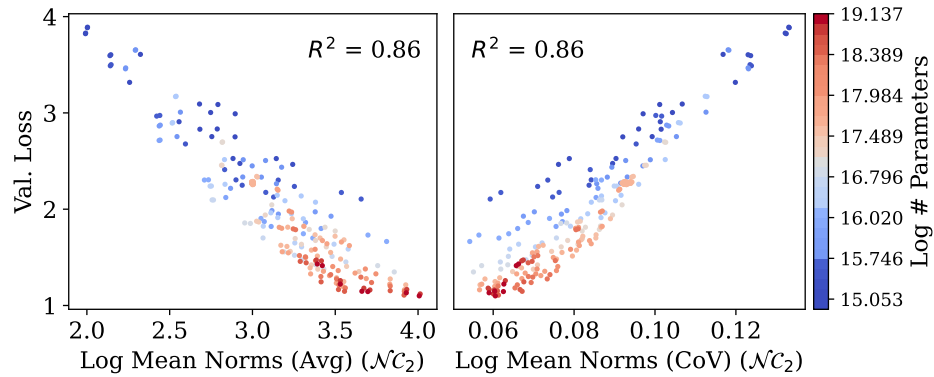

Figure 20: Generalization (validation loss) shows some correlation with logarithmic mean norms in both their average (left) and variations (i.e. equinormness, $\mathcal{NC}_2$) (right).
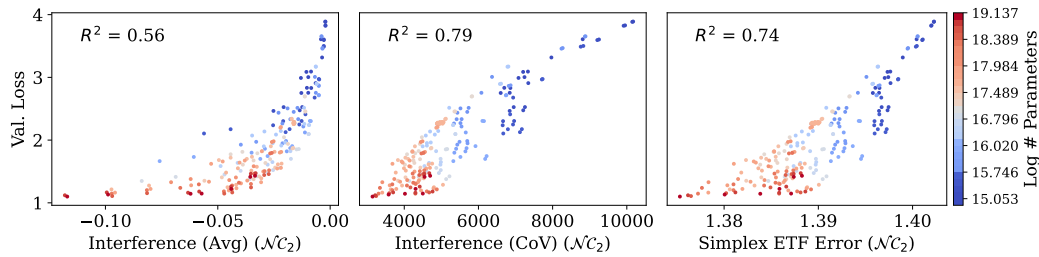

Figure 21: Generalization (validation loss) correlated with average interference (left) and its variation (i.e. equiangularity, $\mathcal{NC}_2$) (centre). We also computed the empirical measure from [5] (right).
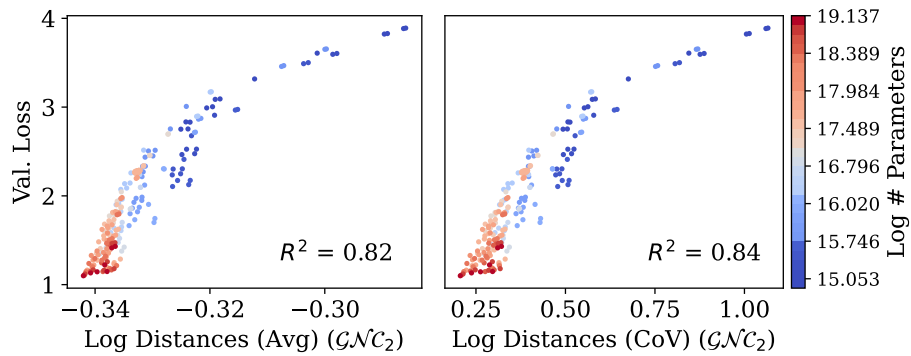

Figure 22: Validation loss shows some correlation with average (logarithmic) kernel distances and with their variation (i.e. hyperspherical uniformity, $\mathcal{GNC}_2$) (right).

M   Self-Duality with Scale — (Against $\mathcal{NC}_3$)

Self-duality ($\mathcal{NC}_3$) was originally the convergence of classifiers to means up to rescaling [1]:

$$\left\| \frac{\boldsymbol{w}_c}{\|\boldsymbol{w}_c\|_2} - \hat{\boldsymbol{\mu}}_c \right\|_2 \to 0, \quad \forall c \tag{11}$$

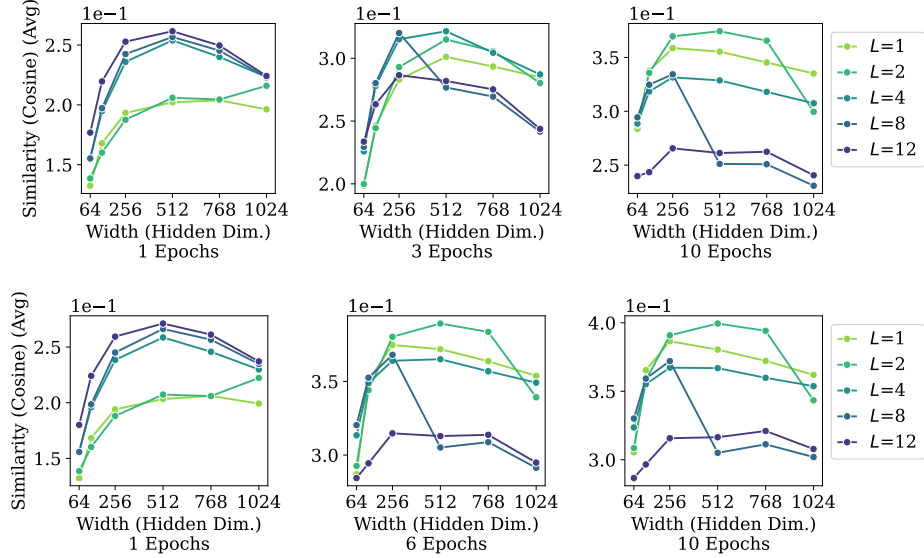Instead, we use class-wise cosine similarity (Equation 8) and its variation ($\mathcal{UNC}3$).

Figure 23: Average classifier alignment increases $\mathcal{NC}3$ when training for 1 (left) through 10 (right) with weight decays $\beta = 0.0005$ (top) and $\beta = 0.1$ (bottom). However, we see no meaningful trend when scaling width d, suggesting that $\mathcal{NC}3$ does not coalesce with language modelling training.
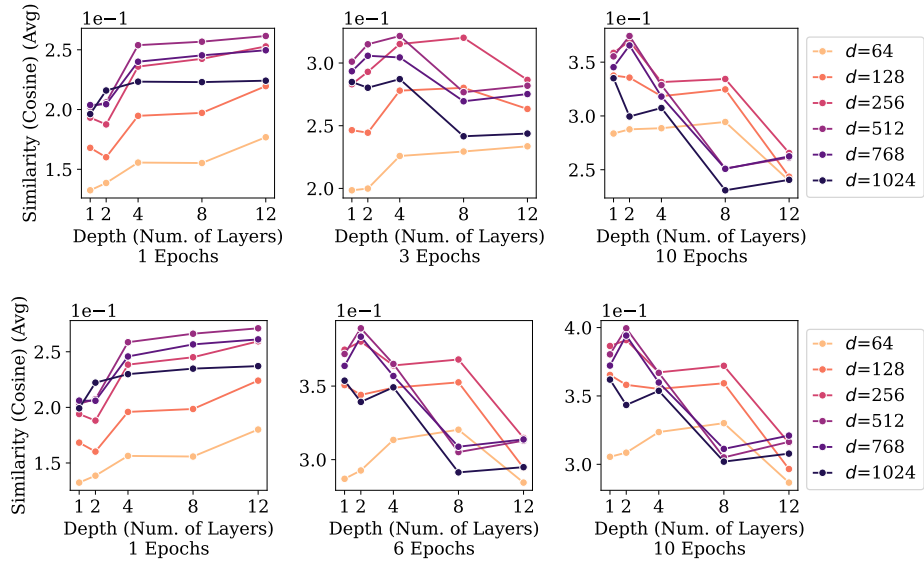

Figure 24: Average classifier alignment increases $\mathcal{NC}3$ when training for 1 (left) through 10 (right) with weight decays $\beta = 0.0005$ (top) and $\beta = 0.1$ (bottom). However, we see no meaningful trend when scaling depth L, suggesting that $\mathcal{NC}3$ does not coalesce with language modelling training.
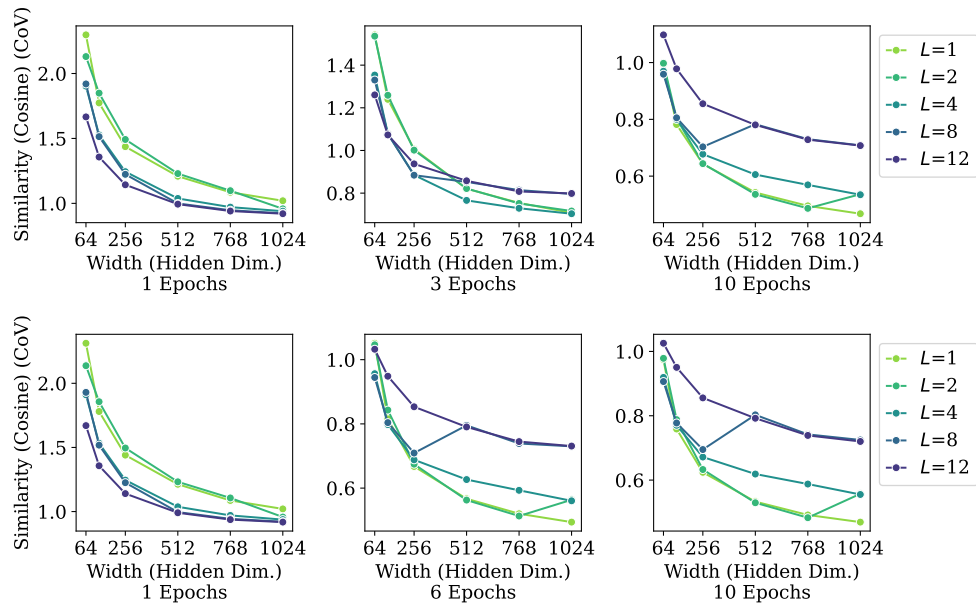
N Uniformity Duality with Scale — $\mathcal{UNC}_3$

Figure 25: Variation in classifier alignment decreases when scaling width (d) in models trained for 1 (left) through 10 (right) with weight decays $\beta = 0.0005$ (top) and $\beta = 0.1$ (bottom).
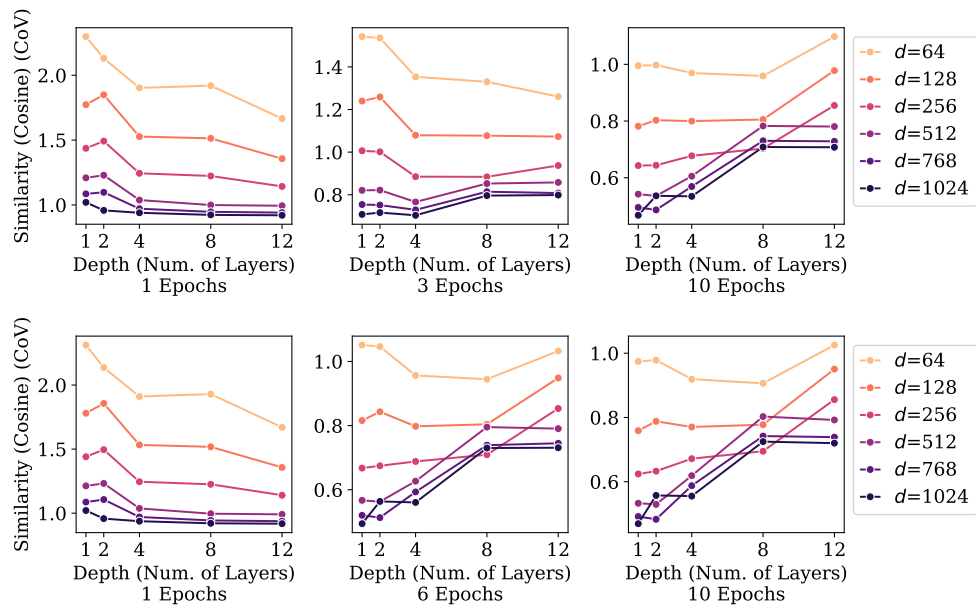

Figure 26: Variation in classifier alignment increases when scaling depth (L) in models trained for 1 (left) through 10 (right) with weight decays $\beta = 0.0005$ (top) and $\beta = 0.1$ (bottom). This negative trend of $\mathcal{UNC}3$ in more learnt models (right) suggests that the link of $(\mathcal{U})\mathcal{NC}3$ with scale and performance is still weak.

O   Correlations of $(\mathcal{U})\mathcal{NC}_3$ with Generalization Performance

Figure 27: Generalization (validation loss) correlated with average dot-product similarity (for interpretability purposes only) (left) and cosine similarity (classifier alignment, $\mathcal{NC}_3$) (centre). We also computed the empirical measure from [5] (right).
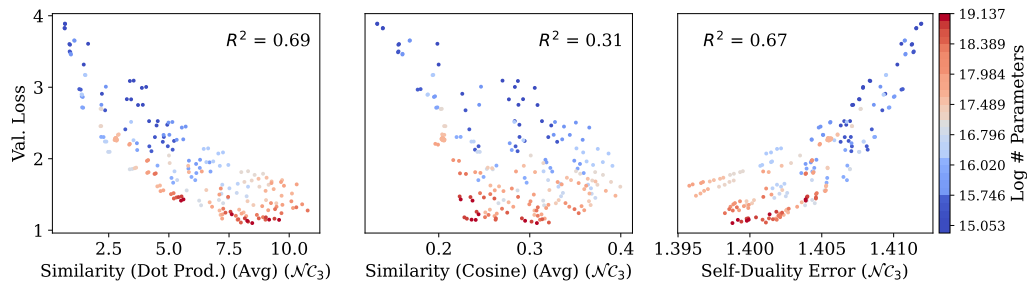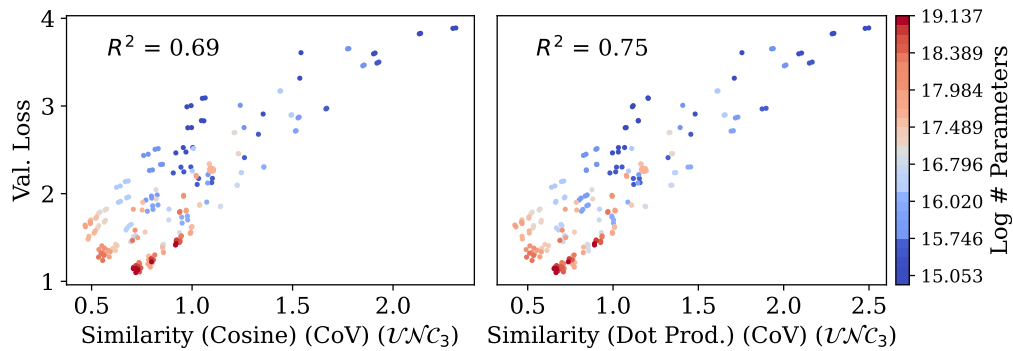

Figure 28: Generalization (validation loss) correlated with variation in dot-product similarity (for interpretability purposes only) (left) and cosine similarity (uniform duality, $\mathcal{UNC}_3$) (right).

P   Classifier Agreement — $\mathcal{NC}_4$

For computational reasons, we compute Equations 9, 10 using a simple decomposition:

$$\underset{c\in\mathbb{V}}{\operatorname{argmin}} \|\boldsymbol{h}_b - \boldsymbol{\mu}_c\|_2 = \underset{c\in\mathbb{V}}{\operatorname{argmin}} \left(\|\boldsymbol{h}_b\|^2 + \|\boldsymbol{\mu}_c\|^2 - 2\boldsymbol{h}_b^\top \boldsymbol{\mu}_c \right), \tag{12}$$

where $b \in [1, B]$ and $c \in \mathbb{V}$ with batch size B and vocabulary $\mathbb{V}$.

Figure 29: Classifier agreement improves when scaling width (d) in models trained for 1 (left) through 10 (right) with weight decays $\beta = 0.0005$ (top) and $\beta = 0.1$ (bottom).

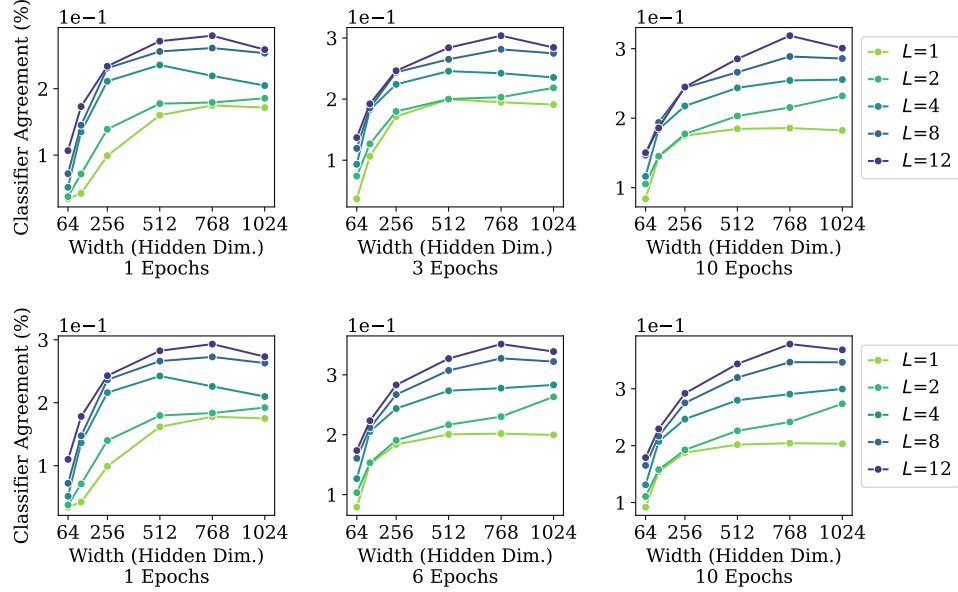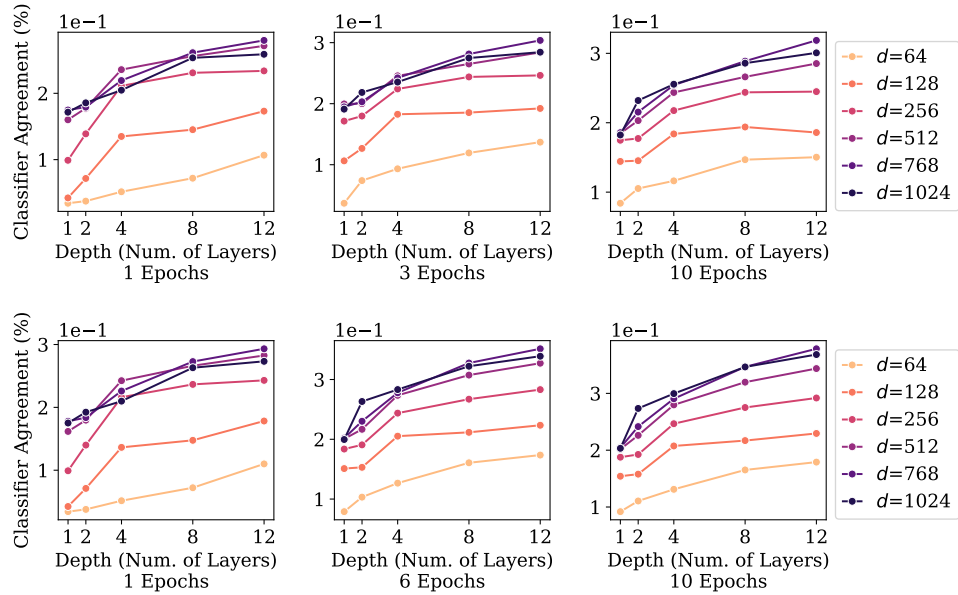

Figure 30: Classifier agreement improves when scaling depth (L) in models trained for 1 (left) through 10 (right) with weight decays $\beta = 0.0005$ (top) and $\beta = 0.1$ (bottom).

Q Examples for Interpretability

This section presents token-wise interpretability results from top-layer embeddings from our most learned models (trained for 10 epochs). Our largest one is publicly available: `https://huggingface.co/rhubarbwu/TinyStories-12x1024_10L`.
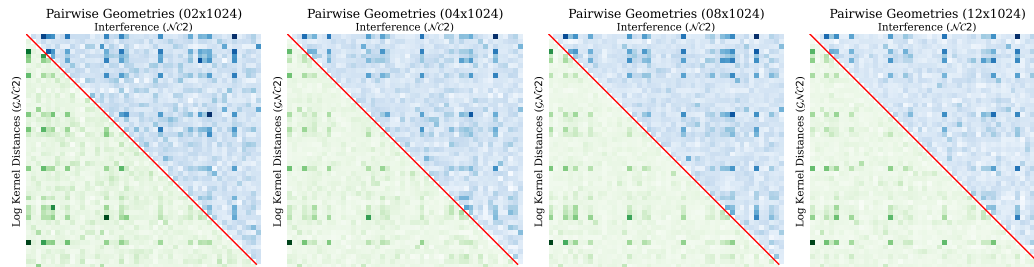
Figure 31: Pairwise interactions between token mean embedding vectors across models of fixed width ($d = 1024$) and increasing depths ($L = 2, 4, 8, 12$). Interference ($\mathcal{NC}2$) decreases on average but only slightly becomes more uniform (top-right, blue). In contrast, logarithmic kernel distances ($\mathcal{GNC}2$) decrease and become more evenly spread, with some outlier pairs (bottom-left, green).
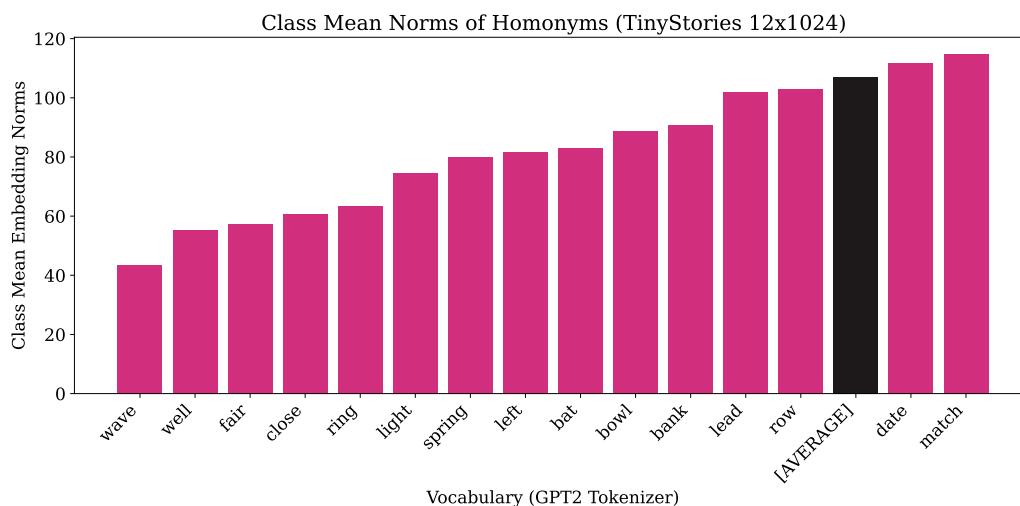
Figure 32: Under `TinyStories-12x1024_10L`, these fifteen homonyms have much shorter mean embedding norms (i.e. closer to the global centre) than the average token. This is expected since homonyms typically present conflicts and interference.
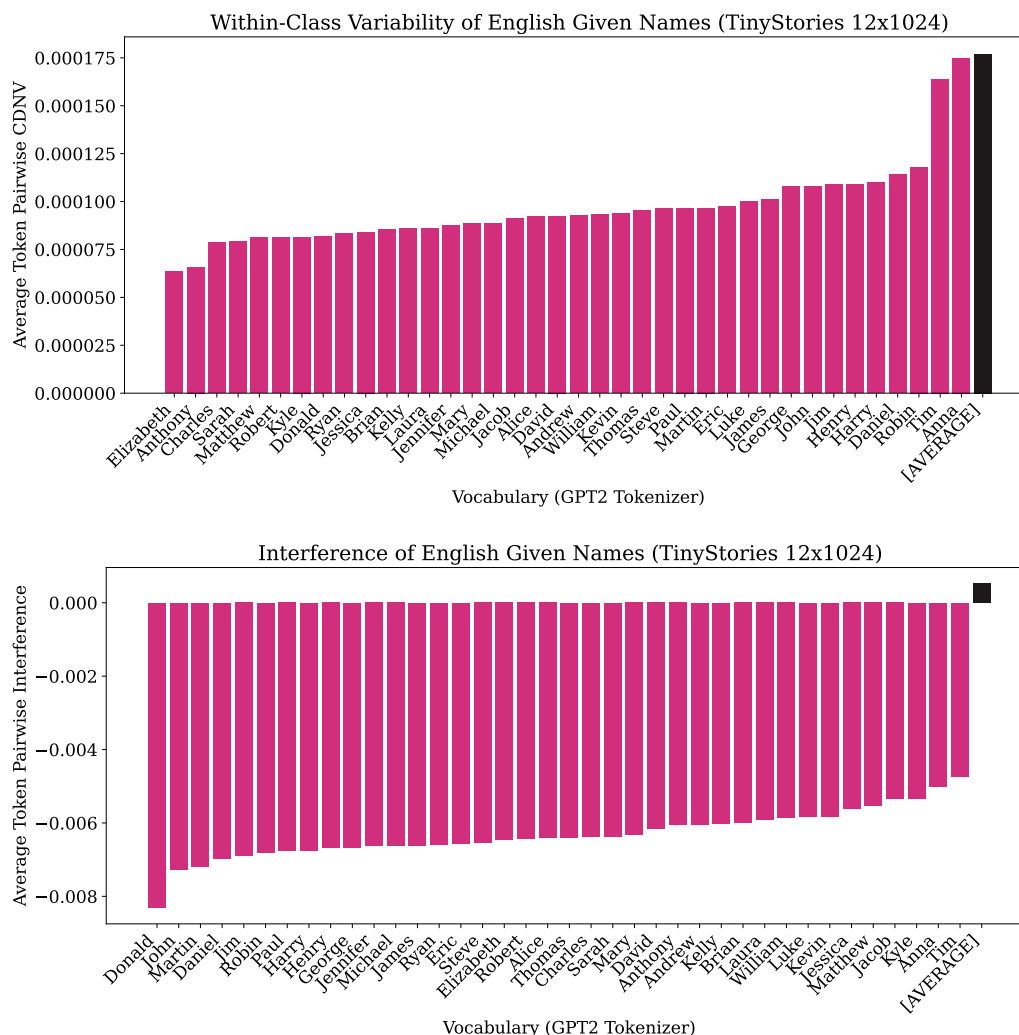
Figure 33: Under `TinyStories-12x1024_10L`, the average within-class variability (top) and inter-ference (bottom) of some English first names were far below those of the average token. This might be because names are distinct and are not typically used in the same contexts as other words (aside from articles). The only names to have CDNV close to that of the average token are "Anna" and "Tim". Note that the positive interference of the average token (right) is not a typo.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims of this paper listed in the abstract and introduction are simply the emergence of the $\mathcal{NC}$ phenomena and its relationship with generalization in CLMs, which are the scope and results of the paper.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations of the experiments themselves (e.g. small models, limited sample size) as well as the possibility that our formulations could be improved in the future (there could be more suitable metrics). See Sections 6, 7.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The equations and equivalent decompositions used are all clear and numbered in the paper. Their sources are all provided as references. We do not introduce new proofs or non-trivial expressions. See Section 3.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Our methods for training the models, collecting embeddings, and analyzing metrics are clearly described in the paper, with supporting details in Appendices A, B, C. Links to code are provided in the Abstract and Section 3.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Links to the main code and auxiliary library are provided in the Abstract and Section 3, respectively. Main code repository includes training, collection, and analysis scripts, as well as a README briefly documenting how to use them.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: Appendices A, B, C, D provide the necessary information on model, data, training, and collection specifications to reproduce our experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We trained multiple copies of the same models and observed patterns consistent with the ones presented in our paper. We also state the number of data points and permutation trials we performed for our significance test. See Figure 1 and Sections 4 and 5.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: We provide details on the number of dataloader/preprocessing works, the specific GPU model used and the amount of RAM and training/collection time required. See Appendices A, B, C, D.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: We have read the Code of Ethics and determined that our paper does not violate the guidelines.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The paper includes a discussion on how $\mathcal{NC}$ might be extended to measure fairness in LLMs. Naturally this may inform positive or adversarial manipulations of LLMs.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: We do not use pretrained models or scraped data. Our dataset is synthetic, public and transparent with no risks. The training procedure we use is the standard causal language modeling regime.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The pre-training dataset and its licence are provided in Appendix A and the authors of basic GPT Neo architectures are credited in Appendix B.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Timestamped CSV files of results that we reported in our paper are included in the repository linked in the Abstract. A link to a reference model is provided in Appendix Q.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: We do not work with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: We do not work with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.