WHAT DOES NEURAL COLLAPSE INDICATE FOR GENERALIZATION?: A SINGULAR LEARNING THEORY PERSPECTIVE

Anonymous authors
Paper under double-blind review

ABSTRACT

Neural Collapse (NC), which describes the emergence of highly structured class representations in the terminal phase of training, has been widely studied in connection with generalization. However, its practical role remains unclear, as several studies report that NC can hinder generalization in downstream tasks. Existing analyses have mainly focused on NC at the level of representation geometry, leaving unclear what this geometric simplification implies for the complexity of the underlying model. We study this question through Singular Learning Theory (SLT), which quantifies local singular complexity using the finite-temperature empirical Local Learning Coefficient (LLC). In this work, we establish a theoretical bridge between NC and SLT by characterizing local singular complexity through the LLC under exact and approximate NC. Empirically, we validate the predicted stability of local LLC under small deviations from exact-NC and the changes in local loss geometry near exact-NC in Polynomial Neural Networks, and observe similar complexity-related trends across broader architectures. This complexity-based view also offers a new perspective on in-domain generalization under NC. Furthermore, we show that deterministic NC statistics can serve as efficient sampling-free indicators of changes in model complexity during training.

1 INTRODUCTION

Neural Collapse (NC) theory has been proposed as a framework to interpret representation phenomena in deep learning (Papyan et al., 2020). NC has been widely used to analyze generalization. From the perspective of representation learning, for instance, NC can enhance generalization ability through margin maximization (Zhu et al., 2021; Wang et al., 2023). However, empirical findings remain divided. While some studies report that NC contributes to generalization in downstream tasks (Kim & Kim, 2024; Wang et al., 2026; Su et al., 2024), others observe that it can hinder performance without converging to a consistent conclusion (Han et al., 2025; Hui et al., 2022). Due to this ambiguity, applying NC across diverse tasks remains limited despite its demonstrated potential.

To clarify the role of NC in generalization, Singular Learning Theory (SLT), which studies singular models using algebra and geometry (Watanabe, 2007; 2009; Lau et al., 2025; Amari et al., 2006), provides a principled analytical tool. In the NC phenomenon, model complexity is related to its simple geometric structures (Munn et al., 2024). SLT quantifies such geometric model complexity using the real log canonical threshold (RLCT) (Watanabe, 2009) and the local learning coefficient (LLC) (Lau et al., 2025), both grounded in local loss geometry. This perspective motivates an SLT-based analysis of NC and its relation to Bayesian generalization.

In this paper, we establish a theoretical connection between NC and local singular complexity. We characterize the exact-NC RLCT under a local normal form and connect approximate-NC LLC bounds to Bayesian generalization under statistical compatibility. Under additional statistical compatibility assumptions, we connect these results to in-domain Bayesian generalization. For clarity of analysis, we derive these results using Polynomial Neural Networks (PNNs) within a neuroalgebraic geometry framework. Empirically, we validate the resulting finite-temperature stability relation and local singularity mechanism on PNNs. Furthermore, we observe related NC and LLC behavior across

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

diverse standard architectures, including ReLU-MLPs, VGGs, ResNets, and modular arithmetic Transformers.

MAIN CONTRIBUTIONS

Our key contributions are:

- **Theoretical Connection**: We first establish theoretical links of exact-NC states to singularities of SLT by exploiting the analytic structure of PNNs to obtain a local analytic representation.
- **Complexity Bound of ϵ-NC**: We further extend the relation between NC and SLT to incomplete NC state, namely ϵ-NC, formulating a bound of deviations of the empirical loss and finite-temperature LLC from the exact-NC state.
- **Generalization Interpretation**: We provide more quantifiable relation of NC to generalization based on the lens of SLT, proving that the Bayesian Generalization Error of NC model is determined by singularity of the model.
- **Empirical Validation**: We empirically validate the relation between NC and SLT within PNNs, and extend it to broader and practical networks and setups, including ReLU MLP, VGG, ResNet, and Transformer.

1.1 RELATED WORKS

Neural Collapse. Following the observation of NC during terminal-phase training (Papyan et al., 2020), the Unconstrained Feature Model (UFM) (Mixon et al., 2022) enabled Simplex Equiangular Tight Frame (ETF) optimality results for various losses (Zhu et al., 2021), nonlinear settings (Tirer & Bruna, 2022), and multiple hidden layers (Súkeník et al., 2023). We instead study how representation geometry is realized by PNN parameters, without claiming that NC necessarily follows from architecture. Generalization accounts based on SGD's effective-depth bias (Galanti et al., 2022) and geometric complexity (Munn et al., 2024) coexist with downstream failures under strict NC (Hui et al., 2022). Our SLT analysis characterizes in-domain singular complexity; downstream loss geometry requires separate analysis.

Singular Learning Theory. Deep-network singularity (Watanabe, 2007) and phenomena such as double descent (Belkin et al., 2019) motivate alternatives to regular asymptotics. SLT (Watanabe, 2009; 2018) characterizes complexity through the RLCT, complementing flatness-based perspectives (Hochreiter & Schmidhuber, 1997; Keskar et al., 2017) without equating sharpness or rank with RLCT. Localized LLC estimation (Lau et al., 2025) and its association with Transformer phase transitions (Hoogland et al., 2025) motivate our analysis of relation between NC and LLC.

Neuroalgebraic Geometry and Polynomial Neural Networks. PNN capacity (Kileel et al., 2019), optimization landscapes (Shahverdi et al., 2025), subnetwork singularities (Shahverdi et al., 2026), and identifiability (Usevich et al., 2025) are studied within neuroalgebraic geometry (Marchetti et al., 2025). Their polynomial parameterization enables our analytic treatment of NC; non-polynomial architectures remain empirical extensions.

2 BACKGROUND

2.1 SINGULAR LEARNING THEORY

Singular Models and the RLCT. Regular identifiable models have positive-definite Fisher information and locally quadratic loss geometry, yielding the classical complexity coefficient $\lambda = D/2$ for parameter dimension D. Deep neural networks are generally non-identifiable and may exhibit singular loss geometry, for which regular asymptotics need not apply (Watanabe, 2007). SLT replaces the regular dimension term by the RLCT, which captures the local vanishing structure of the excess risk together with the associated parameter-space volume. We use *local singular complexity* to denote the complexity quantified by the RLCT of a local excess-loss germ.

Under standard realizable SLT assumptions at unit Bayesian temperature,

$$\mathbb{E}[G_n] = \frac{\lambda_{\text{stat}}}{n} + o(n^{-1}), \tag{1}$$

where G_n is the Bayes predictive excess log-loss and the outer expectation is over the training sample (Watanabe, 2009). Here λ_{stat} denotes the RLCT of the population excess-risk or KL germ, hereafter called the population RLCT. A local RLCT is the standard asymptotic invariant associated with a chosen excess-loss germ.

Local Learning Coefficient. In this paper, LLC denotes the finite-temperature empirical Gibbs quantity $\widehat{\lambda}_t$ defined in Eq. (5), following localized LLC estimation (Lau et al., 2025). For a fixed analytic empirical loss and fixed localization, $\widehat{\lambda}_t$ converges to that loss's local RLCT as $t \to \infty$ under Lemma 4.1. Connecting the LLC to the population RLCT λ_{stat} requires the statistical compatibility assumptions in Appendix A.4.

2.2 Neuroalgebraic Geometry

Polynomial Neural Networks and Evaluation Maps. A PNN composes linear maps with polynomial activations, for example $f_\theta(x) = W_L\sigma(W_{L-1}\sigma(\cdots\sigma(W_1x)))$ with $\sigma(z) = z^p$. For fixed inputs, the network outputs and squared-error loss are polynomial functions of the parameters.

For a training set, the evaluation map

$$\Phi_n(\theta) = \big(f_\theta(x_1), \ldots, f_\theta(x_n)\big) \tag{2}$$

is therefore polynomial. Its fibers consist of parameter values producing the same empirical outputs, and function-preserving symmetries generate non-identifiable directions within such fibers. The zero-residual set is a real algebraic set, while the image of Φ_n is semialgebraic. Fiber dimension or rank deficiency of $d\Phi_n$ alone does not determine the local RLCT of the empirical squared-error excess loss; the local vanishing orders of this loss must also be taken into account. These properties motivate the local analytic treatment used in Section 4.

2.3 Neural Collapse

Four Axioms of NC. NC describes a collection of geometric phenomena observed near the terminal phase of classifier training (Papyan et al., 2020):

- **NC1 (Variability Collapse)**: Within-class variability vanishes, so representations collapse to their class means.
- **NC2 (Simplex ETF)**: Globally centered class means form a Simplex Equiangular Tight Frame (ETF) up to a common scale and rotation.
- **NC3 (Self-Duality)**: Classifier vectors align with the centered class means up to a common positive scale.
- **NC4 (Nearest Center Decision)**: Classification agrees with the nearest-class-center rule.

The quantitative NC errors and ETF projection used in our analysis are introduced in Section 3.

3 Setup for Localized Comparisons

3.1 Model and Training Objective

Following Shahverdi et al. (2026), consider a depth-L PNN with dimensions $\mathbf{d} = (d_0, \ldots, d_L)$, where $d_L = K$ and $d_{L-1} \geq K - 1$ for ETF feasibility. The hidden layers share the fixed polynomial activation $\sigma_c(z) = \sum_{j=0}^{p} c_j z^j$, with $p > 1$. For $\theta = (W_1, \ldots, W_L)$, $W_l \in \mathbb{R}^{d_l \times d_{l-1}}$, write $\phi_\theta(x) \in \mathbb{R}^{d_{L-1}}$ for the penultimate representation and $f_\theta(x) = W_L\phi_\theta(x)$. For fixed inputs, both f_θ and its squared-error loss are polynomial in θ.

Given class-balanced data $\hat{\mathcal{D}}_n = \{(x_i, y_i)\}_{i=1}^n$, define the empirical MSE loss

$$L_n(\theta) = \frac{1}{2n}\sum_{i=1}^{n} \|f_\theta(x_i) - y_i\|_2^2. \tag{3}$$

Its gap from the NC-projected loss below quantifies NC1 and NC2 effects, with class balance giving equal class weights in the reference.

3.2 Localized Gibbs Measure and LLC

Fix a localization center θ_0, a compact semianalytic neighborhood U of θ_0, and $\gamma > 0$. Set $m_U = \inf_{\theta \in U} L_n(\theta)$ and $\mathcal{K}_n = L_n - m_U$. For $t > 0$, use the t-independent localizer $\rho_\gamma(\theta \mid \theta_0) = \exp[-\gamma\|\theta - \theta_0\|_2^2/2]$ to define

$$P_{t,\gamma,U}(d\theta) = \frac{e^{-tL_n(\theta)}\rho_\gamma(\theta \mid \theta_0)\mathbf{1}_U(\theta)}{Z_U(t)}\, d\theta, \qquad Z_U(t) = \int_U e^{-tL_n(\theta)}\rho_\gamma(\theta \mid \theta_0)\, d\theta. \tag{4}$$

The LLC is

$$\widehat{\lambda}_{t,\gamma,U}(\theta_0) \coloneqq t\big(\mathbb{E}_{P_{t,\gamma,U}}[L_n] - m_U\big) = t\,\mathbb{E}_{P_{t,\gamma,U}}[\mathcal{K}_n]. \tag{5}$$

With γ, U, and θ_0 fixed, abbreviate this as $\widehat{\lambda}_t$. Both losses use the same localization so that their LLC comparison reflects the change in loss; further details appear in Appendix A.3.

3.3 NC-Projected Reference and Regularity

The NC reference removes within-class variation and projects the centered class means onto a scaled simplex ETF. Let $\mathcal{C}_k$ contain the n/K samples in class k, and set $h_i = \phi_\theta(x_i)$, $\mu_k = |\mathcal{C}_k|^{-1}\sum_{i\in\mathcal{C}_k} h_i$, and $\delta_i = h_i - \mu_k$ for $i \in \mathcal{C}_k$. Thus $\sum_{i\in\mathcal{C}_k}\delta_i = 0$. Write $W = W_L$, $M = [\mu_1, \ldots, \mu_K]$, $\mu_G = K^{-1}\sum_k \mu_k$, and $\widetilde{M} = M - \mu_G\mathbf{1}^\top$.

Select a nearest centered scaled simplex ETF $\widetilde{M}^*$ to $\widetilde{M}$ in Frobenius norm on the nondegenerate branch $a > 0$, and set $M^* = [\mu_1^*, \ldots, \mu_K^*] \coloneqq \mu_G\mathbf{1}^\top + \widetilde{M}^*$. This preserves the global mean and gives $M - M^* = \widetilde{M} - \widetilde{M}^*$. Define the NC deviations by

$$\epsilon_{NC1}(\theta) \coloneqq \frac{1}{n}\sum_{i=1}^n \|\delta_i\|_2^2, \qquad \epsilon_{NC2}(\theta) \coloneqq \frac{1}{K}\|\widetilde{M} - \widetilde{M}^*\|_F^2. \tag{6}$$

Exact NC means $\epsilon_{NC1} = \epsilon_{NC2} = 0$.

With $Y = [y_1, \ldots, y_K]$ collecting the class targets, the NC-projected loss is

$$L_n^{NC}(\theta) \coloneqq \frac{1}{2K}\|WM^* - Y\|_F^2. \tag{7}$$

Define P_t^{NC} and $Z_U^{NC}(t)$ by replacing L_n with L_n^{NC} in equation 4, keeping the same neighborhood and localizer. Centering the projected LLC at its own local infimum gives

$$m_U^{NC} \coloneqq \inf_{\theta\in U} L_n^{NC}(\theta), \qquad \widehat{\lambda}_t^{NC} \coloneqq t\big(\mathbb{E}_{P_t^{NC}}[L_n^{NC}] - m_U^{NC}\big). \tag{8}$$

For the projected RLCT interpretation, additionally require $M \mapsto M^*$ to be single-valued and analytic on U, with U isolating the minimizing germ of $L_n^{NC} - m_U^{NC}$. The nearest scaled ETF projection is unique and analytic when $\operatorname{rank}(\widetilde{M}) = K - 1$, including near every positive-scale exact-NC point; see Appendix A.3. The deterministic loss bound and finite-temperature LLC comparison require a well-defined projection (measurable for Gibbs integration), but not analyticity.

The loss-deviation estimate assumes finite constants $B_W, B_R > 0$ such that, on U,

$$\|W\|_{\mathrm{op}} \le B_W, \qquad K^{-1/2}\|WM^* - Y\|_F \le B_R. \tag{9}$$

These bounds exist on compact U when the projection is continuous, and their sizes determine the coefficients in Proposition 4.7.

4 THEORETICAL ANALYSIS OF NEURAL COLLAPSE IN LENS OF SINGULARITY

4.1 PREPARATORY LEMMAS

We record two preliminary facts used below. First, for a fixed loss and localization, the finite-temperature LLC convergence to the local RLCT as $t \to \infty$, connecting the LLC to local singular complexity. Second, we decompose the empirical loss, to isolate its NC-dependent terms used in the loss-deviation bound of Proposition 4.7. This decomposition separates the within-class variation removed by NC1 from the class-mean term that agrees with the NC-projected loss under NC2. Proofs are collected in Appendix C.1.

Lemma 4.1 (LLC Convergence for a Fixed Local Loss)**.** *Under the localized Gibbs setup of Section 3, fix the loss, neighborhood U, and localizer as $t \to \infty$. Suppose that the corresponding excess loss has RLCT λ and multiplicity $\mathfrak{m}$, and that*

$$Z_U(t) = e^{-tm_U} C\, t^{-\lambda} (\log t)^{\mathfrak{m}-1} (1 + \eta(t)), \qquad C > 0, \qquad \eta(t) = o(1). \tag{10}$$

If

$$t\,\eta'(t) = o\left(\frac{1}{\log t}\right), \tag{11}$$

then

$$\widehat{\lambda}_t = \lambda - \frac{\mathfrak{m}-1}{\log t} + o\left(\frac{1}{\log t}\right). \tag{12}$$

In particular, $\widehat{\lambda}_t \to \lambda$ as $t \to \infty$.

Lemma 4.2 (NC-Adapted MSE Decomposition)**.** *For every $\theta \in U$, the empirical loss decomposes exactly as*

$$L_n(\theta) = \underbrace{\frac{1}{2K}\sum_{k=1}^{K} \|W\mu_k - y_k\|_2^2}_{L_{\mathrm{cls}}(\theta)} + \underbrace{\frac{1}{2n}\sum_{i=1}^{n} \|W(h_i - \mu_{y_i})\|_2^2}_{L_{\mathrm{within}}(\theta)}. \tag{13}$$

In particular, exact NC1 removes L_{within}. Exact NC2 identifies the centered matrix $\widetilde{M}$ with its Simplex ETF representative $\widetilde{M}^*$, so $M = M^*$ while preserving μ_G. Thus exact NC1 and exact NC2 gives $L_n = L_n^{NC}$ at such a configuration.

4.2 QUARTIC LOSS BEHAVIOR AND LOCAL RLCT UNDER EXACT-NC

We now analyze how the centered empirical loss $\mathcal{K}_n$ increases around an exact-NC solution. Exact-NC can appear as a higher-order structure in the loss landscape: Proposition 4.3 constructs a local curve along which the excess loss has quartic leading order. Under a separate mixed-order normal-form condition, Theorem 4.5 computes the local RLCT. Proofs are collected in Appendix C.2.

Proposition 4.3 (A Sufficient Construction of a Quartic NC Unfolding)**.** *Let $W_0 = (W_{1,0}, \ldots, W_{L,0})$ be an exact-NC zero-residual solution satisfying NC1, NC2 and NC3, with $d_{L-1} > K - 1$. Fix a class k containing distinct samples $i_+, i_- \in \mathcal{C}_k$. Let $z_i := h_{L-2,0}(x_i)$ and $a_i := W_{L-1,0} z_i$. Assume that the penultimate preactivations $a_1, \ldots, a_n$ are linearly independent, which in particular requires $d_{L-1} \geq n$, and that $\sigma'(a_{ij}) \neq 0$ for every coordinate j and $i \in \{i_+, i_-\}$. Then there exists an analytic parameter curve $W(s)$ varying only W_{L-1}, with $W(0) = W_0$ and $W_L(s) = W_{L,0}$, such that*

$$\frac{d}{ds} f_{W(s)}(x_i)\bigg|_{s=0} = 0 \quad \textit{for all } i, \qquad \frac{d^2}{ds^2} f_{W(s)}(x_{i_\pm})\bigg|_{s=0} \neq 0. \tag{14}$$

Consequently, for some $c > 0$,

$$\mathcal{K}_n(W(s)) = cs^4 + O(s^5). \tag{15}$$

Remark 4.4 (Interpolation condition). The linear-independence assumption is a convenient sufficient condition for the explicit interpolation construction. More generally, the prescribed first- and second-order preactivation jets are realizable whenever they satisfy all linear relations among the reference preactivations, equivalently,

$$\ker(\mathsf{A}_0) \subseteq \ker(\mathsf{B}_1) \cap \ker(\mathsf{B}_2). \tag{16}$$

When $n > d_{L-1}$, full column rank is unavailable, and the sparse two-sample construction used above does not in general guarantee this compatibility. Extending the construction to generic rank-deficient preactivation configurations is left open.

While Proposition 4.3 gives an analytic curve with quartic leading-order loss, this one-dimensional unfolding does not by itself imply a decomposition of the full parameter space into independent higher-order coordinates. We therefore treat the mixed-order normal form below as an additional local structural assumption and characterize the corresponding RLCT conditionally on that form.

Theorem 4.5 (RLCT of a Mixed-Order Local Normal Form)**.** *Let θ_0 be an exact-NC zero-residual solution. Suppose that, in local analytic coordinates (u, v, ξ) around θ_0, the excess loss satisfies*

$$\mathcal{K}_n(\theta) \asymp \|u\|_2^2 + \sum_{j=1}^{q} b_j |v_j|^{2m_j}, \qquad b_j > 0, \quad m_j \in \mathbb{Z}_{\geq 2}, \quad r + q > 0, \tag{17}$$

where $u \in \mathbb{R}^r$, $v \in \mathbb{R}^q$, and ξ denotes locally flat coordinates. Then the local RLCT is

$$\lambda_{NC} = \frac{r}{2} + \sum_{j=1}^{q} \frac{1}{2m_j}. \tag{18}$$

Consequently, under the remainder condition of Lemma 4.1, the corresponding fixed-loss LLC satisfies $\widehat{\lambda}_t \to \lambda_{NC}$ as $t \to \infty$.

We next use this local normal form as the basis for a restricted comparison: specified perturbations in which some higher-order directions acquire quadratic curvature.

Theorem 4.6 (RLCT Increase under Specified Quadratic Unfoldings)**.** *Under Theorem 4.5, let $\varnothing \neq S \subseteq \{1, \ldots, q\}$. Let $\mathcal{K}_{\mathrm{pert}}$ be the local excess loss at a perturbed local minimizer θ_{pert}. Suppose that, in local analytic coordinates around θ_{pert},*

$$\mathcal{K}_{\mathrm{pert}}(\theta) \asymp \|u\|_2^2 + \sum_{j \in S} \kappa_j v_j^2 + \sum_{j \notin S} b_j |v_j|^{2m_j}, \qquad \kappa_j > 0, \tag{19}$$

with b_j, m_j as in Eq. (17). Then

$$\lambda_{\mathrm{pert}} = \frac{r}{2} + \sum_{j \in S} \frac{1}{2} + \sum_{j \notin S} \frac{1}{2m_j}, \tag{20}$$

and hence

$$\lambda_{\mathrm{pert}} - \lambda_{NC} = \sum_{j \in S} \frac{m_j - 1}{2m_j} > 0. \tag{21}$$

Within the specified quadratic unfolding family, adding quadratic curvature to higher-order loss directions increases the local RLCT above λ_{NC}. A fixed-feature classifier-head specialization is given in Appendix B.

4.3 LLC Deviation Bound under ϵ-NC

In this section, we move from the exact-NC RLCT analysis to a LLC stability result under ϵ-NC, which is more practical state of trained models. To extend our theory, we control how NC1 and NC2 deviations affect the gap between the empirical loss and its NC-projected counterpart. This provides the loss-deviation bound needed to control the LLC deviation. Proofs are collected in Appendix C.3.

Proposition 4.7 (Empirical Loss Deviation under ϵ-NC)**.** *For every $\theta \in U$, the deviation from the exact-NC projected loss is bounded by*

$$\left| L_n(\theta) - L_n^{NC}(\theta) \right| \leq \Delta_{NC}(\theta), \tag{22}$$

where

$$\Delta_{NC}(\theta) = \frac{B_W^2}{2} \epsilon_{NC1}(\theta) + B_R B_W \sqrt{\epsilon_{NC2}(\theta)} + \frac{B_W^2}{2} \epsilon_{NC2}(\theta). \tag{23}$$

The LLC compares Gibbs expectations over U, so the pointwise loss estimate must hold uniformly across this neighborhood. For $j \in \{1, 2\}$, define

$$\overline{\epsilon}_{NCj}(U) \coloneqq \sup_{\theta \in U} \epsilon_{NCj}(\theta), \tag{24}$$

and

$$\overline{\Delta}_{NC}(U) \coloneqq \frac{B_W^2}{2}\overline{\epsilon}_{NC1}(U) + B_R B_W \sqrt{\overline{\epsilon}_{NC2}(U)} + \frac{B_W^2}{2}\overline{\epsilon}_{NC2}(U). \tag{25}$$

Proposition 4.7 then gives $|L_n - L_n^{NC}| \le \overline{\Delta}_{NC}(U)$ on U, the perturbation bound used in Theorem 4.8.

Theorem 4.8 (LLC Stability under ϵ-NC)**.** *Let*

$$L_s = L_n^{NC} + s\big(L_n - L_n^{NC}\big), \qquad s \in [0, 1]. \tag{26}$$

Suppose that the corresponding localized Gibbs measures satisfy

$$t^2 \operatorname{Var}_{P_{t,s}}[L_s] \le C_H \tag{27}$$

uniformly in s. Then

$$\left|\widehat{\lambda}_t - \widehat{\lambda}_t^{NC}\right| \le (2 + \sqrt{C_H})\, t\, \overline{\Delta}_{NC}(U). \tag{28}$$

At fixed inverse temperature t and under the stated heat-capacity bound, small uniform NC deviations keep the LLC close to its NC-projected counterpart. Asymptotic recovery of the exact-NC generalization coefficient requires the compatibility and joint scaling conditions of Theorem 4.9.

4.4 Bayesian Generalization of ϵ-NC

We finally connect the LLC deviation bound under ϵ-NC to Bayesian generalization by relating the empirical LLC comparison to the statistical RLCT in the SLT expansion of the Bayes generalization error. Under additional statistical compatibility assumptions, small ϵ-NC deviation yields a Bayesian generalization expansion with coefficient λ_{NC}. Proofs are collected in Appendix C.4.

Let $t_n = n\beta_n$ with $\beta_n > 0$, $\beta_n \to 0$, and $t_n \to \infty$. Assume the statistical compatibility conditions of Appendix A.4, and apply Theorem 4.8 to the normalized pair (ℓ_n, ℓ_n^{NC}) on U_n. Suppose

$$\overline{\Delta}_{NC}^{(\ell)}(U_n) \le \Delta_n^{\mathrm{env}} \tag{29}$$

almost surely and let $C_H(n)$ be a deterministic uniform heat-capacity bound.

Theorem 4.9 (Bayesian Generalization under ϵ-NC)**.** *Under the preceding conditions, if*

$$(2 + \sqrt{C_H(n)})\, t_n \Delta_n^{\mathrm{env}} \to 0, \tag{30}$$

then

$$\mathbb{E}[G_n] = \frac{\lambda_{NC}}{n} + o(n^{-1}). \tag{31}$$

More generally,

$$\left|\mathbb{E}[G_n] - \frac{\lambda_{NC}}{n}\right| \le (2 + \sqrt{C_H(n)})\beta_n \Delta_n^{\mathrm{env}} + o(n^{-1}). \tag{32}$$

Under statistical compatibility and the joint scaling condition, approximate NC recovers λ_{NC} as the coefficient of the leading n^{-1} term in the expected Bayes generalization error.

Remark 4.10. If the heat-capacity or NC-deviation bounds are data-dependent, the deterministic envelope may be replaced by the corresponding integrable expectation; see Appendix A.4.

5 Empirical Validation

5.1 Finite-Deviation Bounds under ϵ-NC

We test the loss and LLC bounds of Proposition 4.7 and Theorem 4.8 using PNN checkpoints. Observed $|L_n - L_n^{NC}|$ remains below Δ_{NC} (Figure 1), and actual/projected LLC differences, each recentered at its own local infimum, satisfy the stability bound under measured heat-capacity control, though the bound is substantially conservative. We also report normalized localized free-energy differences as an intermediate quantity linking the loss and LLC bounds. ReLU MLP, VGG, and ResNet experiments extend beyond the formal PNN assumptions; checkpoint and sampling diagnostics appear in Appendix D.

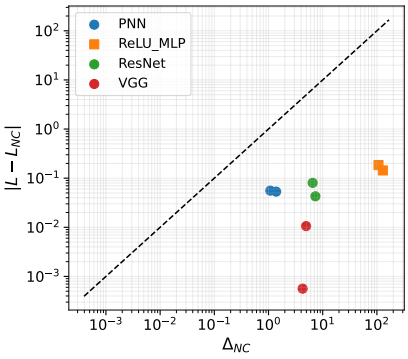


Figure 1: Empirical loss deviation under ϵ-NC. Observed $|L_n - L_n^{NC}|$ remains below the NC-based bound Δ_{NC} across all evaluated architectures.

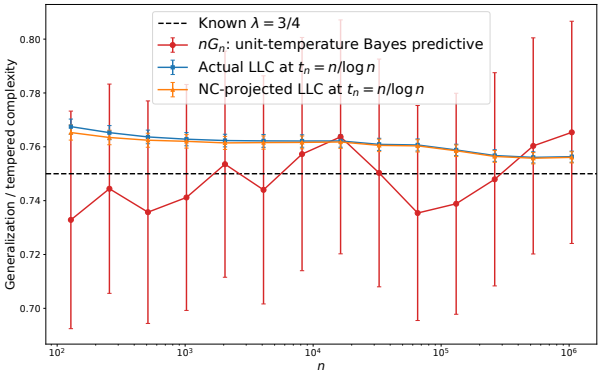
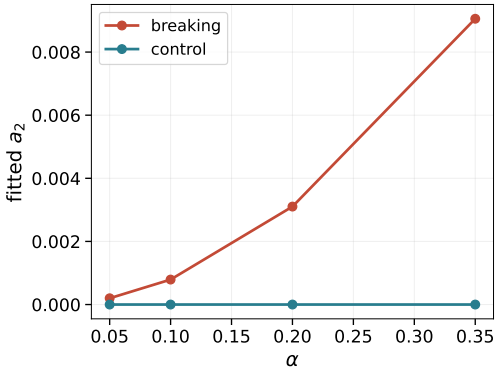
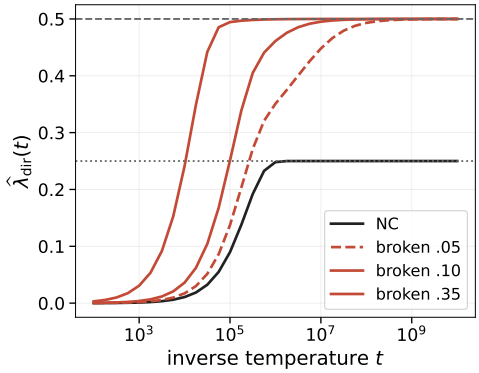


Figure 2: Bayesian generalization and tempered LLCs in a realizable synthetic PNN. The unit-temperature nG_n and tempered actual and NC-projected LLCs remain near the known RLCT $\lambda = 3/4$ across sample sizes.



(a) Loss coefficients along the selected activation direction. Breaking and Control have matched parameter distances from the reference. (Breaking: function-preserving anisotropic scaling that violates NC2, Control: function-preserving isotropic scaling that preserves NC1 and NC2, Reference: the original exact-NC solution.)



(b) The LLC of the loss restricted to this direction moves from $1/4$ toward $1/2$, with stronger perturbations producing the transition at smaller inverse temperatures.

Figure 3: Testing the quadratic-unfolding mechanism along a selected direction near an exact-NC solution. The restricted loss changes from quartic to quadratic under the tested anisotropic perturbations, with its LLC moving from $1/4$ toward $1/2$.

5.2 Bayesian Generalization in a Realizable Synthetic Model

We examine the SLT interpretation of Theorem 4.9 in a realizable synthetic model generated by a fixed ground-truth PNN. Figure 2 compares nG_n, the Bayesian generalization error scaled by training sample size n, with the actual and NC-projected LLCs across sample sizes. SLT predicts $\mathbb{E}[nG_n] \to \lambda_{\text{stat}} = 3/4$, the known population RLCT. The results support the predicted n^{-1} generalization scaling with coefficient $3/4$ over the tested sample sizes.

5.3 Quartic-to-Quadratic Transition under Controlled Perturbations

We test the quadratic-unfolding mechanism of Theorem 4.6 along a selected direction near an exact-NC solution. The tested anisotropic perturbations add a positive quadratic term to the restricted excess loss, while norm-matched controls retain approximately quartic behavior. One-dimensional Gibbs integration yields LLCs approaching $1/4$ for quartic losses and $1/2$ for quadratic losses (Figure 3), supporting the predicted increase in the one-dimensional RLCT when quadratic curvature is introduced.

6 Discussion and Implications

Neuroalgebraic Scope. Neuroalgebraic geometry has primarily been studied in idealized or small models (Shahverdi et al., 2026; Hashimoto et al., 2026; Usevich et al., 2025; Shahverdi et al., 2025; Marchetti et al., 2025). We extend this perspective to NC in PNNs, while non-PNN architectures remain empirical extensions; related directions include mode collapse (Kossale et al., 2022) and memorization (Gu et al., 2025).

Representation Geometry and Singular Complexity. Complementing representation and optimization accounts (Zhu et al., 2021; Wang et al., 2023; Mixon et al., 2022; Galanti et al., 2022) and sharpness-based complexity (Munn et al., 2024), our analysis distinguishes NC representation geometry from local analytic loss structure, whose RLCT is invariant under invertible local analytic reparameterization. Under the NC-adapted local normal form, exact-NC exhibits higher-order singular directions whose quadratic unfolding increases the local RLCT, as supported along a verified compensated activation direction in Experiment 5.3. An anisotropic control instead breaks NC2 while remaining on the same function-preserving $GL(3)$ orbit and thus does not realize the quadratic unfolding of Theorem 4.6. Thus, the theorem concerns changes in local loss structure rather than arbitrary changes in representation geometry.

Low-Cost Monitoring of Local Complexity. Beyond PNNs, our Transformer experiments suggest NC statistics as inexpensive indicators of local complexity regimes. Dense NC measurements track regime changes, including the grokking transition, without repeated SGLD-based LLC estimation. In the representative run in Figure 4, a pronounced change in the causal moving average of Δ_{NC} occurs within one measurement interval of the independently defined grokking transition. We interpret this coincidence as low-cost regime monitoring rather than a unique grokking detector.

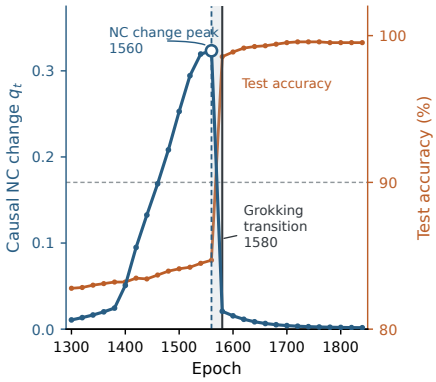


Figure 4: Representative zoom-in of low-cost NC monitoring near grokking. A pronounced change in the causal moving average of Δ_{NC} occurs near the independently defined generalization transition. Multi-seed timing and robustness results appear in Appendix D.

In-Domain and Out-of-Domain Generalization. Theorem 4.9 is an in-domain result, with G_n and λ_{stat} defined relative to the data-generating distribution underlying training. Under statistical compatibility, approximate NC connects λ_{stat} to λ_{NC}, yielding $\mathbb{E}[G_n] = \lambda_{NC}/n + o(n^{-1})$. Distribution shift instead defines a distinct target population KL or excess-risk germ whose minimizers and RLCT need not coincide with their in-domain counterparts. Thus, even with a fixed Bayesian parameter prior, the NC characterization need not transfer to Out-of-Distribution predictive risk, providing a singular-learning perspective on why NC can support in-domain generalization without implying comparable Out-of-Distribution generalization.

7 Conclusion

This work connects NC to SLT through local RLCT analysis and LLC deviation bounds. Under the stated local normal-form conditions, we compute the local RLCT at exact NC and show that specified quadratic unfoldings increase it. For approximate NC, we bound the LLC gap between the empirical and NC-projected losses using NC1 and NC2 deviations. Under statistical compatibility and joint scaling conditions, we identify the exact-NC local RLCT as the leading coefficient of the expected Bayesian generalization error. Experiments provide finite-sample support for the LLC deviation bound and the Bayesian generalization predictions. Deterministic NC statistics further provide inexpensive candidate indicators for monitoring changes in local complexity regimes during training. Together, these results establish a concrete SLT-based connection between NC geometry and local singular complexity.

AI USE STATEMENT

In this work, we used generative AI tools for assist in the writing of proofs, design or provide feedback on research methodology or experiments, support qualitative and thematic data analysis, interpret results. We have not used generative AI tools for help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, propose or refine hypotheses, and Generate synthetic data sets, assist with translation, implement methods, clean and reformat dataset are not applicable to this work. Additionally, we used generative AI tools for create or modify scientific figures or images, suggest experimental parameters, create or edit software code, summarize or analyses existing literature, edit a research paper to improve readability, draft parts of a research paper. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Shun-ichi Amari, Hyeyoung Park, and Tomoko Ozeki. Singularities affect dynamics of learning in neuromanifolds. *Neural Computation*, 18(5):1007–1065, 2006. doi: 10.1162/neco.2006.18.5.1007.

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias-variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019. doi: 10.1073/pnas.1903070116.

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/2202.09028.

Xiangming Gu, Chao Du, Tianyu Pang, Chongxuan Li, Min Lin, and Ye Wang. On memorization in diffusion models. *Transactions on Machine Learning Research*, 2025. URL https://openreview.net/forum?id=V2wQ0T5q4P.

Ting Han, Linara Adilova, Henning Petzka, Jens Kleesiek, and Michael Kamp. Flatness is necessary, neural collapse is not: Rethinking generalization via grokking. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. URL https://openreview.net/forum?id=lbtOctHDQ3.

Yuka Hashimoto, Sho Sonoda, Isao Ishikawa, and Masahiro Ikeda. Why High-rank Neural Networks Generalize?: An Algebraic Framework with RKHSs. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/e36675ab46e276069ec9ea0b61125d0b-Abstract-Conference.html.

Sepp Hochreiter and Jürgen Schmidhuber. Flat minima. *Neural computation*, 9(1):1–42, 1997.

Jesse Hoogland, George Wang, Matthew Farrugia-Roberts, Liam Carroll, Susan Wei, and Daniel Murfet. Loss landscape degeneracy and stagewise development in transformers. *Transactions on Machine Learning Research*, 2025.

Like Hui, M. Belkin, and Preetum Nakkiran. Limitations of neural collapse for understanding generalization in deep learning. In *arXiv.org*, 2022.

Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. In *International Conference on Learning Representations*, 2017.

Joe Kileel, Matthew Trager, and Joan Bruna. On the expressive power of deep polynomial neural networks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/a0dc078ca0d99b5ebb465a9f1cad54ba-Paper.pdf.

Hoyong Kim and Kangil Kim. Fixed Non-negative Orthogonal Classifier: Inducing Zero-mean Neural Collapse with Feature Dimension Separation. In *International Conference on Learning Representations*, May 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/c07d71ff0bc042e4b9acd626a79597fa-Abstract-Conference.html.

Youssef Kossale, Mohammed Airaj, and Aziz Darouichi. Mode collapse in generative adversarial networks: An overview. In *2022 8th International Conference on Optimization and Applications (ICOA)*, 2022. doi: 10.1109/ICOA55659.2022.9934291. URL https://ieeexplore.ieee.org/document/9934291/.

Edmund Lau, Zach Furman, George Wang, Daniel Murfet, and Susan Wei. The local learning coefficient: A singularity-aware complexity measure. In Yingzhen Li, Stephan Mandt, Shipra Agrawal, and Emtiyaz Khan (eds.), *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 244–252. PMLR, 2025. URL https://proceedings.mlr.press/v258/lau25a.html.

Giovanni Luca Marchetti, Vahid Shahverdi, Stefano Mereta, Matthew Trager, and Kathlén Kohn. Position: Algebra unveils deep learning - an invitation to neuroalgebraic geometry. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 81759–81773, 2025. URL https://proceedings.mlr.press/v267/marchetti25a.html.

Dustin G. Mixon, Hans Parshall, and Jianzong Pi. Neural collapse with unconstrained features. *Sampling Theory, Signal Processing, and Data Analysis*, 20(2):11, 2022. doi: 10.1007/s43670-022-00027-5.

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*, volume 37, 2024. doi: 10.52202/079017-2131. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/7b24015f3af598e1d9179f6e06353780-Abstract-Conference.html.

Vardan Papyan, XY 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, 2020.

Vahid Shahverdi, Giovanni Luca Marchetti, and Kathlén Kohn. On the geometry and optimization of polynomial convolutional networks. In *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 604–612, 2025. URL https://proceedings.mlr.press/v258/shahverdi25a.html.

Vahid Shahverdi, Giovanni Luca Marchetti, and Kathl'en Kohn. Learning on a razor's edge: Identifiability and singularity of polynomial neural networks. In *International Conference on Learning Representations (ICLR)*, 2026.

Jingtong Su, Ya Shi Zhang, Nikolaos Tsilivis, and Julia Kempe. On the robustness of neural collapse and the neural collapse of robustness. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=OyXS4ZIqd3.

Peter Súkeník, Marco Mondelli, and Christoph H. Lampert. Deep neural collapse is provably optimal for the deep unconstrained features model. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-2306. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/a60c43ba078b723d3d517d28c50ded4c-Abstract-Conference.html.

Tom Tirer and Joan Bruna. Extended unconstrained features model for exploring deep neural collapse. In *International Conference on Machine Learning*, 2022.

Konstantin Usevich, Ricardo Augusto Borsoi, Clara D'erand, and Marianne Clausel. Identifiability of deep polynomial neural networks. In *Advances in Neural Information Processing Systems*, 2025.

Ke Wang, Vidya Muthukumar, and Christos Thrampoulidis. Benign overfitting in multiclass classification: All roads lead to interpolation. *IEEE Transactions on Information Theory*, 69(12): 7909–7952, 2023. doi: 10.1109/TIT.2023.3320098.

Zheng Wang, Wanhao Yu, Li Yang, and Sen Lin. Rethinking continual learning with progressive neural collapse. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/4ed3645a042ec800a474a6a312575c5e-Abstract-Conference.html.

Sumio Watanabe. Almost all learning machines are singular. In *IEEE Symposium on Foundations of Computational Intelligence*, 2007. doi: 10.1109/FOCI.2007.371500.

Sumio Watanabe. *Algebraic geometry and statistical learning theory*, volume 25. Cambridge university press, 2009.

Sumio Watanabe. A widely applicable bayesian information criterion. *Journal of Machine Learning Research*, 14:867–897, 2013.

Sumio Watanabe. *Mathematical Theory of Bayesian Statistics*. Chapman and Hall/CRC, 2018.

Zhihui Zhu, Tianyu Ding, Jinxin Zhou, Xiao Li, Chong You, Jeremias Sulam, and Qing Qu. A geometric analysis of neural collapse with unconstrained features. In *Neural Information Processing Systems*, 2021.

A Background Detail

A.1 Singular Learning Theory

RLCT and Local Singular Asymptotics. For a nonnegative analytic excess-loss germ $\mathcal{K}$ and a positive analytic density φ, singular learning theory associates the local zeta function

$$\zeta(z) = \int_U \mathcal{K}(\theta)^z \varphi(\theta)\, d\theta. \tag{33}$$

The real log canonical threshold (RLCT) λ is the positive number such that $-\lambda$ is the largest pole of the meromorphic continuation of $\zeta(z)$, with pole order $\mathfrak{m}$ called its multiplicity (Watanabe, 2009). Resolution of singularities expresses the loss and volume element in local charts as

$$\mathcal{K}(g(z)) = A(z) \prod_j z_j^{2k_j}, \qquad |\det Dg(z)| = B(z) \prod_j |z_j|^{\nu_j}, \tag{34}$$

where the relevant units are positive after restricting charts. For a positive analytic density, λ is the minimum of $(\nu_j + 1)/(2k_j)$ over relevant charts and indices with $k_j > 0$. Thus, the RLCT captures both the vanishing order of the loss and the local volume geometry around its minimizers. Regular quadratic models have $\lambda = D/2$, whereas singular models generally require the full local analytic structure rather than Fisher-information or evaluation-map rank alone (Watanabe, 2009).

Localized Gibbs Distribution and LLC. The local learning coefficient (LLC) provides a finite-temperature probe of local singular complexity through a Gibbs distribution concentrated around a trained parameter (Lau et al., 2025). In our localized setting, we define

$$P_t(d\theta) = \frac{e^{-tL(\theta)} \rho_\gamma(\theta \mid \theta_0) \mathbf{1}_U(\theta)\, d\theta}{Z_U(t)}, \tag{35}$$

where t is the inverse temperature, U restricts the local neighborhood, and ρ_γ is a fixed localizer. The corresponding finite-temperature empirical LLC is

$$\widehat{\lambda}_t = t\left(\mathbb{E}_{P_t}[L] - m_U\right), \qquad m_U = \inf_U L. \tag{36}$$

It measures the Gibbs-averaged excess loss above the local minimum, scaled by the inverse temperature. Under the fixed-loss asymptotic conditions of Lemma 4.1, $\widehat{\lambda}_t$ converges to the local RLCT as $t \to \infty$. SGLD approximates this Gibbs expectation rather than automatically providing a consistent estimator of a statistical RLCT. The actual and NC-projected losses are each centered at their own local infimum.

Free Energy. The localized partition function $Z_U(t)$ determines the normalized free energy

$$\mathcal{F}_U(t) = -\frac{1}{t}\log Z_U(t). \tag{37}$$

For a fixed analytic loss with RLCT λ and multiplicity $\mathfrak{m}$, singular learning theory gives the asymptotic expansion (Watanabe, 2009; 2013)

$$-\log Z_U(t) = tm_U + \lambda \log t - (\mathfrak{m}-1)\log\log t + O(1). \tag{38}$$

Thus, the RLCT appears as the coefficient of the leading logarithmic complexity correction to the minimum loss. In our localized analysis, this expansion concerns a fixed local loss and should be distinguished from the sample-dependent Bayesian free energy of a statistical model.

Bayesian Generalization. In a statistical model, Bayesian prediction averages the predictive distribution over the posterior rather than selecting a single fitted parameter. The Bayes generalization error G_n measures the predictive excess log-loss on a fresh sample relative to the population baseline. Under the realizable singular-learning assumptions at unit Bayesian temperature, its expectation satisfies (Watanabe, 2009; 2018)

$$\mathbb{E}[G_n] = \frac{\lambda_{\mathrm{stat}}}{n} + o(n^{-1}), \tag{39}$$

where λ_{stat} is the RLCT of the corresponding population KL or excess-risk germ. This statistical RLCT is conceptually distinct from the RLCT of an arbitrary fixed empirical loss germ. Accordingly, connecting a finite-temperature empirical LLC to Bayesian generalization requires the additional statistical compatibility assumptions stated in our Bayesian result.

A.2 Neuroalgebraic Geometry

Evaluation Maps, Fibers, and Symmetry. For a PNN, the evaluation map $\Phi_n(\theta) = (f_\theta(x_1), \ldots, f_\theta(x_n))$ is polynomial, placing its parameterization geometry within the algebraic and semialgebraic framework studied for polynomial neural networks (Marchetti et al., 2025). Its fibers, including the zero-residual set $V(S) = \{\theta : g(\theta) = 0\ \forall g \in S\}$ for residual polynomials S, are real algebraic sets, while its image is semialgebraic. A drop below the generic rank of $d\Phi_n$ identifies a rank-deficient locus of the parameterization, but does not by itself identify an algebraic-set singularity or determine the loss RLCT. NC constrains representation geometry, whereas our full-network RLCT characterization is conditional on the local loss normal form assumed in Theorem 4.5.

A fiber consists of parameters representing the same empirical output. Non-identifiability and parameter symmetries can produce nontrivial fibers and flat directions (Usevich et al., 2025). Theorem 4.5 instead assumes an explicit local loss normal form in which flat coordinates contribute finite factors while transverse vanishing orders determine the RLCT.

A.3 Localization and NC Projection

Localization Convention. Following localized LLC estimation in singular learning theory (Lau et al., 2025), we use a neighborhood U and a fixed localizer to define the localized Gibbs measure. For practical estimation, β_n may vanish slowly, e.g. $\beta_n = 1/\log n$, while $t_n = n\beta_n \to \infty$. This temperature choice alone does not ensure consistency along a sequence of changing empirical losses. The neighborhood U restricts the support of the localized Gibbs measure, while $\rho_\gamma(\theta \mid \theta_0)$ assigns fixed additional weight near θ_0 and is not scaled by t_n. A localizer of the form $\exp[-t_n\gamma\|\theta-\theta_0\|_2^2/2]$ would instead introduce quadratic concentration at scale $t_n^{-1/2}$ and can suppress higher-order singular directions. We therefore use the fixed localizer in Eq. (4) to preserve the singular asymptotics encoded by the RLCT.

Regularity of the NC Projection. For the projected LLC or RLCT, assume that the selected nearest scaled and aligned ETF representative is locally unique on U and that the map $M \mapsto M^*$ is analytic on the chosen local branch. For singular asymptotics, Lemma 4.1 applies to the analytic excess germ $L_n^{NC} - m_U^{NC}$ on a neighborhood isolating the relevant minimizing germ. These are additional assumptions required for the projected singular-learning interpretation. The pointwise deterministic bound requires only a well-defined projection, while the partition-function bound additionally requires measurability and finite integrals. Neither the finite-temperature Gibbs expectations nor the interpolation inequality requires a singular asymptotic expansion.

A.4 BAYESIAN GENERALIZATION

Gaussian Statistical Realization and Bayes Predictive Generalization. Only the actual network defines test-time predictions. Under a fixed isotropic Gaussian observation model with covariance $\sigma_y^2 I$, define $\ell_n := L_n/\sigma_y^2$ and $\ell_n^{NC} := L_n^{NC}/\sigma_y^2$; L_n remains the empirical loss. The average Gaussian negative log-likelihood is ℓ_n up to an additive constant. Unit Bayesian temperature uses energy ℓ_n at $t = n$, equivalently energy L_n at effective inverse temperature $t = n/\sigma_y^2$.

Define the normalized uniform loss bound $\overline{\Delta}_{NC}^{(\ell)}(U) := \overline{\Delta}_{NC}(U)/\sigma_y^2$. For the LLC comparison in Theorem 4.9, let $t_n = n\beta_n$, where $\beta_n > 0$, $\beta_n \to 0$, and $t_n \to \infty$. Write $\widehat{\lambda}_{t_n}^{(\ell)}$ and $\widehat{\lambda}_{t_n}^{NC,(\ell)}$ for the LLCs constructed from (ℓ_n, ℓ_n^{NC}) at temperature t_n, with the same neighborhood and fixed localizer, centered at their respective infima m_U/σ_y^2 and m_U^{NC}/σ_y^2. This tempering is used only for the LLC comparison and does not alter the unit-temperature Bayesian predictive distribution defining G_n.

Fixed positive scaling of an excess loss leaves its RLCT unchanged. Let G_n be the actual network's Bayes predictive excess log-loss relative to the realizable population baseline. The projected prototype loss defines no test-time conditional distribution or Bayes generalization error. Theorem 4.9 requires separate statistical assumptions.

Statistical Compatibility of the Tempered LLC. The following assumptions apply only to the Bayesian corollary. Assume a realizable analytic statistical model for the actual network with a fixed positive prior, unit Bayesian temperature, and the regularity and sampling assumptions needed for the expected SLT expansion $\mathbb{E}[G_n] = \lambda_{\mathrm{stat}}/n + o(n^{-1})$. Here λ_{stat} is the RLCT of the population excess-risk/KL germ, not the RLCT of an arbitrary empirical loss germ. The selected local component is assumed to capture this statistical asymptotic; the fixed localizer must not alter that component's leading generalization term. Class-balanced sampling, even together with a separately assumed zero-residual reference, does not establish these statistical assumptions.

Along $n \to \infty$, let

$$t_n = n\beta_n, \qquad \beta_n \to 0, \qquad t_n \to \infty. \tag{40}$$

For the normalized energy ℓ_n specified in Section 3, assume explicitly that the recentered tempered LLC $\widehat{\lambda}_{t_n}^{(\ell)}$ recovers λ_{stat}. In addition, fix an exact-NC analytic reference with RLCT λ_{NC} and assume that $\widehat{\lambda}_{t_n}^{NC,(\ell)}$ recovers the same reference invariant. To pass to data expectations, require

$$\mathbb{E}\left[\left|\widehat{\lambda}_{t_n}^{(\ell)} - \lambda_{\mathrm{stat}}\right|\right] = o(1), \qquad \mathbb{E}\left[\left|\widehat{\lambda}_{t_n}^{NC,(\ell)} - \lambda_{NC}\right|\right] = o(1). \tag{41}$$

Hence the sufficient condition used in Theorem 4.9 is

$$(2 + \sqrt{C_H(n)})t_n\Delta_n^{\mathrm{env}} \to 0. \tag{42}$$

Under this condition, $\lambda_{\mathrm{stat}} - \lambda_{NC} \to 0$, which together with the unit-temperature SLT expansion gives

$$\mathbb{E}[G_n] = \frac{\lambda_{NC}}{n} + o(n^{-1}). \tag{43}$$

If $\overline{\Delta}_{NC}^{(\ell)}(U_n)$ is itself deterministic, one may take $\Delta_n^{\mathrm{env}} = \overline{\Delta}_{NC}^{(\ell)}(U_n)$.

If the heat-capacity or NC-deviation bound is data-dependent, the deterministic envelope assumption may instead be replaced by the integrability condition

$$\mathbb{E}\Big[(2 + \sqrt{C_H(n)})t_n\overline{\Delta}_{NC}^{(\ell)}(U_n)\Big] < \infty. \tag{44}$$

The corresponding sufficient condition is

$$\mathbb{E}\Big[(2 + \sqrt{C_H(n)})t_n\overline{\Delta}_{NC}^{(\ell)}(U_n)\Big] \to 0. \tag{45}$$

Conditions for the Bayesian Generalization Bound For Theorem 4.9, apply Theorem 4.8 to the normalized energy pair (ℓ_n, ℓ_n^{NC}) at temperature $t = t_n$ on neighborhoods U_n. Assume that the interpolating Gibbs measures satisfy

$$t_n^2 \operatorname{Var}_{P_{t_n,s}}[\ell_{n,s}] \le C_H(n), \qquad s \in [0,1], \tag{46}$$

almost surely in the data, where

$$\ell_{n,s} = \ell_n^{NC} + s(\ell_n - \ell_n^{NC}), \tag{47}$$

and $C_H(n) < \infty$ is deterministic.

Further, suppose that the normalized NC-deviation bound admits a deterministic envelope

$$\overline{\Delta}_{NC}^{(\ell)}(U_n) \le \Delta_n^{\mathrm{env}} < \infty \qquad \text{almost surely.} \tag{48}$$

Then Theorem 4.8 gives

$$\left|\widehat{\lambda}_{t_n}^{(\ell)} - \widehat{\lambda}_{t_n}^{NC,(\ell)}\right| \le (2 + \sqrt{C_H(n)}) t_n \Delta_n^{\mathrm{env}}. \tag{49}$$

Combining this inequality with the L^1 compatibility conditions Eq. (41) yields

$$|\lambda_{\mathrm{stat}} - \lambda_{NC}| \le (2 + \sqrt{C_H(n)}) t_n \Delta_n^{\mathrm{env}} + o(1). \tag{50}$$

B Additional Theoretical Results

Corollary B.1 (Quadratic Classifier-Head Special Case)**.** *As a quadratic special case of Theorem 4.5, if exact-NC penultimate features are fixed and only the terminal classifier is trainable near a zero-residual solution, then* $r = K\,\mathrm{rank}(M^*)$ *and* $\lambda_{\mathrm{head}} = K\,\mathrm{rank}(M^*)/2$*. For a centered Simplex ETF with* $\mathrm{rank}(M^*) = K - 1$*, this becomes* $\lambda_{\mathrm{head}} = K(K-1)/2$*.*

Proof. For the classifier-head special case, write $W = W_0 + \Delta W$ at a zero-residual solution with fixed features M^*. The excess loss is $(2K)^{-1}\|\Delta W M^*\|_F^2$. Right multiplication by M^* has rank $\mathrm{rank}(M^*)$ for each of the K output rows. A linear change of head coordinates therefore leaves $K\,\mathrm{rank}(M^*)$ nondegenerate quadratic directions and the remaining directions flat, giving $\lambda_{\mathrm{head}} = K\,\mathrm{rank}(M^*)/2$. This calculation holds conditionally on the features. □

Remark B.2. The quantity λ_{head} is a conditional classifier-head RLCT with fixed features, not the full-network LLC or RLCT. In general, the centroid-preserving M^* need not have rank $K-1$. The centered special case requires $\mu_G = 0$ and targets compatible with zero residual: in a biasless network $WM^* = Y$ then forces $Y\mathbf{1} = 0$, excluding ordinary one-hot targets in that special case.

Remark B.3 (Asymptotic Proximity to the Exact-NC Regime). For a sequence indexed by n, allow $t = t_n \to \infty$, $U = U_n$, and $C_H = C_H(n)$. Corollary 4.8 ensures LLC proximity if $(2 + \sqrt{C_H(n)}) t_n \overline{\Delta}_{NC}(U_n) \to 0$. The simpler condition $t_n \overline{\Delta}_{NC}(U_n) \to 0$ suffices when $C_H(n) = O(1)$. Independently of heat capacity, $t_n \overline{\Delta}_{NC}(U_n) = o(\log t_n)$ makes the normalized localized free-energy difference $o(\log t_n / t_n)$. Interpreting this relative to an RLCT term along changing local loss functions requires a uniform singular expansion with nonvanishing RLCT coefficient; Proposition 4.1 supplies an expansion only for a fixed local loss.

Proposition B.4 (ϵ-NC Stability of the Normalized Localized Free Energy)**.** *Let* $Z_U^{NC}(t)$ *denote the localized partition function obtained by replacing* L_n *with* L_n^{NC} *in Eq. (4), and let* $\mathcal{F}_U(t) = -t^{-1}\log Z_U(t)$ *and* $\mathcal{F}_U^{NC}(t) = -t^{-1}\log Z_U^{NC}(t)$*. Then*

$$\left|\mathcal{F}_U(t) - \mathcal{F}_U^{NC}(t)\right| \le \overline{\Delta}_{NC}(U). \tag{51}$$

C Proofs

C.1 Proof of Sectoin 4.1: Preliminaries

Proof of Lemma 4.1.

Proof. The neighborhood and the positive analytic localizer are fixed, so differentiation under the integral gives $-\partial_t \log Z_U(t) = \mathbb{E}_{P_t}[L]$. The excess-energy partition function is $e^{tm_U} Z_U(t)$, hence

$$\widehat{\lambda}_t = -t\,\partial_t \log Z_U(t) - t m_U. \tag{52}$$

The singular expansion in the proposition gives

$$\log Z_U(t) = -tm_U - \lambda \log t + (\mathfrak{m} - 1) \log \log t + \log C + \log(1 + \eta(t)). \tag{53}$$

Consequently,

$$\widehat{\lambda}_t = \lambda - \frac{\mathfrak{m} - 1}{\log t} - \frac{t\eta'(t)}{1 + \eta(t)}. \tag{54}$$

The hypotheses $\eta(t) = o(1)$ and $t\eta'(t) = o(1/\log t)$ imply Eq. (12). A fixed analytic density that is positive near the minimizing germ is an analytic unit (a nonvanishing analytic factor) and changes neither the RLCT nor its multiplicity. No step varies the empirical sample or its loss germ with t. □

Proof of Lemma 4.2.

Proof. For $i \in \mathcal{C}_k$, $Wh_i - y_k = (W\mu_k - y_k) + W\delta_i$. Therefore,

$$\sum_{i \in \mathcal{C}_k} \|Wh_i - y_k\|_2^2 = |\mathcal{C}_k| \|W\mu_k - y_k\|_2^2 + 2(W\mu_k - y_k)^\top W \sum_{i \in \mathcal{C}_k} \delta_i + \sum_{i \in \mathcal{C}_k} \|W\delta_i\|_2^2. \tag{55}$$

The middle term vanishes because $\sum_{i \in \mathcal{C}_k} \delta_i = 0$. Summing over k and using $|\mathcal{C}_k| = n/K$ proves Eq. (13). The identity is independent of an ETF projection. Under exact NC1 the within-class term vanishes, and under exact NC2, $\widetilde{M} = \widetilde{M}^*$, hence $M = \mu_G \mathbf{1}^\top + \widetilde{M}^* = M^*$. □

C.2 Proof of Sectoin 4.2: Local RLCT Characterization under Exact-NC

Proof of Proposition 4.3.

Proof. By NC2 and NC3,

$$W_{L,0} = \alpha \widetilde{M}_0^\top, \qquad \operatorname{rank}(\widetilde{M}_0) = K - 1. \tag{56}$$

Hence $d > K - 1$ implies

$$\ker W_{L,0} \neq \{0\}. \tag{57}$$

Choose a nonzero vector

$$v \in \ker W_{L,0}. \tag{58}$$

Let

$$\rho := \|\widetilde{\mu}_{k,0}\|_2 > 0. \tag{59}$$

We prescribe the first- and second-order feature variations by

$$\delta_{i_+} = v, \qquad \delta_{i_-} = -v, \tag{60}$$

and

$$\eta_{i_+} = \eta_{i_-} = -\frac{\|v\|_2^2}{\rho^2} \widetilde{\mu}_{k,0}, \tag{61}$$

while setting $\delta_i = \eta_i = 0$ for all other training samples.

These are precisely the first two jets of a symmetric transverse splitting of the two collapsed features along the sphere containing the centered simplex ETF.

For each selected sample, define vectors $b_i, d_i \in \mathbb{R}^d$ coordinatewise by

$$b_{ij} := \frac{\delta_{ij}}{\sigma'(a_{ij})}, \tag{62}$$

and

$$d_{ij} := \frac{\eta_{ij} - \sigma''(a_{ij}) b_{ij}^2}{\sigma'(a_{ij})}. \tag{63}$$

For all remaining samples, set $b_i = d_i = 0$. Let

$$\mathsf{B}_1 := [b_1, \ldots, b_n], \qquad \mathsf{B}_2 := [d_1, \ldots, d_n]. \tag{64}$$

Since A_0 has full column rank,

$$\mathsf{A}_0^\dagger \mathsf{A}_0 = I_n. \tag{65}$$

Thus the matrices

$$C := \mathsf{B}_1 \mathsf{A}_0^\dagger, \qquad D := \mathsf{B}_2 \mathsf{A}_0^\dagger \tag{66}$$

satisfy

$$C\mathsf{A}_0 = \mathsf{B}_1, \qquad D\mathsf{A}_0 = \mathsf{B}_2. \tag{67}$$

Define

$$G(s) := I + sC + \frac{s^2}{2} D \tag{68}$$

and perturb only the penultimate weight matrix by

$$W_{L-1}(s) := G(s) W_{L-1,0}. \tag{69}$$

Then

$$W_{L-1}(s) z_i = a_i + s b_i + \frac{s^2}{2} d_i. \tag{70}$$

Expanding the fixed polynomial activation gives

$$h_i(s) = h_i(0) + s\delta_i + \frac{s^2}{2}\eta_i + O(s^3). \tag{71}$$

Since the final classifier is fixed,

$$f_{W(s)}(x_i) - f_{W_0}(x_i) = s W_{L,0}\delta_i + \frac{s^2}{2} W_{L,0}\eta_i + O(s^3). \tag{72}$$

By construction,

$$W_{L,0}\delta_i = 0 \qquad \text{for every } i. \tag{73}$$

For the two selected samples,

$$W_{L,0}\eta_{i_\pm} = -\frac{\|v\|_2^2}{\rho^2} W_{L,0}\widetilde{\mu}_{k,0}. \tag{74}$$

Using NC3 and the simplex ETF identity from NC2,

$$W_{L,0}\widetilde{\mu}_{k,0} = \alpha \widetilde{M}_0^\top \widetilde{\mu}_{k,0} = \alpha a^2 \left(e_k - \frac{1}{K}\mathbf{1} \right) \neq 0. \tag{75}$$

Hence the second-order prediction variation is nonzero.

Since W_0 is a zero-residual solution, the two selected samples therefore contribute a strictly positive fourth-order term to the empirical squared loss, while all first-order prediction variations vanish. Thus

$$\mathcal{K}_n(W(s)) = cs^4 + O(s^5) \tag{76}$$

for some $c > 0$. □

Proof of Theorem 4.5.

Proof. Define the comparison loss

$$Q(u, v) := \|u\|_2^2 + \sum_{j=1}^{q} b_j |v_j|^{2m_j}. \tag{77}$$

By the local equivalence assumption in Eq. (17), after restricting to a sufficiently small coordinate neighborhood of θ_0, there exist constants $c_-, c_+ > 0$ such that

$$c_- Q(u, v) \leq \mathcal{K}_n(\theta(u, v, \xi)) \leq c_+ Q(u, v). \tag{78}$$

Since the coordinate change is analytic and locally invertible, its Jacobian determinant is bounded above and below by positive constants on a sufficiently small coordinate box. The same is true for

the fixed localizer. Hence the local partition integral is bounded above and below by positive constant multiples of

$$I_c(t) := \int \exp\left[-ct\left(\|u\|_2^2 + \sum_{j=1}^{q} b_j |v_j|^{2m_j}\right)\right] du\, dv\, d\xi, \qquad c > 0. \tag{79}$$

Therefore it is sufficient to determine the power-law decay of $I_c(t)$ as $t \to \infty$.

Because Q is separable in the coordinates (u, v), the integral factorizes into a quadratic part, one higher-order part for each v_j, and a flat factor from ξ.

For the regular coordinates $u \in \mathbb{R}^r$, make the change of variables

$$u = t^{-1/2} a. \tag{80}$$

Then

$$du = t^{-r/2}\, da, \tag{81}$$

and hence

$$\int e^{-ct\|u\|_2^2}\, du \asymp t^{-r/2}. \tag{82}$$

Thus each quadratic coordinate contributes $1/2$ to the local RLCT.

For a singular coordinate v_j, consider

$$\int \exp\left(-ctb_j |v_j|^{2m_j}\right)\, dv_j. \tag{83}$$

With the rescaling

$$v_j = t^{-1/(2m_j)} z_j, \tag{84}$$

we have

$$dv_j = t^{-1/(2m_j)}\, dz_j. \tag{85}$$

Therefore

$$\int \exp\left(-ctb_j |v_j|^{2m_j}\right)\, dv_j \asymp t^{-1/(2m_j)}. \tag{86}$$

Hence the coordinate v_j contributes

$$\frac{1}{2m_j} \tag{87}$$

to the local RLCT.

Finally, the coordinates ξ do not appear in Q. Their integration over a fixed local coordinate box therefore contributes only a finite, strictly positive constant and does not change the power of t.

Multiplying the contributions from all coordinates gives

$$I_c(t) \asymp t^{-\left(\frac{r}{2} + \sum_{j=1}^{q} \frac{1}{2m_j}\right)}. \tag{88}$$

Since the true local partition integral is bounded above and below by positive constant multiples of such comparison integrals, it has the same leading power. Consequently,

$$\lambda_{NC} = \frac{r}{2} + \sum_{j=1}^{q} \frac{1}{2m_j}. \tag{89}$$

Moreover, the separated comparison integral has no logarithmic factor. The two-sided comparison above therefore rules out an additional $(\log t)^{\mathfrak{m}-1}$ factor in the local partition asymptotics, so the corresponding pole multiplicity is

$$\mathfrak{m} = 1. \tag{90}$$

The LLC convergence statement then follows from Proposition 4.1 under its stated remainder condition. □

Proof of Theorem 4.6.

Proof. Let $Q_S(u,v) = \|u\|_2^2 + \sum_{j\in S}\kappa_j v_j^2 + \sum_{j\notin S} b_j |v_j|^{2m_j}$ be the comparison germ. As in the preceding normal-form proof, on sufficiently small coordinate boxes the positive analytic density, invertible coordinate Jacobian, and comparison factors are bounded above and below by positive constants. They preserve the partition integral's power exponent. With all $\kappa_j > 0$ fixed, rescale $u = t^{-1/2}\widetilde{u}$, $v_j = t^{-1/2}\widetilde{v}_j$ for $j \in S$, and $v_j = t^{-1/(2m_j)}\widetilde{v}_j$ otherwise. The factorized comparison integral has order

$$t^{-r/2-|S|/2-\sum_{j\notin S}(2m_j)^{-1}}, \tag{91}$$

with a finite positive factor from the unchanged flat ξ fibers. This proves Eq. (20). Subtracting Eq. (18) gives Eq. (21), strictly positive since $S \neq \varnothing$ and $m_j \geq 2$. The comparison constants need not be uniform as a quadratic coefficient tends to zero; this RLCT comparison holds each local loss fixed, not a uniform finite-temperature limit. □

C.3 Proof of Sectoin 4.3: LLC Deviation Bound under ϵ-NC

Proof of Proposition 4.7.

Proof. By Proposition 4.2, $L_n = L_{\text{cls}} + L_{\text{within}}$. For the within-class component,

$$L_{\text{within}} = \frac{1}{2n}\sum_{i=1}^{n}\|W(h_i - \mu_{y_i})\|_2^2 \leq \frac{B_W^2}{2}\epsilon_{NC1}. \tag{92}$$

For the class-level component, let $D = M - M^* = \widetilde{M} - \widetilde{M}^*$; the centroids cancel exactly. Since $M = M^* + D$,

$$L_{\text{cls}}(M) - L_{\text{cls}}(M^*) = \frac{1}{K}\langle WM^* - Y, WD\rangle_F + \frac{1}{2K}\|WD\|_F^2. \tag{93}$$

By Cauchy–Schwarz, $\|W\|_{\text{op}} \leq B_W$, $K^{-1/2}\|WM^* - Y\|_F \leq B_R$, and $K^{-1}\|D\|_F^2 = \epsilon_{NC2}$,

$$|L_{\text{cls}}(M) - L_{\text{cls}}(M^*)| \leq B_R B_W\sqrt{\epsilon_{NC2}} + \frac{B_W^2}{2}\epsilon_{NC2}. \tag{94}$$

Combining the two bounds gives

$$\left|L_n(\theta) - L_n^{NC}(\theta)\right| \leq \Delta_{NC}(\theta). \tag{95}$$

□

Proof of Theorem 4.8.

Proof. Let $\Delta L = L_n - L_n^{NC}$, $\Delta_U = \overline{\Delta}_{NC}(U)$, and $L_s = L_n^{NC} + s\Delta L$ for $s \in [0,1]$. Use the common fixed localizer and neighborhood for $P_{t,s}$, and first interpolate the uncentered quantity $A_s = t\mathbb{E}_{P_{t,s}}[L_s]$. Bounded losses on the chosen analytic local branch justify differentiation under the integral, giving

$$\frac{dA_s}{ds} = t\mathbb{E}_{P_{t,s}}[\Delta L] - t^2\,\mathrm{Cov}_{P_{t,s}}(L_s, \Delta L). \tag{96}$$

Theorem 4.7 gives $|\Delta L| \leq \Delta_U$. Thus $|\mathbb{E}_{P_{t,s}}[\Delta L]| \leq \Delta_U$ and $\mathrm{Var}_{P_{t,s}}(\Delta L) \leq \mathbb{E}_{P_{t,s}}[(\Delta L)^2] \leq \Delta_U^2$. By Cauchy–Schwarz and the heat-capacity assumption,

$$t^2|\,\mathrm{Cov}_{P_{t,s}}(L_s, \Delta L)| \leq t^2\sqrt{\mathrm{Var}_{P_{t,s}}(L_s)}\sqrt{\mathrm{Var}_{P_{t,s}}(\Delta L)} \leq \sqrt{C_H}\,t\Delta_U. \tag{97}$$

Integrating the derivative bound over s therefore yields

$$|A_1 - A_0| \leq (1 + \sqrt{C_H})t\Delta_U. \tag{98}$$

Separately, $L_n^{NC} - \Delta_U \leq L_n \leq L_n^{NC} + \Delta_U$ on U implies, on taking infima,

$$m_U^{NC} - \Delta_U \leq m_U \leq m_U^{NC} + \Delta_U, \qquad |m_U - m_U^{NC}| \leq \Delta_U. \tag{99}$$

Since $\widehat{\lambda}_t = A_1 - tm_U$ and $\widehat{\lambda}_t^{NC} = A_0 - tm_U^{NC}$, the triangle inequality gives

$$|\widehat{\lambda}_t - \widehat{\lambda}_t^{NC}| \leq |A_1 - A_0| + t|m_U - m_U^{NC}| \leq (2 + \sqrt{C_H})t\Delta_U. \tag{100}$$

One unit in the coefficient 2 comes from the direct expectation perturbation, and the other from the difference of the two local infima. No derivative of an interpolated minimum is needed. □

C.4 Proof of Sectoin 4.4: Bayesian Generalization of ϵ-NC

Proof of Theorem 4.9.

Proof. Apply Theorem 4.8 to the normalized energy pair (ℓ_n, ℓ_n^{NC}) at $t = t_n$:

$$\left|\widehat{\lambda}_{t_n}^{(\ell)} - \widehat{\lambda}_{t_n}^{NC,(\ell)}\right| \le (2 + \sqrt{C_H(n)}) t_n \overline{\Delta}_{NC}^{(\ell)}(U_n). \tag{101}$$

The triangle inequality and the L^1 compatibility assumptions Eq. (41), followed by expectation over the data, give

$$|\lambda_{\text{stat}} - \lambda_{NC}| \le (2 + \sqrt{C_H(n)}) t_n \Delta_n^{\text{env}} + o(1). \tag{102}$$

This step uses the separate sequence-level assumptions, not the limit for a fixed local loss in Proposition 4.1.

Dividing by n and using $t_n/n = \beta_n$ gives

$$\frac{|\lambda_{\text{stat}} - \lambda_{NC}|}{n} \le (2 + \sqrt{C_H(n)}) \beta_n \Delta_n^{\text{env}} + o(n^{-1}). \tag{103}$$

The assumed actual-model SLT expansion at unit Bayesian temperature is

$$\mathbb{E}[G_n] = \frac{\lambda_{\text{stat}}}{n} + o(n^{-1}). \tag{104}$$

Combining this expansion with Eq. (103) proves Eq. (32).

If

$$(2 + \sqrt{C_H(n)}) t_n \Delta_n^{\text{env}} \to 0, \tag{105}$$

then the right-hand side of Eq. (102) is $o(1)$, so $\lambda_{\text{stat}} = \lambda_{NC}$ in the limit, and therefore Eq. (31) follows from the unit-temperature SLT expansion.

If the bound is data-dependent, take expectation of the product on the right-hand side of Eq. (101); its integrability and the corresponding vanishing condition give the same conclusion. No projected predictive distribution or projected Bayes generalization error is invoked. □

Proof of Proposition B.4.

Proof. Taking the supremum over $\theta \in U$ yields $|L_n - L_n^{NC}| \le \overline{\Delta}_{NC}(U)$ on U. Therefore,

$$e^{-t\overline{\Delta}_{NC}(U)} e^{-tL_n^{NC}(\theta)} \le e^{-tL_n(\theta)} \le e^{t\overline{\Delta}_{NC}(U)} e^{-tL_n^{NC}(\theta)}. \tag{106}$$

Multiplying by the common localizer $\rho_\gamma(\theta \mid \theta_0)$ and integrating over U gives

$$e^{-t\overline{\Delta}_{NC}(U)} Z_U^{NC}(t) \le Z_U(t) \le e^{t\overline{\Delta}_{NC}(U)} Z_U^{NC}(t). \tag{107}$$

Taking logarithms and multiplying by $-t^{-1}$ proves

$$\left|\mathcal{F}_U(t) - \mathcal{F}_U^{NC}(t)\right| \le \overline{\Delta}_{NC}(U). \tag{108}$$

□

D Additional Experimental Details

D.1 Main-Text Experimental Setup

D.1.1 Finite-Deviation Bounds under ϵ-NC

We compare the empirical loss L_n with the NC-projected loss L_n^{NC} at PNN checkpoints trained on MNIST. At each checkpoint, we compute ϵ_{NC1} and ϵ_{NC2} using the centroid-preserving ETF projection described in the main text. The actual and projected LLCs are recentered at their respective local infima,

$$m_U = \inf_{\theta \in U} L_n(\theta), \qquad m_U^{NC} = \inf_{\theta \in U} L_n^{NC}(\theta). \tag{109}$$

To evaluate the finite-temperature stability bound, we interpolate between the projected and actual losses as

$$L_s = L_n^{NC} + s\big(L_n - L_n^{NC}\big), \qquad s \in \{0, 0.25, 0.5, 0.75, 1\}. \tag{110}$$

For each interpolation point, we use four pCN chains with 400 warm-up steps and 2,000 retained draws per chain. The heat-capacity diagnostic is computed from $t^2 \operatorname{Var}_{P_{t,s}}[L_s]$ along this interpolation.

D.1.2 Bayesian Generalization Experiment

We use a realizable two-parameter tied-weight polynomial neural network with trainable parameters (u, v) and Gaussian observation noise with standard deviation $\sigma = 0.2$. The population excess-risk germ is

$$K(u, v) = \frac{u^2 + \frac{2}{3}v^4 + \frac{2}{3}v^6}{2\sigma^2}, \tag{111}$$

while the NC-projected germ is

$$K_{NC}(u, v) = \frac{u^2 + \frac{2}{3}v^4}{2\sigma^2}. \tag{112}$$

Both germs have the known RLCT

$$\lambda = \frac{3}{4}. \tag{113}$$

The Bayes predictive generalization error G_n is evaluated at unit Bayesian temperature. In contrast, the LLC comparison uses the tempered sequence

$$\beta_n = \frac{1}{\log n}, \qquad t_n = n\beta_n = \frac{n}{\log n}. \tag{114}$$

Thus, we compare the scaled Bayes predictive quantity nG_n with the tempered actual and NC-projected LLCs against the known reference coefficient $3/4$. Tempering is applied only to the LLC diagnostic and does not modify the unit-temperature Bayes predictive distribution.

The final confirmatory experiment uses 1,500 fresh independent datasets at eight pre-specified sample sizes,

$$n \in \{128, 512, 2048, 8192, 32768, 131072, 524288, 1048576\}. \tag{115}$$

Error bars denote Student-t 95% confidence intervals for the mean across datasets.

D.1.3 Controlled NC-Breaking Experiment

We compare an exact-NC reference, NC-preserving controls, and anisotropic NC-breaking perturbations. For each condition, we evaluate the exact empirical loss along a one-dimensional transported perturbation coordinate s. The integration domain is

$$s \in [-r, r], \qquad r \in \{0.10, 0.30, 0.50\}, \tag{116}$$

and the inverse-temperature grid consists of 33 logarithmically spaced values between

$$t = 10^2 \qquad \text{and} \qquad t = 10^{10}. \tag{117}$$

The one-dimensional partition integrals are evaluated using Gauss–Legendre quadrature, with 4,096 nodes per radius 0.1. To identify the local vanishing order, we fit the loss in a small neighborhood of $s = 0$ to the even polynomial

$$a_2 s^2 + a_4 s^4 + a_6 s^6 + a_8 s^8. \tag{118}$$

D.2 Robustness of the ϵ-NC Bound

We evaluate the ϵ-NC stability relation over a sensitivity grid of independently trained PNN checkpoints. We use three MNIST PNN training seeds, $\{101, 202, 303\}$, and retain checkpoints at epochs $\{20, 100, 200\}$. The sensitivity grid varies inverse temperature, localization strength, and neighborhood radius as

$$t \in \{25, 50, 100, 200, 400\}, \qquad \gamma \in \{2 \times 10^4, 4 \times 10^4, 8 \times 10^4\}, \qquad r \in \{2, 3, 4\}. \tag{119}$$

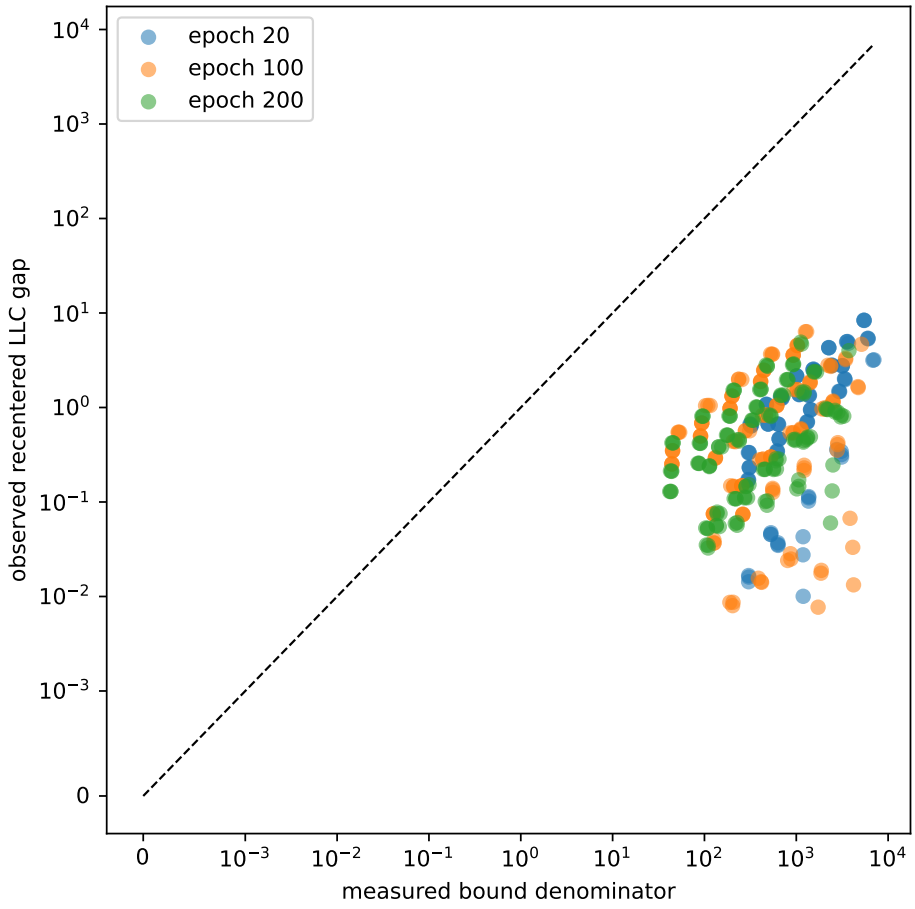


Figure 5: Observed recentered LLC gaps versus the measured finite-grid bound denominator across reliable PNN sensitivity-grid cells. The dashed line denotes equality. All reliable cells lie below the line, indicating that the measured inequality is satisfied with a substantial margin over the tested settings. The denominator is based on sampled interpolation and neighborhood diagnostics and should not be interpreted as a certified continuous supremum.

This gives $3 \times 3 \times 5 \times 3 \times 3 = 405$ experimental cells. For each cell, the actual and NC-projected losses are minimized separately within the corresponding radius-r Euclidean neighborhood. We interpolate between the projected and actual losses using

$$L_s = L_n^{NC} + s\big(L_n - L_n^{NC}\big), \qquad s \in \{0, 0.25, 0.5, 0.75, 1\}. \tag{120}$$

At each interpolation point, we run four pCN chains with 400 warm-up steps and 2,000 retained draws per chain. The actual and projected LLC estimates are recentered at their respective constrained minima,

$$\widehat{\lambda}_t = t\left(\mathbb{E}_{P_{t,1}}[L_n] - m_U\right), \qquad \widehat{\lambda}_t^{NC} = t\left(\mathbb{E}_{P_{t,0}}[L_n^{NC}] - m_U^{NC}\right). \tag{121}$$

We estimate the interpolation heat-capacity factor on the measured grid as

$$C_H^{\mathrm{grid}} = \max_{s \in \{0,0.25,0.5,0.75,1\}} t^2 \operatorname{Var}_{P_{t,s}}[L_s]. \tag{122}$$

For each cell, we evaluate the empirical finite-grid ratio

$$\rho = \frac{\left|\widehat{\lambda}_t - \widehat{\lambda}_t^{NC}\right|}{\left(2 + \sqrt{C_H^{\mathrm{grid}}}\right) t \Delta_{NC}^{\mathrm{grid}}}. \tag{123}$$

Here, $\Delta_{NC}^{\mathrm{grid}}$ denotes the largest measured pointwise Δ_{NC} among the checkpoint, the estimated actual and projected minima, and every twentieth retained draw. The quantity $\Delta_{NC}^{\mathrm{grid}}$ is a sampled finite-grid diagnostic and is not identified with the continuous neighborhood supremum $\overline{\Delta}_{NC}(U)$ appearing in the theoretical bound. The quadratic localization term is $\gamma\|\theta - \theta_0\|^2/2$ and is not multiplied by t. Sampler proposal scales and reliability criteria were fixed before inspecting the resulting bound ratios. We report only cells satisfying the pre-specified local-minimum and scalar MCMC reliability criteria in the primary comparison.

Across 362 reliable cells in the 405-cell PNN sensitivity grid, every measured recentered LLC gap remained below the corresponding finite-grid ϵ-NC stability bound, with a maximum ratio of $\rho = 0.0106$ (Figure 5). The substantial margin indicates that the measured bound is conservative over the tested temperatures, localization strengths, and neighborhood radii. The ratio magnitude varied strongly with localization, and numerical reliability decreased at the highest temperature, so the experiment does not establish localizer-independent behavior. Moreover, the finite interpolation and sampled-neighborhood diagnostics do not certify the continuous heat-capacity and neighborhood-supremum assumptions of the theorem.

D.3 Low-Cost Training-Regime Monitoring in Transformers

We investigate whether dense NC-derived statistics can serve as inexpensive indicators of training-regime changes in modular-arithmetic Transformers. We use five independent training seeds, $\{101, 202, 303, 404, 505\}$, with the same architecture, optimization procedure, data split, and 10,000-epoch training horizon. Training and test accuracies and NC statistics are evaluated every 20 epochs. For each seed, the grokking transition epoch t_g is defined independently of the NC statistics as the first measured epoch satisfying

$$\mathrm{Acc}_{\mathrm{train}} \geq 0.99, \qquad \mathrm{Acc}_{\mathrm{test}} \geq 0.90. \tag{124}$$

We compute dense trajectories of ϵ_{NC1}, ϵ_{NC2}, and Δ_{NC} throughout training. Because the Transformer is trained with cross-entropy, these quantities are used only as representation-geometry monitoring statistics and are not interpreted as instances of the MSE LLC bound. We use Δ_{NC} as the primary monitoring statistic and treat ϵ_{NC1} and ϵ_{NC2} as secondary decomposition diagnostics. For each metric M_t, we first compute

$$x_t = \log_{10}(M_t + 10^{-12}). \tag{125}$$

We then construct a one-sided causal exponential moving average with span five checkpoints. We denote the resulting causal smoothed trajectory by s_t and define the signed and absolute one-step changes as

$$d_t = s_t - s_{t-1}, \qquad q_t = |d_t|. \tag{126}$$

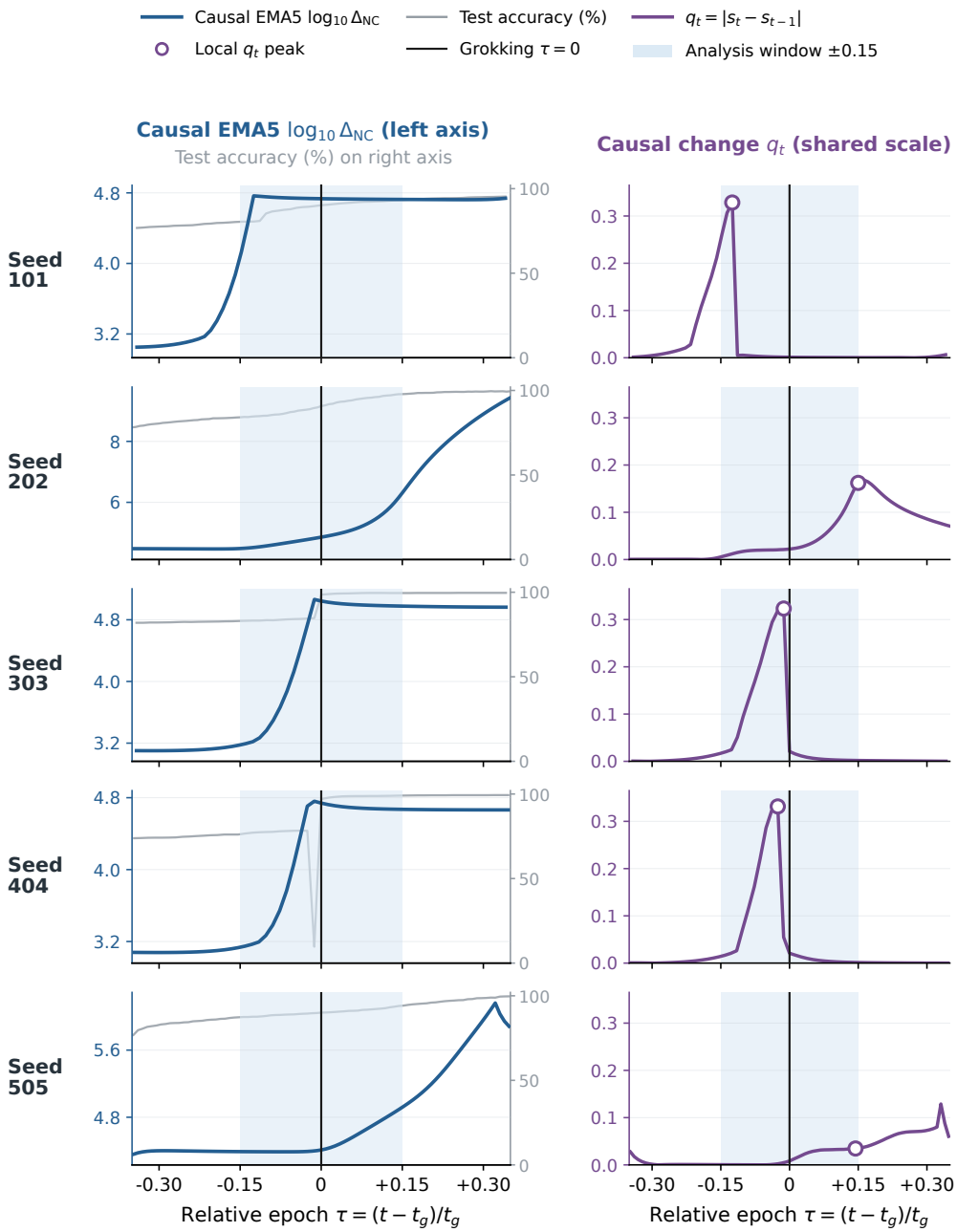


Figure 6: Causal NC monitoring around the grokking transition in five independent modular-arithmetic Transformer runs. Each row shows test accuracy, the causal EMA-smoothed $\log_{10}\Delta_{NC}$ trajectory, and the corresponding one-sided change magnitude $q_t = |s_t - s_{t-1}|$. The vertical line marks the independently defined grokking transition, the shaded region denotes the primary transition window, and the open circle marks the largest local q_t within that window. Pronounced transition-local changes occur in several runs, whereas other runs exhibit broader or weaker local activity. These trajectories are interpreted as representation-geometry monitoring signals rather than as LLC estimates.

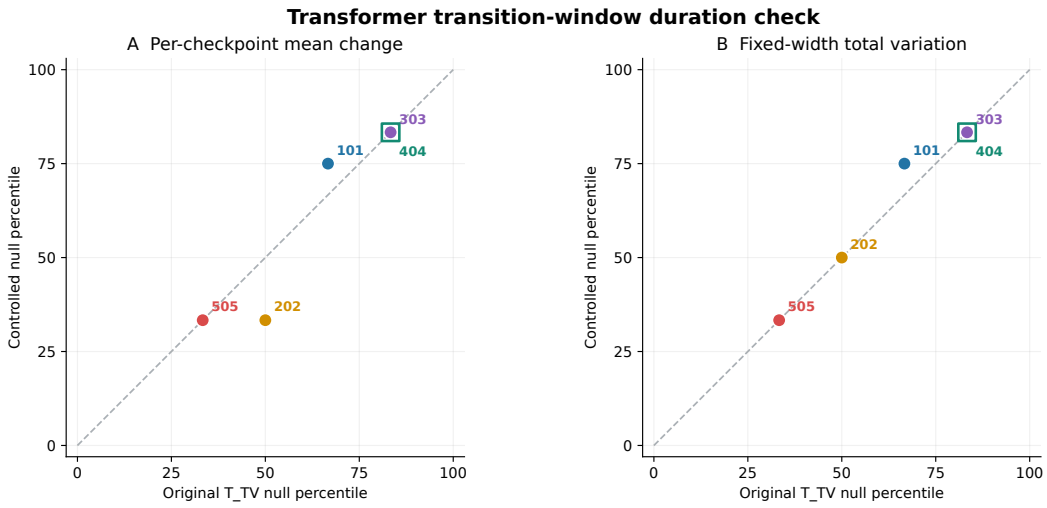


Figure 7: Robustness of transition-local NC activity to unequal pseudo-window duration. The horizontal axis gives the empirical percentile of the original T_{TV} relative to twelve off-transition pseudo-centers, while the vertical axis gives the corresponding percentile after correcting for window duration. Panel A uses the per-checkpoint statistic T_{mean}, and Panel B recomputes total variation over fixed-duration windows. Seeds 303 and 404 retain the highest transition-local ranks under both corrections, with seed 101 remaining the next clearest case, whereas seeds 202 and 505 remain comparatively non-specific or weak. The qualitative seed-level pattern is therefore not explained by the varying numbers of observations in the original normalized pseudo-windows.

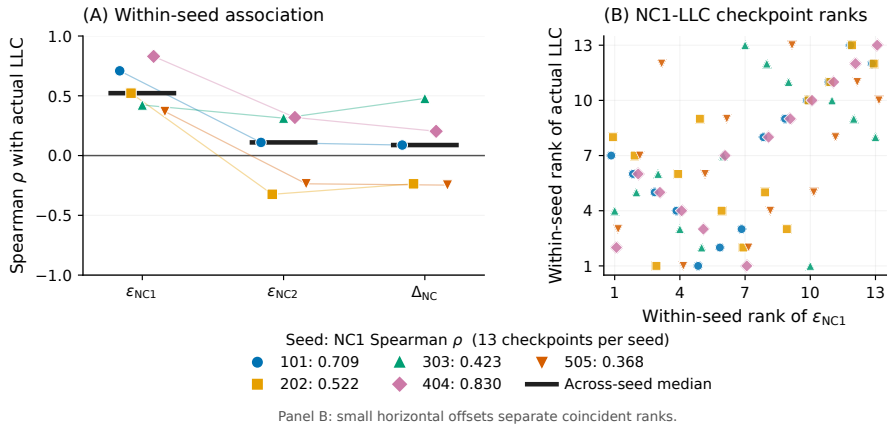


Figure 8: Checkpoint-level association between deterministic NC statistics and finite-temperature LLC across five Transformer runs. Panel A reports within-seed Spearman correlations over thirteen transition-aligned checkpoints per seed, with black horizontal segments indicating the across-seed medians. ϵ_{NC1} is positively associated with actual LLC in all five runs, with seed-wise Spearman correlations ranging from 0.368 to 0.830 and an across-seed median of 0.522, whereas ϵ_{NC2} and Δ_{NC} exhibit heterogeneous associations across seeds. Panel B shows the corresponding within-seed checkpoint ranks for ϵ_{NC1} and actual LLC. These comparisons characterize empirical checkpoint-level association and do not treat the NC statistics as numerical estimators of LLC.

The primary transition neighborhood is defined in normalized training time as

$$\left| \frac{t - t_g}{t_g} \right| \leq 0.15. \tag{127}$$

We additionally use windows of 0.10 and 0.20 as sensitivity checks. Within each transition window, we measure the largest local change and the accumulated local change as

$$T_{\text{peak}} = \max_{t \in W(t_g)} q_t, \qquad T_{\text{TV}} = \sum_{t \in W(t_g)} q_t. \tag{128}$$

We define the transition-local event time t_{local} as the epoch attaining T_{peak} within the prescribed window. We summarize its timing relative to grokking using the normalized lag

$$\delta_{\text{local}} = \frac{t_{\text{local}} - t_g}{t_g}. \tag{129}$$

Negative values of δ_{local} indicate that the NC change precedes the grokking threshold, whereas positive values indicate that it follows the threshold. To assess whether transition-local activity is unusual relative to other stages of training, we compare the true transition window with twelve preselected off-transition pseudo-centers per seed. For every pseudo-center, we apply the same causal signal construction and the same normalized timing window. We compare the observed T_{peak} and T_{TV} values with their corresponding pseudo-transition distributions using empirical ranks and exceedance counts. Because normalized pseudo-transition windows have absolute widths proportional to their pseudo-centers, their numbers of included checkpoints can differ. We therefore use two duration controls for T_{TV}. First, we define the per-checkpoint activity

$$T_{\text{mean}} = \frac{1}{|W|} \sum_{t \in W} q_t. \tag{130}$$

Second, within each seed we fix the absolute half-width to that of the true transition window, $H = 0.15 t_g$, and define

$$W_{\text{fixed}}(b) = \{t : |t - b| \leq H\}, \qquad T_{\text{TV}}^{\text{fixed}}(b) = \sum_{t \in W_{\text{fixed}}(b)} q_t, \tag{131}$$

where b denotes either the true transition or an admissible pseudo-transition center. These controls distinguish unusually concentrated NC activity from an apparent increase caused only by a longer observation window. We also evaluate smoothing spans of three and seven checkpoints to assess sensitivity to the temporal smoothing scale. Sparse finite-temperature LLC measurements provide a separate contextual comparison and are not used to define the monitoring events or their success criteria. For each seed, we use thirteen preselected transition-aligned checkpoints and compare the finalized actual LLC estimates with the simultaneously measured NC statistics. Because between-seed offsets can induce misleading pooled associations, we compute checkpoint-level Spearman correlations separately within each seed and summarize their cross-seed consistency rather than pooling all 65 observations into a single coefficient. All 65 LLC measurements used in this comparison passed the confirmatory scalar and local-minimum reliability checks.

Figure 6 shows the causal Δ_{NC} trajectories and transition-local changes for each seed. Pronounced local changes occur near the grokking transition in seeds 101, 303, and 404, whereas the remaining runs exhibit broader or weaker activity. Because normalized pseudo-transition windows vary in absolute duration, we additionally repeat the integrated-change comparison using per-observation and fixed-duration statistics. Figure 7 shows that this correction preserves the same qualitative seed-level pattern. The sparse LLC comparison reveals a more specific relationship between representation geometry and finite-temperature local complexity. As shown in Figure 8, ϵ_{NC1} has a positive within-seed rank association with actual LLC in all five runs, with Spearman correlations of 0.709, 0.522, 0.423, 0.830, and 0.368 for seeds 101, 202, 303, 404, and 505, respectively. The across-seed median correlation is 0.522. By contrast, the associations of ϵ_{NC2} and Δ_{NC} with LLC are heterogeneous and change sign across seeds. Thus the experiment identifies a reproducible monotonic association between within-class collapse and finite-temperature LLC at the sampled checkpoints, but does not support interpreting the combined Δ_{NC} statistic as an LLC estimator or calibrated proxy.

Across five independent grokking runs, causal Δ_{NC} trajectories exhibited pronounced transition-local changes in three seeds. Seeds 303 and 404 showed the highest integrated transition-local activity

relative to the descriptive pseudo-transition controls, while seed 101 exhibited a sharp local change with somewhat weaker accumulated activity. One additional run showed a broader change spanning the transition but weak transition specificity, whereas the remaining run showed only weak transition-local activity. The same qualitative seed-level pattern persists after normalizing for the number of observations and after imposing fixed-duration comparison windows. Separately, sparse reliable LLC measurements show a consistent positive checkpoint-level rank association with ϵ_{NC1} across all five seeds, but not with ϵ_{NC2} or the combined Δ_{NC} statistic. Taken together, these results show that dense NC statistics can expose representation-level training changes at substantially finer temporal resolution than sparse LLC sampling, while their relationship to finite-temperature local complexity is metric- and seed-dependent. They therefore motivate NC statistics as inexpensive representation-geometry monitoring signals rather than as unique detectors of grokking or numerical estimators of LLC. Additional NC events occur elsewhere during training, and the transition-local comparisons use only twelve descriptive pseudo-transition windows per seed. Accordingly, the experiment establishes neither a unique grokking signature nor a statistically calibrated online grokking detector.