

---

# Flatness is Necessary, Neural Collapse is Not: Rethinking Generalization via Grokking

---

**Ting Han**

Lamarr Institute, TU Dortmund, Germany  
and Institute for AI in Medicine, UK Essen  
ting.han@tu-dortmund.de

**Linara Adilova**

RC Trust, TU Dortmund, Germany

**Henning Petzka**

Ruhr University Bochum, Germany

**Jens Kleesiek**

Institute for AI in Medicine, UK Essen

**Michael Kamp**

Lamarr Institute, TU Dortmund, Germany  
and Institute for AI in Medicine, UK Essen  
michael.kamp@tu-dortmund.de

## Abstract

Neural collapse, i.e., the emergence of highly symmetric, class-wise clustered representations, is frequently observed in deep networks and is often assumed to reflect or enable generalization. In parallel, flatness of the loss landscape has been theoretically and empirically linked to generalization. Yet, the causal role of either phenomenon remains unclear: Are they prerequisites for generalization, or merely by-products of training dynamics? We disentangle these questions using grokking, a training regime in which memorization precedes generalization, allowing us to temporally separate generalization from training dynamics and we find that while both neural collapse and relative flatness emerge near the onset of generalization, only flatness consistently predicts it. Models encouraged to collapse or prevented from collapsing generalize equally well, whereas models regularized away from flat solutions exhibit delayed generalization. Furthermore, we show theoretically that neural collapse implies relative flatness under classical assumptions, explaining their empirical co-occurrence. Our results support the view that relative flatness is a potentially necessary and more fundamental property for generalization, and demonstrate how grokking can serve as a powerful probe for isolating its geometric underpinnings.

## 1 Introduction

Overparameterized neural networks continue to challenge classical learning theory. Despite their capacity to memorize arbitrary labels [Zhang et al., 2016], they often generalize well on natural data and, in some cases, only begin to generalize after complete memorization, a phenomenon known as grokking [Power et al., 2022]. This apparent contradiction has motivated a renewed interest in identifying geometric signatures of generalization. Two candidates have emerged as particularly prominent: the flatness of the loss landscape [Keskar et al., 2017, Jiang et al., 2020, Petzka et al., 2021] and the neural collapse (NC) phenomenon [Papayan et al., 2020, Mixon et al., 2022, Zhou et al., 2022]. Both tend to appear late in training and are frequently associated with good generalization, yet their precise causal roles remain unclear.

In this paper, we challenge the common conjecture that NC is necessary for generalization [Mixon et al., 2022, Zhou et al., 2022, Sůkeník et al., 2023, Zhu et al., 2021]. Our key idea is to leverage grokking as a unique observational window: since generalization emerges only after prolonged memorization, it enables a clean separation between training dynamics and generalization. This allows us to ask—do neural collapse and flatness merely correlate with generalization, or do they contribute to it?

We measure both phenomena in the penultimate layer. For flatness, we adopt the relative flatness metric proposed by Petzka et al. [2021], which considers the Hessian trace normalized by the weight norm, a quantity theoretically and empirically shown to align with generalization. To make this computable in large state-of-the-art neural networks, we employ the alternative closed-form upper bound introduced by [Walter et al., 2024], which is valid in the penultimate layer under cross-entropy loss. For NC, we measure the class-wise clustering of penultimate-layer representations through empirical variance and mean properties, tracking its emergence throughout training. In grokking experiments, we find that flatness emerges alongside generalization, never before. NC, in contrast, appears earlier during memorization, suggesting that it may result from, rather than enable, generalization (cf. Fig. 1).

To isolate their functional roles, we perturb training dynamics. When we suppress collapse-style clustering, networks still generalize robustly; test accuracy remains high despite high clustering values. Flatness, however, is unaffected. In contrast, regularizing against flatness, i.e., encouraging sharp solutions, reliably delays generalization. These findings point to flatness as a more fundamental driver of generalization. To explain the frequent co-occurrence of NC and flatness, we provide a theoretical result: under classical NC assumptions, collapse *implies* relative flatness. This unifies previously observed correlations under a single geometric framework, but also shows that flatness can arise via other mechanisms. Neural collapse is thus a sufficient but non-exclusive path toward flatness. Finally, we revisit the theoretical framework of Petzka et al. [2021], which asserts that generalization requires not only flatness, but also *representativeness*, i.e., the alignment between learned features and the true data distribution. Our experiments confirm this dependency. Representativeness, when approximated, improves only as generalization sets in, highlighting that neither flatness nor collapse alone ensures generalization without meaningful feature alignment.

Together, these results place neural collapse in a new light. While it often accompanies generalization, it is not a prerequisite. Instead, it may serve as an inductive bias toward flatness under standard training dynamics. Relative flatness, by contrast, appears empirically necessary and theoretically justified, especially when labels are locally constant and features are representative—assumptions which appear to hold in our experimental settings.

Our work provides new insight into the geometry of generalization:

- We show empirically that relative flatness, but not neural collapse, is potentially necessary for generalization;
- We prove that neural collapse implies relative flatness under classical assumptions;
- We demonstrate that delayed generalization can be actively induced by manipulating flatness, even in real-world architectures.

These insights challenge prevailing narratives about generalization in deep learning and suggest new levers for training, regularization, and model diagnostics.

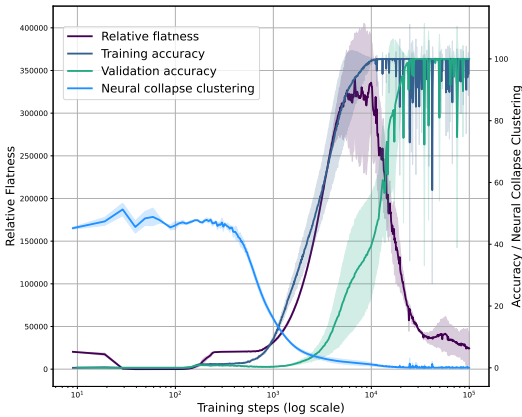


Figure 1: Neural collapse clustering and relative flatness in grokking. While both correlate with generalization, neural collapse emerges early during memorization, whereas flatness only drops sharply when generalization begins, highlighting flatness as a better indicator of generalization onset.

## 2 Related Work

Despite substantial empirical success, the mechanisms driving generalization in deep neural networks remain poorly understood. Two pronounced geometric perspectives, *neural collapse* and *loss surface flatness*, have emerged as phenomena empirically correlated with generalization, but their interrelation and causal roles remain ambiguous. In this work, we leverage the properties of grokking to disentangle necessity from sufficiency and to assess which of these phenomena are fundamentally required for generalization.

**Neural Collapse and Generalization** *Neural collapse* (NC) describes a geometric convergence during training a classifier where penultimate layer features from the same class collapse to a single mean, the means form a simplex equiangular tight frame, and classifier weights align accordingly [Papayan et al., 2020]. This phenomenon has been widely observed at late training stages in overparameterized networks [Han et al., 2021, Graf et al., 2021, Rangamani and Banburski-Fahey, 2022, Wu and Papayan, 2024, Zhou et al., 2025] and has become a widely accepted indicator for generalization and robustness. Under idealistic assumption of an unconstrained features model, i.e., a model which optimizes features as free variables, NC has been shown to be the global optimal solution for various loss functions [Mixon et al., 2022, Zhou et al., 2022, Súkeník et al., 2023, Zhu et al., 2021]. Similar results for compatibility of zero test error and collapse of variances (so only NC1 condition) were shown for mean-field networks [Wu and Mondelli, 2025]. Practical applications have leveraged NC properties for downstream performance. For instance, Wu et al. [2025] use a feature separation loss based on NC to improve out-of-distribution detection. Similarly, Munn et al. [2024] show that lower geometric complexity in pre-trained representations fosters NC and leads to better few-shot generalization in transfer learning. Galanti et al. [2021] shows theoretically and empirically that NC is the reason for transfer learning to work better. These findings suggest that NC may be *sufficient* for generalization in certain regimes. However, the open question remains: is NC *necessary* outside of the unconstrained features models? For example, NC is a debated matter in the context of imbalanced learning, where it was observed that minority collapse can happen, i.e., the means of the minority classes merge thus preventing classification [Fang et al., 2021]. At the same moment it is possible to induce NC by enforcing it through training, even in the unbalanced data setup [Yang et al., 2022]. For deep linear networks it was also theoretically shown that global optimum exhibits NC, but even more: Under unbalanced data NC gets means proportional to the class size [Dang et al., 2023]. Hui et al. [2022] argues for measuring NC on the test set for understanding generalization properties and demonstrates that test set NC is hurting generalization performance on the downstream tasks, i.e., representations with strong NC on the test set are bad for transfer learning. Also, large enough signal-to-noise ratio was observed to be an important criterion for NC to signify good generalization [Hong and Ling, 2024].

Overall, the existing research suggests that NC is common, but might not be required for generalization: a network can generalize without NC and some setups (like noisy labels) can break the link between generalization and NC. Our work addresses this question directly by constructing settings in which models generalize without exhibiting NC, demonstrating that it is not a prerequisite for generalization.

**Flatness and Generalization** The connection between loss surface flatness and generalization has long been suspected [Hochreiter and Schmidhuber, 1997], and more recently has been supported by empirical studies showing that flatness-based measures often outperform alternatives such as norm-, margin- or optimization-based metrics [Jiang et al., 2020, Keskar et al., 2017]. However, classical flatness measures, typically based on the Hessian of the loss with respect to parameters, are known to be sensitive to reparameterizations that leave the function and generalization behavior unchanged [Dinh et al., 2017] and also very cumbersome to compute for large state-of-the-art models. To resolve this, Petzka et al. [2021] introduced a reparameterization-invariant measure of *relative flatness*, grounded in a theoretical framework that connects flatness of an individual layer to generalization under assumptions of representative data and locally constant labels. This formulation not only explains prior empirical correlations, but also provides theoretically grounded conditions under which flatness predicts generalization. Walter et al. [2024] further made this measure more practical by deriving closed-form computation formulas in standard classification settings. In our work, we empirically verify that relative flatness drops precisely when generalization emerges during grokking. Moreover, we demonstrate that minimizing relative flatness can actively

induce generalization even in the absence of NC. In doing so, we confirm and extend the theoretical framework of Petzka et al. [2021], while disentangling flatness from NC.

**Grokking as a Window into Generalization** *Grokking* refers to the sudden emergence of generalization long after perfect training performance is achieved, as first described by Power et al. [2022]. This phase transition, which is often induced by implicit or explicit regularization, offers a unique opportunity to isolate potential causes for the emergence of generalization. Recent work explains grokking through two-phase dynamics: an initial kernel regime focused on memorization, followed by feature learning that supports generalization [Mohamadi et al., 2023, Lyu et al., 2024, Kumar et al., 2024]. Analytical studies reveal closed-form grokking solutions in modular arithmetic tasks [Gromov, 2024, Žunkovič and Ilievski, 2022], while mechanistic studies identify sparse and dense subnetworks competing during training [Nanda et al., 2023, Gouki et al., 2024]. Most of these studies focus on shallow networks or toy tasks. However, recent work shows that grokking also arises in deeper architectures and real-world datasets [Murty et al., 2023, Humayun et al., 2024], suggesting greater relevance. We build on this foundation by using grokking as an experimental lens: By tracking the timing of flatness, NC, and generalization, we are able to distinguish which factors are causally connected and which are merely correlated.

### 3 Geometric Foundations of Generalization

In this work, we investigate the interplay between generalization and two notable phenomena in modern deep learning: *neural collapse*, and *flatness* of the loss surface. These concepts are frequently observed to co-occur late in training, but their individual roles and causal relations remain poorly understood. Our goal is to disentangle their contributions to model performance by leveraging the grokking setting, in which memorization precedes generalization, allowing us to observe the emergence of each phenomenon in isolation.

**Generalization** A central objective in machine learning is to ensure that a trained model not only performs well on the training data, but also generalizes to unseen data drawn from the same distribution. This difference in performance is formalized via the *generalization gap*. Let  $f : \mathcal{X} \rightarrow \mathcal{Y}$  be a model from a model class  $\mathcal{F}$  trained with a twice-differentiable loss function  $\ell : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}_+$  on a finite training set  $S \subseteq \mathcal{X} \times \mathcal{Y}$ , drawn i.i.d. from a data distribution  $\mathcal{D}$  over the input space  $\mathcal{X}$  and output space  $\mathcal{Y}$ . The generalization gap is defined as  $\mathcal{E}_{\text{gen}}(f, S) := \mathcal{E}(f) - \mathcal{E}_{\text{emp}}(f, S)$ , where  $\mathcal{E}(f) := \mathbb{E}_{(x,y) \sim \mathcal{D}} [\ell(f(x), y)]$  is the risk and  $\mathcal{E}_{\text{emp}}(f, S) := \frac{1}{|S|} \sum_{(x,y) \in S} \ell(f(x), y)$  is the empirical risk. The model generalizes well when  $\mathcal{E}_{\text{gen}}(f, S)$  is small, indicating close alignment between training and test performance.

**Neural Collapse and the NCC Measure** *Neural collapse* (NC) is a geometric phenomenon characterizing the late stages of training in deep classification networks [Papayan et al., 2020, Mixon et al., 2022]. It manifests as an alignment of the learned representations such that: (1) within each class, features in the penultimate layer collapse to their class mean; (2) the class means themselves form a simplex equiangular tight frame, a maximally spaced configuration; (3) the classifier weights align with these class means; and (4) the classifier effectively becomes a nearest-neighbor model.

This structure has been theoretically motivated and empirically observed in both training and unseen test samples, including new classes, particularly in transfer learning settings [Galanti et al., 2021]. To quantify the emergence of NC, we adopt the simplified Neural Collapse Clustering (NCC) measure proposed by Galanti et al. [2021], which captures the essential characteristics of neural collapse, i.e., tight intra-class clustering and strong inter-class separation, without requiring all four formal NC conditions. This is the measure we use throughout our analysis.

**Definition 3.1** (NCC Measure). Let  $\phi(x)$  denote the penultimate-layer representation of input  $x$ , and  $D_c$  the set of training samples from class  $c$ . Then

$$\text{NCC} := \sum_{c \neq c'} \frac{V_c + V_{c'}}{2 \|\mu_c - \mu_{c'}\|^2},$$

where  $\mu_c := \frac{1}{|D_c|} \sum_{x \in D_c} \phi(x)$ ,  $V_c = \sum_{x \in D_c} \|\phi(x) - \mu_c\|^2$  are the mean, respectively the variance of representations of class  $c$ .

A low NCC value indicates tight intra-class clustering and well-separated class means, characteristic of NC. In Section 5, we confirm that the NCC measure correlates with the angular separation of means, thus capturing all the characteristics of NC.

**Relative Flatness** Flatness of the loss landscape, understood as the insensitivity of the training loss to small perturbations in the parameter space, has long been linked to generalization research [Hochreiter and Schmidhuber, 1997, Keskar et al., 2017]. However, classical measures based on the Hessian or curvature are sensitive to reparameterizations, making them unreliable in modern architectures. To overcome this, Petzka et al. [2021] introduced a reparameterization-invariant notion of *relative flatness*, grounded in a theory of robust generalization under locally constant labels.

Consider a model that decomposes as  $f(x, \mathbf{w}) = g(\mathbf{w}\phi(x))$ , where  $\phi$  is a fixed feature map (e.g., the penultimate layer),  $\mathbf{w} \in \mathbb{R}^{d \times m}$  is a weight matrix, and  $g$  is a twice-differentiable function. The relative flatness is defined via the trace-weighted Hessian along the directions spanned by  $\mathbf{w}$ .

**Definition 3.2** (Relative Flatness). Let  $g, \ell$  be twice differentiable, and  $S$  a sample set. Then:

$$\kappa_{\text{Tr}}^{\phi}(\mathbf{w}) := \sum_{s, s'=1}^d \langle \mathbf{w}_s, \mathbf{w}_{s'} \rangle \cdot \text{Tr}(H_{s, s'}(\mathbf{w}, \phi(S))),$$

where  $\mathbf{w}_s$  denotes the  $s$ -th row of  $\mathbf{w}$ ,  $\langle \cdot, \cdot \rangle$  is the scalar product, and  $H_{s, s'}(\mathbf{w}, \phi(S))$  is the Hessian of the empirical loss with respect to  $\mathbf{w}_s$  and  $\mathbf{w}_{s'}$  evaluated at  $\phi(S)$ .

This measure is invariant to neuron-wise rescaling and composition with orthogonal transformations, making it robust across different parameterizations. Under mild assumptions—namely, that the training data is representative and the labels are locally constant in feature space—relative flatness predicts generalization. Note that a small value of  $\kappa_{\text{Tr}}^{\phi}(\mathbf{w})$  is indicative of a flat solution and good generalization. In Section 5, we further show that NC implies low relative flatness, establishing a theoretical dependency between the two phenomena.

**On the Role of Representativeness** Both neural collapse and relative flatness have been proposed as geometric signatures of generalization. However, neither property is meaningful in isolation: their predictive value depends critically on the relationship between training and test distributions in feature space. In particular, if the training features are not representative of the full data distribution, then both NC and flatness can emerge without leading to generalization. For example, consider a network that perfectly collapses the training set into a class-wise simplex configuration and assigns test samples to incorrect cluster means. (On discrete data, such a network function can be easily constructed.) Prediction will then fail despite the presence of collapse. This limitation also holds for relative flatness: a flat but non-representative model does not generalize [cf. Sec. 2, Petzka et al., 2021]. While NC lacks a rigorous definition of representativeness, relative flatness is grounded in a formal framework that makes the required assumptions explicit. In particular, the theory of  $\varepsilon$ -representativeness [Petzka et al., 2021] defines a condition under which robustness in feature space becomes necessary for generalization. Under this assumption, relative flatness serves as an upper bound on feature robustness and provides a tractable diagnostic for generalization. This perspective clarifies that our analysis operates under the assumption of representative features—an assumption that is empirically supported in our settings, as we demonstrate in Appendix D.

## 4 Neural Collapse and Flatness in Delayed Generalization

Grokking refers to the surprising phenomenon in which a neural network, after memorizing the training set with zero generalization for an extended period, suddenly begins to generalize with high accuracy, often after tens or hundreds of thousands of additional optimization steps [Power et al., 2022]. This temporal separation between memorization and generalization provides a unique opportunity to disentangle the factors that underlie generalization. In our setting, we use grokking, or delayed generalization, not as a curiosity, but as a causal probe: a mechanism for identifying which geometric properties emerge coincidentally with generalization, and which ones may be functionally necessary for it.

First, we study grokking on symbolic algorithmic tasks such as modular arithmetic (e.g.,  $x + y \bmod p$ ), where networks are trained to predict the output of binary operations over abstract tokens.

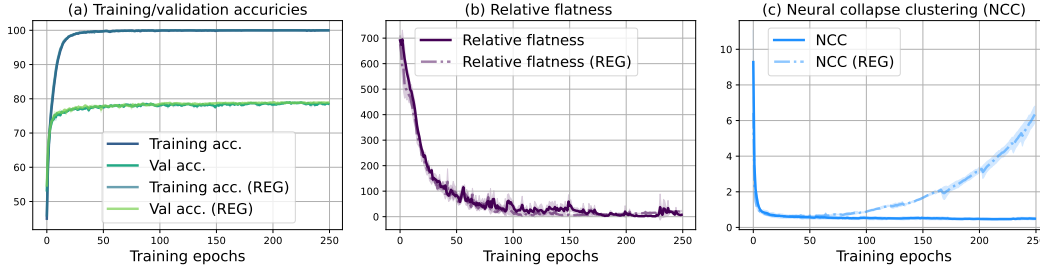


Figure 2: Results of neural collapse clustering regularization on CIFAR-10. We display both unregularized and regularized (REG) training dynamics for comparison. Increasing NCC does not affect generalization or relative flatness, indicating that NC is not necessary for generalization. Figure (a) shows the training and validation accuracies, and y-axis represents accuracy. Figure (b) presents the relative flatness values during training, and y-axis represents measurement of relative flatness. Figure (c) illustrates the NCC development, and y-axis represents NCC value.

These tasks are small, fully observable, and require exact generalization beyond memorized samples, making them an ideal testbed. We follow the original setup of Power et al. [2022], training a 2 layer-transformer using the AdamW optimizer with a learning rate of  $10^{-4}$  and weight decay of 1.0. Each experiment runs for  $10^6$  steps with a 50/50 train/validation split. All results are averaged over three seeds. Figure 1 shows the evolution of train and validation accuracy, as well as relative flatness (Def. 3.2), and neural collapse clustering (Def. 3.1) throughout training. We observe that both measures decrease sharply when the network start generalizing, however, NCC already decreases during the memorization phase. Relative flatness remains high during the memorization phase and decreases sharply near the point at which generalization begins. This temporal coincidence suggests that both NCC and relative flatness are correlated with generalization, but representations start exhibiting collapse behaviour before generalization is satisfying.

However, temporal correlation is not causation. The simultaneous decrease of NCC and flatness may reflect shared dependence on generalization, or confounders, rather than causal influence. The grokking setting thus makes it clear that we must go further: it is not enough to ask whether these quantities track generalization, we must ask whether they are causing it. In the following sections, we address this question empirically. In Section 5, we show that neural networks can generalize without exhibiting collapse. In Section 6, we provide empirical evidence that relative flatness is necessary for generalization under mild assumptions.

## 5 Neural Collapse is not Necessary

Neural collapse (NC) is often observed in models that generalize well, leading to the impression that it may play a functional role in generalization. However, correlation does not imply necessity. In this section, we demonstrate both empirically and theoretically that NC is not required for generalization. While collapse can accompany good generalization, it is neither sufficient nor necessary: it is one path that leads to flatness, and through flatness, to generalization.

To substantiate this claim, we explicitly suppress the emergence of NC during training without affecting generalization. We train a ResNet-18 on CIFAR-10 using a regularized loss of the form:

$$\mathcal{L}_{\text{NC-REG}} = \mathcal{L}_{\text{CE}} - \lambda \cdot \text{NCC},$$

where  $\mathcal{L}_{\text{CE}}$  is the standard cross-entropy loss and  $\text{NCC}$  is the NC clustering measure (Def. 3.1). The NCC term penalizes tight clustering of penultimate-layer features by

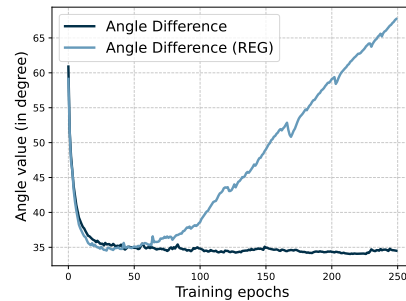


Figure 3: Pairwise cluster angles vs. optimal 10-simplex angles. Under NCC regularization, angles drift from the optimal configuration, while in standard training they remain stable. "REG" indicates use of the NCC regularizer.

shrinking the distances between class means and increasing intra-class variance, thereby actively discouraging the geometric structure characteristic of NC. The training hyperparameter setups and additional experimental details are in Appendix B.

Figure 2 presents the results. As shown in panel (a), both training and validation accuracies remain unaffected by the regularizer. Panel (b) shows that relative flatness, our proxy for generalization, also remains stable. However, panel (c) reveals that NC is effectively suppressed: after a brief drop in the NCC measure, it steadily increases throughout training. This provides clear evidence that generalization can occur in the complete absence of NC, and that collapse is not a necessary condition for generalization. In Figure 3 we confirm that the NCC measure correctly captures cluster angles as well. Under neural collapse, angles are close to the optimal 10-simplex, yet when we regularize with the NCC measure, angles deviate strongly from this optimum.

Why, then, does neural collapse often accompany generalization? We argue that collapse is one geometric pathway to flat solutions, which are more directly linked to generalization. In fact, under mild assumptions, NC in the penultimate layer implies relative flatness<sup>1</sup>. This relationship is formalized below.

We assume that the classical neural collapse [Papayan et al., 2020] in the limit holds for a network function  $f$ . That is, the following four conditions apply to the neural network  $f(x) = \text{softmax}(w\phi(x) + b)$  with  $\phi(x)$  as the representation of the penultimate layer.

**Assumptions 5.1** (Neural Collapse [Papayan et al., 2020]). Let  $\phi(x)$  denote the penultimate layer feature representation of an input  $x \in \mathcal{X}$ , and for a dataset  $D \subset \mathcal{X}$ , let  $D_c \subseteq D$  denote the set of inputs belonging to class  $c \in \{1, \dots, k\}$ . Then the four neural collapse criteria are

- (NC1) Feature representations collapse to class means:  $\phi(x) = \mu_c$  for all  $x \in D_c$ .
- (NC2) All class means  $\mu_c$  have equal distance to the global mean  $\mu_g$ , i.e.,  $\|\mu_c - \mu_g\|_2 = M$ , and they form a centered equiangular tight frame (ETF), so that  $(\mu_j - \mu_g)^\top (\mu_c - \mu_g) = -\frac{M^2}{k-1}$  for  $j \neq c$ .
- (NC3) The classifier weights align with class means, i.e.  $w_j = \lambda(\mu_j - \mu_g)$  for all  $j$  and some  $\lambda > 0$ .
- (NC4)  $\text{argmax}_j w_j^\top, h + b_j = \text{argmin}_j \|h - \mu_j\|$ .

**Proposition 5.2.** Let  $f(x) = \text{softmax}(w\phi(x) + b)$  be a neural network with softmax output and trained with cross-entropy loss, where  $w \in \mathbb{R}^{C \times d}$  denotes the final-layer weight matrix classifying into  $k$  classes and  $\phi(x)$  is the penultimate-layer representation. Assume that the classical neural collapse holds in the limit as above. Then relative flatness  $\kappa_\phi(w)$  is bounded by:

$$\kappa_\phi(w) \leq \lambda^2 k^3 M^4 \cdot \frac{e^{-\lambda M^2 \cdot \frac{k}{k-1}}}{\left(1 + (k-1)e^{-\lambda M^2 \cdot \frac{k}{k-1}}\right)^2}$$

In particular, for sufficiently large  $\lambda$ , this yields the asymptotic bound:

$$\kappa_\phi(w) \lesssim \lambda^2 k^3 M^4 e^{-\lambda M^2 \cdot \frac{k}{k-1}},$$

which decays exponentially in  $\lambda$ .

The proof is provided in Appendix A. The fact that relative flatness indeed decreases to zero in the limit follows from the fact that the conditions of neural collapse are independent of the weight norm and thereby also the magnitude of the scalar  $\lambda$ . As training continues in the NC limit, the training loss can be decreased to zero by increasing  $\lambda$ , because the softmax probabilities converge to  $\hat{y}_c(x_c) = 1$  and  $\hat{y}_j(x_c) = 0, j \neq c$  as  $\lambda \rightarrow \infty$ .

Taken together, our results demonstrate that, while NC can facilitate generalization by inducing flatness, it is not a prerequisite. Flatness remains the more fundamental quantity, and its relevance

<sup>1</sup>Related works often rely on simplifying assumptions, e.g., one-layer behavior or layer peeling [Mixon et al., 2022, Fang et al., 2021], yet our analysis does not require such approximations.

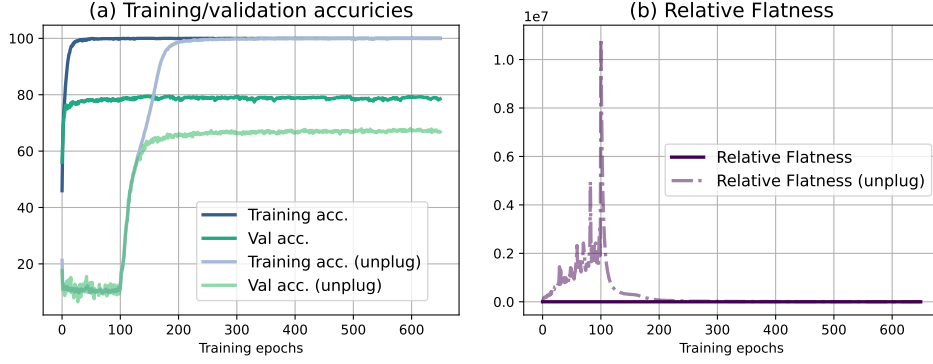


Figure 4: Results of inducing delayed generalization on CIFAR-10 through relative flatness regularization. Delayed generalization occurs immediately after the relative flatness regularizer is removed, as indicated by sharp increases in validation accuracy. Figure (a) shows the training and validation accuracies on CIFAR-10, while Figure (b) presents the measure of relative flatness. “Unplug” indicates that the relative flatness regularizer is applied during training and removed at a specific training epoch.

critically depends on the representativeness of the learned feature space. It remains to investigate whether flatness is a sufficient condition or whether it could be necessary.

## 6 Relative Flatness is Necessary

In this section, we induce delayed generalization, similar to what is observed in grokking. These experiments demonstrate that relative flatness is necessary for generalization. To achieve such behaviour, we design a loss function that includes the following regularization term:

$$\mathcal{L}_{\text{NC-RF}} = \mathcal{L}_{\text{CE}} - \lambda * \kappa_{T_r}^{\phi}(\mathbf{w}) \quad (1)$$

where  $\kappa_{T_r}^{\phi}(\mathbf{w})$  is defined in Definition 3.2, and  $\mathcal{L}_{\text{CE}}$  denotes the cross-entropy loss. By maximizing the relative flatness, we encourage the network to explore sharper regions of the loss landscape rather than flatter ones, which leads to suboptimal generalization. The coefficient  $\lambda$  controls the strength of the regularization. We train ResNet-10 on CIFAR-10 from scratch to evaluate the effectiveness of inducing delayed generalization in general settings by regularizing relative flatness. The results are shown in Figure 4. The training hyperparameter setups and additional experimental details are in Appendix C.

From Figure 4 (a), we observe that the validation accuracy increases sharply after the relative flatness regularizer is removed at epoch 100. Correspondingly, in Figure 4 (b), the measure of relative flatness is incredibly high under the regularization and drops sharply after the regularizer is removed. By removing relative flatness, we can induce delayed generalization in more general tasks and training settings.

In particular, when the relative flatness regularizer is removed, the model still converges to nearly 100% training accuracy but suffers a marked drop in validation performance, plateauing around 67%. In effect, the regularizer steers the optimization toward a suboptimal region of the loss landscape, such that once it is unplugged, the model needs to recover from an unfavorable initialization. This behavior resembles restarting training from a poor initialization: the final model remains reasonably effective but clearly suboptimal. We interpret this phenomenon as further evidence that flatness should arise naturally as part of a balanced optimization process rather than being imposed as a late-stage correction. While additional techniques such as careful learning rate scheduling might partially mitigate the degradation observed after un-plugging, they would not alter the core insight established by these experiments.

Overall, we strongly support our argument that *relative flatness is necessary for generalization* by both empirical evidence and consistent experimental outcomes.



## 7 Discussion and Conclusion

Our results identify relative flatness as a potentially necessary condition for generalization in deep networks, while neural collapse emerges as a correlated but non-essential feature of late-stage training. This distinction reshapes our understanding of the geometry of generalization and invites several important reflections. Our experiments show that NC tends to coincide with high training accuracy but does not reliably track generalization. This suggests that it is more a byproduct of training dynamics than a causal mechanism, a view that aligns with findings from Hui et al. [2022] and Hong and Ling [2024]. Nevertheless, we show that under mild assumptions, neural collapse-style clustering implies relative flatness. As such, collapse may still serve as a practical proxy for flat solutions in some settings.

While NC produces highly symmetric and compact feature representations, this structural simplicity can be problematic in transfer learning and out-of-distribution (OOD) scenarios. For instance, if a network is trained to classify supercategories by collapsing fine-grained class distinctions, it becomes impossible to recover those distinctions later, thus limiting the generality and interpretability of the learned features [Hui et al., 2022].

Flatness itself has also come under scrutiny, particularly in the context of sharpness-aware minimization (SAM) [Foret et al., 2021] and recent critiques by Andriushchenko et al. [2023]. However, many such criticisms target classical flatness metrics, which are sensitive to reparameterization. Our analysis relies on relative flatness, a reparameterization-invariant measure with theoretical guarantees [Petzka et al., 2021, Walter et al., 2024]. While it remains an open question how these recent critiques translate to relative flatness, our findings suggest that it continues to offer a reliable signal for generalization in the settings we consider.

The theoretical link between relative flatness and generalization depends on two key assumptions: that labels are locally constant and that the features at the penultimate layer are representative of the data distribution. In our settings (standard image classification and modular arithmetic tasks) both assumptions are well justified: (i) the assumption of locally constant labels underlies the very notion of adversarial robustness, and its empirical validity is supported by our experimental findings, and (ii) we show in Appendix D that representativeness improves with the onset of generalization. Nonetheless, these assumptions may not hold in other domains such as structured prediction, high-noise regression, or complex real-world tasks, where label semantics may vary significantly under small input changes.

One of our most surprising results is that regularizing networks away from flat minima consistently induces delayed generalization. This underscores the functional importance of flatness, not just its correlation with generalization. It also opens new directions for controlling training dynamics via geometric regularization. However, the extent to which these artificially induced grokking phases mirror the mechanisms of standard grokking remains to be understood.

Finally, while penalizing flatness reliably delays generalization, encouraging flatter minima does not substantially improve it [Adilova et al., 2023]. Even SAM, which improves training in vision tasks, surprisingly does not actually encourage flatter minima [Andriushchenko and Flammarion, 2022]. This asymmetry supports the long-held view that stochastic gradient descent (SGD) already biases toward flat minima [Hochreiter and Schmidhuber, 1997, Keskar et al., 2017], and that relative flatness may serve more effectively as a diagnostic and interpretive tool rather than a universal optimization target.

While our findings are robust across datasets and architectures, our conclusions are necessarily limited in scope. We focus exclusively on classification networks trained with standard objectives and optimizers; whether our results extend to generative models, contrastive pretraining, or large-scale language models remains an open question.

We challenge the assumption that neural collapse is essential for generalization. Through grokking and targeted regularization, we isolate flatness as the necessary factor. Our theoretical and empirical results jointly establish flatness as the geometric driver of generalization.

## References

- Linara Adilova, Amr Abourayya, Jianing Li, Amin Dada, Henning Petzka, Jan Egger, Jens Kleesiek, and Michael Kamp. Fam: Relative flatness aware minimization. In *Topological, Algebraic and Geometric Learning Workshops 2023*, pages 37–49. PMLR, 2023.
- Maksym Andriushchenko and Nicolas Flammarion. Towards understanding sharpness-aware minimization. In *International conference on machine learning*, pages 639–668. PMLR, 2022.
- Maksym Andriushchenko, Francesco Croce, Maximilian Müller, Matthias Hein, and Nicolas Flammarion. A modern look at the relationship between sharpness and generalization. In *International Conference on Machine Learning*, pages 840–902. PMLR, 2023.
- Hien Dang, Tho Tran Huu, Stanley Osher, Nhat Ho, Tan Minh Nguyen, et al. Neural collapse in deep linear networks: From balanced to imbalanced data. In *International Conference on Machine Learning*, pages 6873–6947. PMLR, 2023.
- Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. In *International Conference on Machine Learning*, pages 1019–1028. PMLR, 2017.
- 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.
- Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*, 2021.
- Tomer Galanti, András György, and Marcus Hutter. On the role of neural collapse in transfer learning. In *International Conference on Learning Representations*, 2021.
- Minegishi Gouki, Yusuke Iwasawa, and Yutaka Matsuo. Grokking tickets: Lottery tickets accelerate grokking, 2024. URL <https://openreview.net/forum?id=WSsP7W8tqN>.
- Florian Graf, Christoph Hofer, Marc Niethammer, and Roland Kwitt. Dissecting supervised contrastive learning. In *International Conference on Machine Learning*, pages 3821–3830. PMLR, 2021.
- Andrey Gromov. A simple and interpretable model of grokking modular arithmetic tasks, 2024. URL <https://openreview.net/forum?id=0ZUKLCxwBo>.
- XY Han, Vardan Papayan, and David L Donoho. Neural collapse under mse loss: Proximity to and dynamics on the central path. In *International Conference on Learning Representations*, 2021.
- Sepp Hochreiter and Jürgen Schmidhuber. Flat minima. *Neural computation*, 9(1):1–42, 1997.
- Wanli Hong and Shuyang Ling. Beyond unconstrained features: Neural collapse for shallow neural networks with general data. *arXiv preprint arXiv:2409.01832*, 2024.
- Like Hui, Mikhail Belkin, and Preetum Nakkiran. Limitations of neural collapse for understanding generalization in deep learning. *arXiv preprint arXiv:2202.08384*, 2022.
- Ahmed Imtiaz Humayun, Randall Balestriero, and Richard Baraniuk. Deep networks always grok and here is why, 2024.
- Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. Fantastic generalization measures and where to find them. In *Proceedings of the International Conference on Learning Representations*, 2020.
- Nitish Shirish Kesar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. In *Proceedings of the International Conference on Learning Representations*, 2017.
- Tanishq Kumar, Blake Bordelon, Samuel J. Gershman, and Cengiz Pehlevan. Grokking as the transition from lazy to rich training dynamics. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=vt5mnLVIVo>.

- Kaifeng Lyu, Jikai Jin, Zhiyuan Li, Simon Shaolei Du, Jason D. Lee, and Wei Hu. Dichotomy of early and late phase implicit biases can provably induce grokking. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=XsHqr9dEGH>.
- Dustin G Mixon, Hans Parshall, and Jianzong Pi. Neural collapse with unconstrained features. *Sampling Theory, Signal Processing, and Data Analysis*, 20:11, 2022.
- Mohamad Amin Mohamadi, Zhiyuan Li, Lei Wu, and Danica Sutherland. Grokking modular arithmetic can be explained by margin maximization. In *NeurIPS 2023 Workshop on Mathematics of Modern Machine Learning*, 2023. URL <https://openreview.net/forum?id=QPMfCLnIqf>.
- Michael Munn, Benoit Dherin, and Javier Gonzalvo. The impact of geometric complexity on neural collapse in transfer learning. In *Advances in Neural Information Processing Systems*, 2024.
- Shikhar Murty, Pratyusha Sharma, Jacob Andreas, and Christopher Manning. Grokking of hierarchical structure in vanilla transformers. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 439–448, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-short.38. URL <https://aclanthology.org/2023.acl-short.38>.
- Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. In *The Eleventh International Conference on Learning Representations*, 2023. URL <https://openreview.net/forum?id=9XFSbDPmdW>.
- Vardan Papyan, XY Han, and David L Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. In *Proceedings of the National Academy of Sciences*, volume 117, pages 24652–24663. National Academy of Sciences, 2020.
- Henning Petzka, Michael Kamp, Linara Adilova, Cristian Sminchisescu, and Mario Boley. Relative flatness and generalization. *Advances in neural information processing systems*, 34:18420–18432, 2021.
- Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets. *arXiv preprint arXiv:2201.02177*, 2022.
- 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. IEEE, 2022.
- Peter Šúkeník, Marco Mondelli, and Christoph H Lampert. Deep neural collapse is provably optimal for the deep unconstrained features model. *Advances in Neural Information Processing Systems*, 36:52991–53024, 2023.
- Nils Philipp Walter, Linara Adilova, Jilles Vreeken, and Michael Kamp. The uncanny valley: Exploring adversarial robustness from a flatness perspective. *arXiv preprint arXiv:2405.16918*, 2024.
- Diyuan Wu and Marco Mondelli. Neural collapse beyond the unconstrained features model: Landscape, dynamics, and generalization in the mean-field regime. *arXiv preprint arXiv:2501.19104*, 2025.
- Robert Wu and Vardan Papyan. Linguistic collapse: Neural collapse in (large) language models. *Advances in Neural Information Processing Systems*, 37:137432–137473, 2024.
- Yingwen Wu, Ruiji Yu, Xinwen Cheng, Zhengbao He, and Xiaolin Huang. Pursuing feature separation based on neural collapse for out-of-distribution detection. In *The Thirteenth International Conference on Learning Representations*, 2025.
- 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? *Advances in neural information processing systems*, 35:37991–38002, 2022.

- Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2016.
- Jinxin Zhou, Chong You, Xiao Li, Kangning Liu, Sheng Liu, Qing Qu, and Zhihui Zhu. Are all losses created equal: A neural collapse perspective. In *Advances in Neural Information Processing Systems*, volume 35, pages 31697–31710, 2022.
- Jinxin Zhou, Jiachen Jiang, and Zhihui Zhu. Are all layers created equal: A neural collapse perspective. In *The Second Conference on Parsimony and Learning*, 2025.
- Zhihui Zhu, Tianyu Ding, Jinxin Zhou, Xiao Li, Chong You, Jeremias Sulam, and Qing Qu. A geometric analysis of neural collapse with unconstrained features. *Advances in Neural Information Processing Systems*, 34:29820–29834, 2021.
- Bojan Žunkovič and Enej Ilievski. Grokking phase transitions in learning local rules with gradient descent, 2022.

## A Proof of Proposition 5.2

First, we provide a technical lemma that establishes that, under Neural Collapse, the combined contribution of the bias and the projection of the class mean onto the global mean is identical across classes. This means that the logits are effectively governed solely by the geometric structure of the class means relative to each other.

**Lemma A.1.** *Suppose that the neural collapse conditions (NC1) – (NC4) in the limit apply as in Section 5 (see also below). Then there is a constant  $d$  such that  $b_j + \lambda(\mu_j - \mu_g)^\top \mu_g = d$ .*

*Proof.*

$$\begin{aligned}
\operatorname{argmin}_j \|\phi(x) - \mu_j\| &= \operatorname{argmin}_j \|(\phi(x) - \mu_g) - (\mu_j - \mu_g)\|^2 \\
&= \operatorname{argmin}_j \|\phi(x) - \mu_g\|^2 + \|\mu_j - \mu_g\|^2 - 2(\phi(x) - \mu_g)^\top (\mu_j - \mu_g) \\
&\stackrel{(NC1)}{=} \operatorname{argmin}_j 2M^2 - 2(\phi(x) - \mu_g)^\top (\mu_j - \mu_g) \\
&= \operatorname{argmax}_j \phi(x)^\top (\mu_j - \mu_g) - \mu_g^\top (\mu_j - \mu_g) \\
&\stackrel{(NC3)}{=} \operatorname{argmax}_j \frac{1}{\lambda} \phi(x)^\top w_j - \frac{1}{\lambda} \lambda \mu_g^\top (\mu_j - \mu_g) \\
&= \operatorname{argmax}_j \phi(x)^\top w_j + b_j - (b_j + \lambda \mu_g^\top (\mu_j - \mu_g)) \\
&\stackrel{(NC4)}{=} \operatorname{argmin}_j \|\phi(x) - \mu_j\| - (b_j + \lambda \mu_g^\top (\mu_j - \mu_g))
\end{aligned}$$

But  $\operatorname{argmin}_j \|\phi(x) - \mu_j\| = \operatorname{argmin}_j \|\phi(x) - \mu_j\| - (b_j + \lambda \mu_g^\top (\mu_j - \mu_g))$  for all  $\phi(x)$  can only hold true if  $(b_j + \frac{1}{\lambda} \lambda \mu_g^\top (\mu_j - \mu_g))$  is constant over  $j$ .  $\square$

We now provide a second technical lemma showing that relative flatness is upper bounded by the simplified version  $\|w\|^2 \operatorname{Tr}(H(w))$ , for which we have a closed form expression in terms of logits.

**Lemma A.2.** *Let  $H(w)$  be the Hessian wrt. the penultimate layer weights  $w \in \mathbb{R}^{d \times m}$  and  $\kappa_\phi(w)$  the relative flatness measure. Then*

$$\kappa_\phi(w) \leq \|w\|^2 \operatorname{Tr}(H(w)) .$$

*Proof.* Recall that

$$\kappa_\phi(w) = \sum_{s, s'} \langle w_s, w_{s'} \rangle \operatorname{Tr}(H_{s, s'}) .$$

Applying the Cauchy-Schwarz inequality to each summand yields

$$\langle w_s, w_{s'} \rangle \operatorname{Tr}(H_{s, s'}) \leq \|w_s\| \|w_{s'}\| \sqrt{\operatorname{Tr}(H_{s, s})} \sqrt{\operatorname{Tr}(H_{s', s'})} .$$

Summing over all  $s, s'$  and using symmetrie as well as Cauchy-Schwarz again we get

$$\kappa_\phi(w) \leq \left( \sum_s \|w_s\|^2 \right) \left( \sum_s \operatorname{Tr}(H_{s, s}) \right) = \|w\|^2 \operatorname{Tr}(H(w)) .$$

$\square$

With this we can proof the theorem, which we restate together with the neural collapse criteria for convenience.

**Assumptions A.3** (Neural Collapse [Papayan et al., 2020]). Let  $\phi(x)$  denote the penultimate layer feature representation of an input  $x \in \mathcal{X}$ , and for a dataset  $D \subset \mathcal{X}$ , let  $D_c \subseteq D$  denote the set of inputs belonging to class  $c \in \{1, \dots, k\}$ . Then the four neural collapse criteria are

(NC1) Feature representations collapse to class means:  $\phi(x) = \mu_c$  for all  $x \in D_c$ .

(NC2) All class means  $\mu_c$  have equal distance to the global mean  $\mu_g$ , i.e.,  $\|\mu_c - \mu_g\|_2 = M$ , and they form a centered equiangular tight frame (ETF), so that  $(\mu_j - \mu_g)^\top (\mu_c - \mu_g) = -\frac{M^2}{k-1}$  for  $j \neq c$ .

(NC3) The classifier weights align with class means, i.e.  $w_j = \lambda(\mu_j - \mu_g)$  for all  $j$  and some  $\lambda > 0$ .

(NC4)  $\operatorname{argmax}_j w_j^\top, h + b_j = \operatorname{argmin}_j \|h - \mu_j\|$ .

**Proposition A.4.** *Let  $f(x) = \operatorname{softmax}(w\phi(x) + b)$  be a neural network with softmax output and trained with cross-entropy loss, where  $w \in \mathbb{R}^{C \times d}$  denotes the final-layer weight matrix classifying into  $k$  classes and  $\phi(x)$  is the penultimate-layer representation. Assume that the classical neural collapse holds in the limit as above. Then relative flatness  $\kappa_\phi(w)$  is bounded by:*

$$\kappa_\phi(w) \leq \lambda^2 k^3 M^4 \cdot \frac{e^{-\lambda M^2 \cdot \frac{k}{k-1}}}{\left(1 + (k-1)e^{-\lambda M^2 \cdot \frac{k}{k-1}}\right)^2}$$

In particular, for sufficiently large  $\lambda$ , this yields the asymptotic bound:

$$\kappa_\phi(w) \lesssim \lambda^2 k^3 M^4 e^{-\lambda M^2 \cdot \frac{k}{k-1}},$$

which decays exponentially in  $\lambda$ .

*Proof.* Lemma A.1 shows that the NC conditions imply that there is a constant  $d$  such that  $b_j + \lambda(\mu_j - \mu_g)^\top \mu_g = d$  for all class labels  $j = 1, \dots, k$ . For  $x_c \in D_c$ , the logits of  $f$  are

$$\begin{aligned} w_j^\top \phi(x_c) + b_j &= \lambda(\mu_j - \mu_g)^\top \mu_c + b_j \\ &= \lambda(\mu_j - \mu_g)^\top (\mu_c - \mu_g) + \underbrace{\lambda(\mu_j - \mu_g)^\top \mu_g + b_j}_{=d} = \lambda(\mu_j - \mu_g)^\top (\mu_c - \mu_g) + d, \end{aligned}$$

and the softmax probabilities become:

$$\hat{y}_c(x_c) = \frac{1}{1 + (k-1)e^{-\lambda\delta}}, \quad \hat{y}_j(x_c) = \frac{e^{-\lambda\delta}}{1 + (k-1)e^{-\lambda\delta}}, \quad j \neq c,$$

with margin

$$\delta := M^2 - \left(-\frac{M^2}{k-1}\right) = M^2 \cdot \frac{k}{k-1}.$$

The trace of the Hessian for input  $x_c \in D_c$  is:

$$\begin{aligned} \operatorname{Tr}(H(x_c, w)) &= \sum_{j=1}^k \hat{y}_j(x_c)(1 - \hat{y}_j(x_c)) \cdot \|\phi(x_c)\|^2 = M^2 \cdot \sum_{j=1}^k \hat{y}_j(x_c)(1 - \hat{y}_j(x_c)). \\ &= M^2 \frac{(k-1)e^{-\lambda\delta} \cdot (2 + (k-2)e^{-\lambda\delta})}{(1 + (k-1)e^{-\lambda\delta})^2}. \end{aligned}$$

Thus,

$$\operatorname{Tr}(H(w)) \leq M^2 \cdot \frac{(k-1)ke^{-\lambda\delta}}{(1 + (k-1)e^{-\lambda\delta})^2}.$$

Since  $\|w\|^2 = \lambda^2 \sum_j \|\mu_j - \mu_g\|^2 = \lambda^2 k M^2$ , we obtain with Lemma A.2:

$$\kappa_\phi(w) \leq \lambda^2 k M^2 \cdot M^2 \cdot \frac{(k-1)ke^{-\lambda\delta}}{(1 + (k-1)e^{-\lambda\delta})^2} \leq \lambda^2 k^3 M^4 \cdot \frac{e^{-\lambda M^2 \cdot \frac{k}{k-1}}}{\left(1 + (k-1)e^{-\lambda M^2 \cdot \frac{k}{k-1}}\right)^2}.$$

This yields the desired bound, and for large  $\lambda$ , the denominator tends to 1, giving the asymptotic behavior.  $\square$

## B Training Setups for NCC and Ablation Studies

We evaluate our method on the CIFAR-10 dataset using the standard training (50,000 samples) and test (10,000 samples) splits. Input images are normalized using the channel-wise mean and standard deviation of (0.5, 0.5, 0.5) and (0.5, 0.5, 0.5), respectively, as implemented in `transforms.Normalize`. No data augmentation is applied in the main experiments.

All models are trained using stochastic gradient descent (SGD) with a fixed learning rate of 0.01, a batch size of 64, and no weight decay. The regularization coefficient is set to  $\lambda = 10^{-3}$ . Momentum is 0.9. Training is performed for 250 epochs without the use of a learning rate scheduler or early stopping.

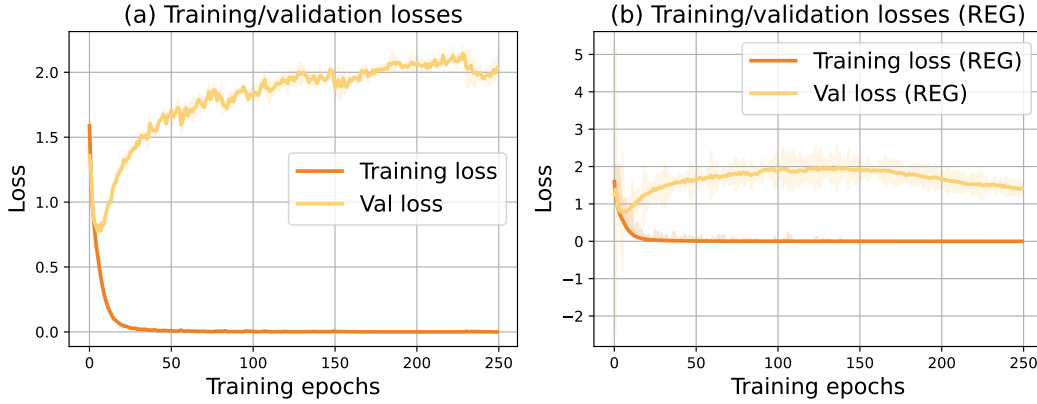


Figure 5: Results of training and validation losses in the NCC experiments.

We use the ResNet-18 architecture as implemented in `torchvision` version 0.19.1, without any modifications. In particular, we retain the original  $7 \times 7$  kernel size in the first convolutional layer and the following max pooling layer, despite the smaller resolution of CIFAR-10 images.

All results in the main experiments are reported as the average over three independent runs with random seeds 1, 42, and 15213 to account for variability due to random initialization. Experiments are conducted using PyTorch 2.4.1 on a single NVIDIA A100 GPU with 80GB of memory. Training each model for 250 epochs takes approximately 10.5 hours.

The loss corresponding to Figure 2 is presented in Figure 5.

**Effect of Different  $\lambda$  Values of NCC Regularizer** In our main experiments, we apply  $\lambda = 1 \times 10^{-3}$  using three random seeds. To study the effect of different  $\lambda$  values, we conduct an ablation study using a range of  $\lambda = \{1 \times 10^{-2}, 1 \times 10^{-3}, 1 \times 10^{-4}, 1 \times 10^{-5}, 1 \times 10^{-6}\}$ , evaluated with Seed 42 due to computational constraints. We have verified that results with a single seed remain consistent with the trends observed using multiple seeds.

The results are shown in Figure 6, which plots accuracy, loss, relative flatness, and neural collapse clustering (NCC) metrics. We observe that a large  $\lambda$  value (e.g.,  $1 \times 10^{-2}$ ) causes training instability: although training accuracy reaches nearly 100% at the end of training, validation accuracy remains below 60%, and both training and validation accuracies collapse to around 10% around epoch 70.

Conversely, when  $\lambda$  is smaller than  $1 \times 10^{-3}$ , the NCC regularizer has little to no effect on the training, as evidenced by negligible differences in performance metrics and NCC values compared to the baseline. When  $\lambda$  is set to  $1 \times 10^{-3}$ , the measure of relative flatness remains low while the NCC metric increases. This supports our argument that neural collapse clustering is not necessary for generalization, as good generalization performance can still be achieved in the absence of strong NCC behavior.

**Effect of Data Augmentation in NCC** To investigate the effect of data augmentation on NCC regularization, we apply standard transformations—*RandomCrop*(32, padding=4) and *RandomHorizon-*

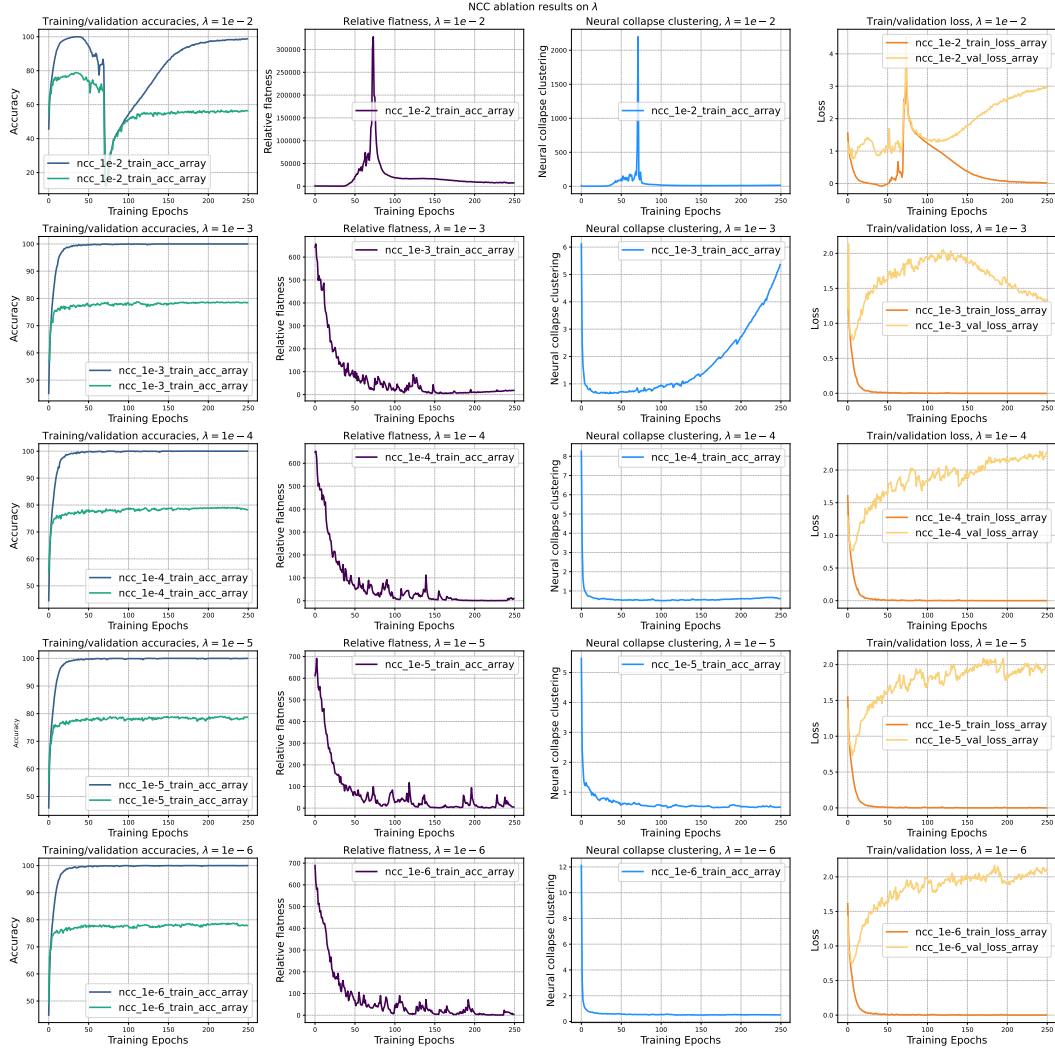


Figure 6: Results of different  $\lambda$  values for the NCC regularizer. Each row contains training/validation accuracies, measure of relative flatness, NCC measurement and training/validation losses for a particular  $\lambda$  value.

*talFlip()*—and conduct experiments using  $\lambda \in \{10^{-2}, 10^{-3}, 10^{-4}\}$ . All inputs are normalized using per-channel means and standard deviations of (0.4914, 0.4822, 0.4465) and (0.2023, 0.1994, 0.2010), respectively.

We limit the range of  $\lambda$  to these three values because prior results indicate that  $\lambda = 10^{-4}$  has negligible effect; smaller values would therefore be redundant. Due to limited resources, we use Seed 42 for this ablation, and based on consistent trends observed across seeds in other experiments, we find this sufficient for reliable comparison.

The results, shown in Figure 7, indicate that with  $\lambda = 10^{-2}$ , training becomes unstable and performance significantly deteriorates. With  $\lambda = 10^{-3}$  and  $10^{-4}$ , data augmentation prevents the NCC metric from increasing as it does in the non-augmented setting—suggesting that NCC regularization is less effective when strong implicit regularization is already applied. In particular,  $\lambda = 10^{-3}$ , which was effective without augmentation (low measure of relative flatness and increased NCC), does not produce the same NCC increase under augmentation. At  $\lambda = 10^{-4}$ , the NCC metric remains low in both augmented and non-augmented cases.



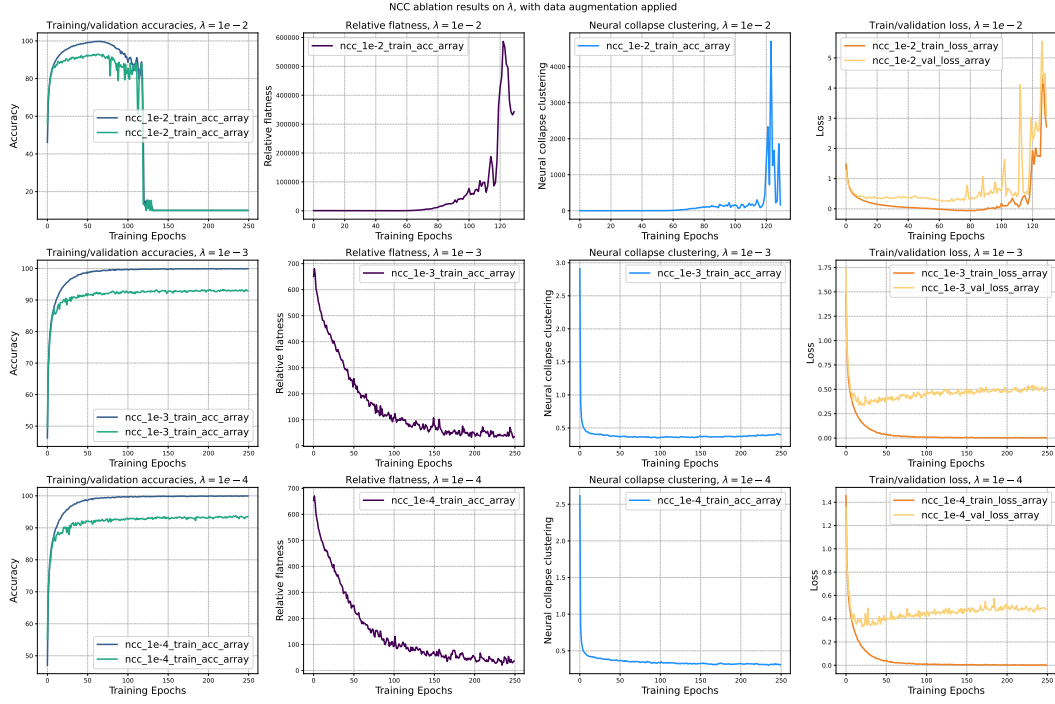


Figure 7: Results of different  $\lambda$  values for the NCC regularizer when data augmentation is applied. Each row contains training/validation accuracies, measure of relative flatness, NCC measurement and training/validation losses for a particular  $\lambda$  value.

In these experiments, we also adopt a ResNet-18 variant tailored for CIFAR-10, following prior work: the first convolutional layer is modified to use a  $3 \times 3$  kernel, and the initial max pooling layer is removed to better suit the smaller input resolution. The model with augmentation achieves higher validation accuracy (94%) compared to the non-augmented case (78%).

## C Training Setups for CIFAR-10

### C.1 CIFAR-10 Training Setups and Ablation Experiments

**Training Setup for Delayed Generalization on CIFAR-10** We use the same training setup described in Section B, with the following modifications: (1) the total number of training epochs is increased to 650; (2) the coefficient of the relative flatness regularizer is set to  $10^{-4}$ . The total training time for 650 epochs is approximately 6 hours. To induce delayed generalization behavior, we remove the relative flatness regularizer after epoch 100. More experimental results will be included.

## D Representativeness

As described in Petzka et al. [2021], representativeness is computed based on the representations from the penultimate layer of the network. Following their procedure, we use kernel density estimation (KDE) to measure representativeness (for further details, please refer to the original paper).

We employ a Gaussian kernel with a bandwidth of 1.0, and assign each training sample a weight of 0.02, chosen empirically. Consistent with the setup of the main experiments (Figure 1), all computations are averaged over three random seeds. The results are presented in Figure 8.

As shown in the results, representativeness decreases to zero when the validation accuracy reaches 100%. Moreover, we observe that improvements in representativeness coincide with the initial rise in validation accuracy during training, providing empirical support for our theoretical findings and arguments.

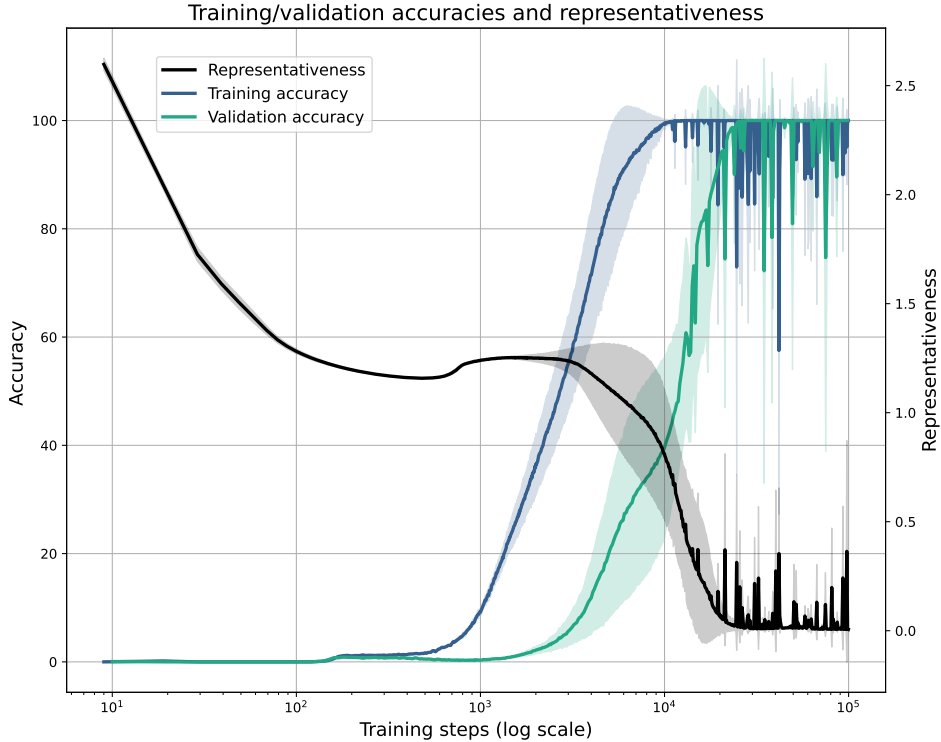


Figure 8: Results of representativeness and training/validation accuracies in grokking.