Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Conditional Confinement Criteria and Empirical Diagnostics for Quadratically Regularized Transformer Objectives

Sayantan Dutta

sayantan.dutta1@gehealthcare.com

GE HealthCare

Abhijit Das

GE HealthCare

Article

Keywords:

Posted Date: November 17th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11379021/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Conditional Confinement Criteria and Empirical Diagnostics for Quadratically Regularized Transformer Objectives

Sayantan Dutta[1*] and Abhijit Das[1]

[1]Science and Technology Organization, GE HealthCare, Bangalore, 560066, Karnataka, India.

*Corresponding author(s). E-mail(s): sayantan.dutta1@gehealthcare.com;

Abstract

Coupled L^2 regularization makes neural-network objectives with non-negative data-fitting losses coercive and their Gibbs densities normalizable, but these properties do not guarantee differential confinement of the parameter-space derivatives. This study establishes this distinction for Transformer objectives. A bilinear-logit counterexample shows that bounded inputs and positive quadratic regularization are insufficient to ensure $\Psi_{s,\lambda}(\theta) = -\Delta\mathcal{F}_\lambda(\theta) + s^{-1}\|\nabla\mathcal{F}_\lambda(\theta)\|^2 \to +\infty$ in every parameter-space direction. We derive a sufficient criterion that balances the asymptotic growth of the data-loss gradient and Hessian trace against the regularization strength and temperature. We then develop a scalable finite-directional diagnostic based on Hessian–vector products and stochastic trace estimation. Experiments with GPT-Neo-125M on Penn Treebank and WikiText-103 reveal substantial variation across directions, checkpoints, temperatures, and regularization implementations. Within the investigated near-reference regime, the negative data-loss Hessian-trace contribution and the constant regularizer trace dominate the positive radial term, and the measured diagnostic does not exhibit uniform quadratic growth across the sampled directions. The results separate global mathematical confinement from finite directional evidence and provide a reproducible framework for investigating asymptotic loss geometry in large Transformer models.

Keywords: Transformer optimization, Coupled L^2 regularization, Differential confinement, Loss geometry, Hessian trace estimation, Langevin dynamics

1 Introduction

Transformers are foundational architectures for natural-language processing, computer vision, and speech processing [3, 4, 12]. Their parameter-space geometry nevertheless remains difficult to characterize. Attention, projection, and feed-forward operators contain multiplicative parameter interactions, while residual and normalization operations introduce additional dependencies. The resulting objectives are high-dimensional, non-convex, and not amenable to global conclusions based solely on bounded model inputs.

Coupled L^2 regularization augments a data-fitting loss $\mathcal{L}(\theta)$ with a quadratic penalty:

$$\mathcal{F}_\lambda(\theta) = \mathcal{L}(\theta) + \frac{\lambda}{2}\|\theta\|^2, \qquad \theta \in \mathbb{R}^d, \quad \lambda > 0. \tag{1}$$

Although coupled L^2 regularization and weight decay coincide for standard stochastic gradient descent under an appropriate update convention, AdamW applies decay separately from the adaptively preconditioned loss gradient [7, 8]. The theoretical results in this work concern the explicit objective in Eq. (1); coupled L^2 regularization and decoupled AdamW weight decay are compared separately in the experiments.

Because cross-entropy is non-negative, the quadratic penalty makes $\mathcal{F}_\lambda$ coercive and yields a Gaussian upper bound on its Gibbs weight. These properties ensure normalizability of the associated Gibbs density and are relevant to Langevin and stochastic-gradient Langevin analyses [11, 13, 14]. Coercivity, however, controls function values rather than the parameter derivatives required by stronger functional-analytic conditions.

In particular, coercivity does not imply the differential confinement condition

$$\Psi_{s,\lambda}(\theta) := -\Delta\mathcal{F}_\lambda(\theta) + \frac{1}{s}\|\nabla\mathcal{F}_\lambda(\theta)\|^2 \longrightarrow +\infty \quad \text{as} \quad \|\theta\| \longrightarrow \infty, \tag{2}$$

where $s > 0$, ∇ is the Euclidean parameter gradient, and Δ is the parameter-space Laplacian. The regularization-induced gradient must dominate the data-loss Hessian trace and any adverse interaction between $\nabla\mathcal{L}(\theta)$ and θ. Bounded token embeddings constrain the inputs but do not uniformly bound parameter derivatives: multiplicative parameter interactions can cause both the logit Jacobian and the Hessian trace to grow with $\|\theta\|$.

This issue is relevant to attention models because their trainable matrices interact multiplicatively. Regularization can induce low-rank structure in such matrix products [6], but that structural effect does not imply Eq. (2). Differential confinement additionally requires control of the asymptotic data-loss gradient and Hessian trace.

Related results analyze Langevin dynamics in mean-field formulations of neural-network training. Chizat established exponential convergence under suitable log-Sobolev conditions, while Nitanda *et al.* derived propagation-of-chaos guarantees for finite-particle approximations [2, 9]. These results operate in spaces of probability measures and do not directly establish parameter-space confinement for finite-width Transformers. The present study instead examines the Euclidean geometry of a finite-dimensional, explicitly regularized Transformer objective.

The distinction between coercivity and differential confinement leads to the central question considered in this work:

> *Under what parameter-space growth conditions does coupled quadratic regularization imply differential confinement, and what aspects of this condition can be investigated computationally in a large Transformer?*

We first show, through a bilinear-logit counterexample, that bounded inputs and positive regularization are insufficient. We then derive a sufficient criterion that explicitly relates the growth of $\|\nabla\mathcal{L}(\theta)\|$ and $\Delta\mathcal{L}(\theta)$ to λ and s.

The contributions of this study are threefold:

1. We separate quadratic coercivity and Gibbs integrability from differential confinement, and construct an explicit counterexample showing that the former properties do not imply the latter.
2. We derive a sufficient confinement criterion that quantifies the competition among data-loss gradient growth, Hessian-trace growth, regularization strength, and temperature.
3. We develop a scalable finite-directional estimator based on Hessian–vector products and stochastic trace estimation, and use it to study directional, checkpoint-level, temperature-dependent, spectral, and parameter-group behavior in GPT-Neo-125M and a ViT-B/16 extension. The measurements are interpreted as finite geometric evidence rather than as a global asymptotic certificate.

The remainder of the article is organized as follows. Section 2 introduces the regularized Transformer objective, presents the counterexample, and derives the conditional confinement criterion. Section 3 develops the geometric interpretation and computational estimation of the differential diagnostic. Section 4 reports the experimental protocol and results. Section 5 discusses the theoretical and empirical implications of the findings. Section 6 summarizes the limitations of the present analysis and outlines directions for future work. Section 7 provides the methodological, implementation, statistical, and reproducibility details. Finally, Section 8 summarizes the principal conclusions and broader significance of the proposed framework.

2 Theoretical Framework

This section distinguishes coercivity, Gibbs integrability, and differential confinement for the coupled L^2-regularized objective. The quadratic penalty guarantees the first two properties without global derivative bounds. Differential confinement is stronger: it depends on the asymptotic competition among the data-loss gradient, Hessian trace, regularization strength, and temperature. After fixing the notation, we establish the unconditional properties, construct a counterexample to unconditional differential confinement, and derive a sufficient criterion under explicit derivative-growth assumptions.

2.1 Notation and parameterization

Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ denote a dataset containing N input–target pairs. Each x_i denotes an input token sequence and y_i denotes the corresponding target token or target sequence, depending on the training convention. The vocabulary size is denoted by V.

Consider a Transformer with L layers. The model parameters include the token and positional embeddings, attention projections, feed-forward weights, normalization parameters, output projections, and their associated biases whenever these quantities are trainable. We collect all trainable parameters into the vector

$$\theta = \operatorname{vec}(\theta_1, \theta_2, \ldots, \theta_L, \theta_{\text{emb}}, \theta_{\text{out}}) \in \mathbb{R}^d, \tag{3}$$

where $\theta_\ell \in \mathbb{R}^{d_\ell}$ contains the parameters of the ℓth Transformer layer, θ_{emb} contains the embedding parameters, and θ_{out} contains the output-layer parameters. The total number of trainable scalar parameters is $d = \sum_{\ell=1}^{L} d_\ell + d_{\text{emb}} + d_{\text{out}}$. If input and output embeddings are tied, their shared parameters are counted only once.

For a vector $u \in \mathbb{R}^d$, $\|u\|$ denotes the Euclidean norm. For a matrix A, $\|A\|_F$, $\|A\|_2$, and $\operatorname{tr}(A)$ denote the Frobenius norm, spectral norm, and trace, respectively. The $d \times d$ identity matrix is denoted by I_d. For a twice continuously differentiable scalar function $f : \mathbb{R}^d \to \mathbb{R}$, we use

$$\nabla f(\theta) = \left(\frac{\partial f}{\partial \theta_1}, \ldots, \frac{\partial f}{\partial \theta_d} \right)^\top \tag{4}$$

for its gradient,

$$\nabla^2 f(\theta) = \left[\frac{\partial^2 f}{\partial \theta_j \, \partial \theta_k} \right]_{j,k=1}^{d} \tag{5}$$

for its Hessian, and

$$\Delta f(\theta) = \operatorname{tr}\left[\nabla^2 f(\theta) \right] = \sum_{j=1}^{d} \frac{\partial^2 f(\theta)}{\partial \theta_j^2} \tag{6}$$

for its parameter-space Laplacian.

The notation $\theta^{(0)}$ denotes a reference parameter vector, such as the initialization or a trained checkpoint. A directional path beginning at $\theta^{(0)}$ is written as

$$\theta(r; u) = \theta^{(0)} + ru, \qquad r \geq 0, \qquad \|u\| = 1, \tag{7}$$

where $u \in \mathbb{R}^d$ is a unit direction and r is the radial displacement. Directional evaluation along paths of this form is used later as a computational probe. Observing a property along finitely many directions does not establish the corresponding global limit over $\mathbb{R}^d$.

2.2 Transformer forward map and cross-entropy loss

For an input sequence x, the Transformer produces a vector of logits

$$z_\theta(x) = (z_{1,\theta}(x), \dots, z_{V,\theta}(x))^\top \in \mathbb{R}^V. \tag{8}$$

The probability assigned to token $v \in \{1, \dots, V\}$ is

$$p_\theta(v \mid x) = \frac{\exp(z_{v,\theta}(x))}{\sum_{k=1}^{V} \exp(z_{k,\theta}(x))}. \tag{9}$$

For a target token y_i, the sample-wise cross-entropy loss is

$$\ell_i(\theta) = -\log p_\theta(y_i \mid x_i) = -z_{y_i,\theta}(x_i) + \log \left[\sum_{k=1}^{V} \exp(z_{k,\theta}(x_i)) \right]. \tag{10}$$

The empirical data-fitting objective is

$$\mathcal{L}(\theta) = \frac{1}{N} \sum_{i=1}^{N} \ell_i(\theta). \tag{11}$$

Throughout the theoretical analysis, we assume that the model uses smooth activation functions and that each normalization operation contains a fixed stabilization constant $\varepsilon_{\mathrm{LN}} > 0$ in its denominator. Under these conditions, the normalization denominator remains strictly positive. For every fixed finite input sequence, the resulting parameter-to-logit map and the cross-entropy loss are twice continuously differentiable with respect to the trainable parameters.

This smoothness assumption concerns differentiability and should not be confused with global boundedness of the derivatives. A smooth function can have a gradient, Hessian, or Hessian trace that grows without bound as $\|\theta\| \to \infty$.

2.3 Quadratically regularized objective

We study the coupled L^2-regularized objective

$$\mathcal{F}_\lambda(\theta) = \mathcal{L}(\theta) + \mathcal{R}_\lambda(\theta), \qquad \mathcal{R}_\lambda(\theta) = \frac{\lambda}{2}\|\theta\|^2, \qquad \lambda > 0. \tag{12}$$

The derivatives of the quadratic component are

$$\nabla \mathcal{R}_\lambda(\theta) = \lambda\theta, \qquad \nabla^2 \mathcal{R}_\lambda(\theta) = \lambda I_d, \qquad \Delta \mathcal{R}_\lambda(\theta) = \lambda d.$$

Consequently,

$$\begin{aligned} \nabla \mathcal{F}_\lambda(\theta) &= \nabla \mathcal{L}(\theta) + \lambda\theta, \\ \nabla^2 \mathcal{F}_\lambda(\theta) &= \nabla^2 \mathcal{L}(\theta) + \lambda I_d, \end{aligned}$$

$$\Delta\mathcal{F}_\lambda(\theta) = \Delta\mathcal{L}(\theta) + \lambda d. \tag{13}$$

Equation (13) shows that, at a fixed parameter vector, coupled L^2 regularization shifts every Hessian eigenvalue by the same amount λ. Denoting the ordered Hessian eigenvalues by $\rho_j(\cdot)$, this shift implies

$$\rho_j\big(\nabla^2\mathcal{F}_\lambda(\theta)\big) = \rho_j\big(\nabla^2\mathcal{L}(\theta)\big) + \lambda, \qquad j = 1, \dots, d, \tag{14}$$

Any additional spectral differences observed between independently trained models must therefore also reflect changes in the learned parameter vector and in the Hessian of the data-fitting loss. Such differences are not produced solely by the instantaneous isotropic term λI_d.

The mathematical objective in Eq. (12) corresponds to coupled L^2 regularization. Decoupled weight decay under an adaptive optimizer generally produces a different parameter update. The theoretical results below therefore apply directly to Eq. (12); coupled regularization and decoupled weight decay are treated separately in the experiments.

2.4 Smoothness, coercivity, and Gibbs integrability

We begin with properties that follow directly from the quadratic penalty and do not require global bounds on $\nabla\mathcal{L}(\theta)$ or $\nabla^2\mathcal{L}(\theta)$.

Proposition 2.1 (Smoothness of the regularized objective) *Suppose that $\mathcal{L} \in C^2(\mathbb{R}^d)$. Then $\mathcal{F}_\lambda \in C^2(\mathbb{R}^d)$ for every $\lambda > 0$. If $\mathcal{L} \in C^\infty(\mathbb{R}^d)$, then $\mathcal{F}_\lambda \in C^\infty(\mathbb{R}^d)$.*

Proof The quadratic function $\mathcal{R}_\lambda(\theta) = \lambda\|\theta\|^2/2$ is a polynomial and is therefore infinitely differentiable. The conclusion follows because sums of functions in C^2, or in C^∞, remain in the corresponding function class. □

Proposition 2.2 (Quadratic coercivity) *For every $\lambda > 0$,*

$$\mathcal{F}_\lambda(\theta) \geq \frac{\lambda}{2}\|\theta\|^2. \tag{15}$$

Consequently,

$$\lim_{\|\theta\|\to\infty} \mathcal{F}_\lambda(\theta) = +\infty. \tag{16}$$

Proof For a softmax model, $p_\theta(y_i \mid x_i) \in (0, 1]$, and hence $\ell_i(\theta) = -\log p_\theta(y_i \mid x_i) \geq 0$. It follows that $\mathcal{L}(\theta) \geq 0$. Therefore,

$$\mathcal{F}_\lambda(\theta) = \mathcal{L}(\theta) + \frac{\lambda}{2}\|\theta\|^2 \geq \frac{\lambda}{2}\|\theta\|^2.$$

The right-hand side tends to $+\infty$ as $\|\theta\| \to \infty$. □

For an inverse-temperature parameter $\beta > 0$, define the partition function

$$Z_{\beta,\lambda} = \int_{\mathbb{R}^d} \exp\left[-\beta \mathcal{F}_\lambda(\theta)\right] d\theta. \tag{17}$$

Whenever $Z_{\beta,\lambda} < \infty$, the corresponding Gibbs probability measure is

$$\mu_{\beta,\lambda}(d\theta) = Z_{\beta,\lambda}^{-1} \exp\left[-\beta \mathcal{F}_\lambda(\theta)\right] d\theta. \tag{18}$$

Proposition 2.3 (Gaussian-dominated Gibbs integrability) *For every $\beta > 0$ and $\lambda > 0$,*

$$Z_{\beta,\lambda} \leq \left(\frac{2\pi}{\beta\lambda}\right)^{d/2} < \infty. \tag{19}$$

Therefore, $\mu_{\beta,\lambda}$ in Eq. (18) is a well-defined probability measure.

Proof By Proposition 2.2,

$$\exp\left[-\beta \mathcal{F}_\lambda(\theta)\right] \leq \exp\left[-\frac{\beta\lambda}{2}\|\theta\|^2\right].$$

Integrating both sides over $\mathbb{R}^d$ gives

$$Z_{\beta,\lambda} \leq \int_{\mathbb{R}^d} \exp\left[-\frac{\beta\lambda}{2}\|\theta\|^2\right] d\theta = \left(\frac{2\pi}{\beta\lambda}\right)^{d/2}.$$

□

Propositions 2.2 and 2.3 establish that positive quadratic regularization supplies global confinement at the level of function values and makes the associated Gibbs density normalizable. Neither result controls the asymptotic behavior of the derivatives of $\mathcal{L}$.

2.5 Differential confinement diagnostic

For $s > 0$, define the scalar field

$$\Psi_{s,\lambda}(\theta) = -\Delta \mathcal{F}_\lambda(\theta) + \frac{1}{s}\left\|\nabla \mathcal{F}_\lambda(\theta)\right\|^2. \tag{20}$$

When λ is clear from the context, we write $\Psi_s(\theta)$.

The differential confinement condition considered in this study is

$$\lim_{\|\theta\| \to \infty} \Psi_{s,\lambda}(\theta) = +\infty. \tag{21}$$

This is a global condition: it must hold along every sequence $\{\theta_n\}_{n \geq 1}$ satisfying $\|\theta_n\| \to \infty$. Equivalently, for every $M > 0$ there must exist $R_M > 0$ such that

$$\|\theta\| \geq R_M \quad \Longrightarrow \quad \Psi_{s,\lambda}(\theta) \geq M. \tag{22}$$

Consequently, behavior observed along finitely many directions or over a bounded radial interval cannot establish Eq. (21).

Using Eq. (13), the diagnostic decomposes as

$$\begin{aligned}\Psi_{s,\lambda}(\theta) &= -\Delta\mathcal{L}(\theta) - \lambda d + \frac{1}{s}\left\|\nabla\mathcal{L}(\theta) + \lambda\theta\right\|^2 \\ &= -\Delta\mathcal{L}(\theta) - \lambda d + \frac{1}{s}\left\|\nabla\mathcal{L}(\theta)\right\|^2 + \frac{2\lambda}{s}\left\langle\nabla\mathcal{L}(\theta), \theta\right\rangle + \frac{\lambda^2}{s}\|\theta\|^2. \end{aligned} \tag{23}$$

The final term in Eq. (23) is a positive quadratic contribution from the regularizer. It does not, however, automatically dominate the other terms. In particular:

1. the inner product $\langle\nabla\mathcal{L}(\theta), \theta\rangle$ can be negative and can grow in magnitude with $\|\theta\|$;
2. the positive part of $\Delta\mathcal{L}(\theta)$ enters $\Psi_{s,\lambda}$ with a negative sign; and
3. the relative importance of the squared-gradient term decreases when s increases.

Consequently, coercivity of $\mathcal{F}_\lambda$ does not by itself imply Eq. (21).

2.6 Why bounded inputs are insufficient

Let $h_i^{(0)}$ denote the embedded representation of the input sequence x_i supplied to the first Transformer layer. Suppose that

$$\|h_i^{(0)}\| \le B \qquad \text{for all } i \tag{24}$$

for some finite $B > 0$. This condition bounds the network inputs but does not uniformly bound the parameter-to-logit Jacobian

$$J_z(\theta; x) = \frac{\partial z_\theta(x)}{\partial\theta}. \tag{25}$$

Indeed, by the chain rule, the gradient of a sample-wise softmax cross-entropy loss can be written as

$$\nabla\ell_i(\theta) = J_z(\theta; x_i)^\top \left[p_\theta(\cdot \mid x_i) - e_{y_i}\right], \tag{26}$$

where $e_{y_i} \in \mathbb{R}^V$ is the one-hot vector associated with the target class. Although the probability residual $p_\theta(\cdot \mid x_i) - e_{y_i}$ is bounded, the Jacobian $J_z(\theta; x_i)$ need not be uniformly bounded in θ. Multiplicative parameterizations provide a simple illustration: if a logit contains a product of two trainable parameters, its derivative with respect to one parameter contains the other parameter and can grow without bound.

The same issue affects the Hessian trace. Bounded inputs do not rule out quadratic or higher-order growth of $\Delta\mathcal{L}(\theta)$ along selected directions. Therefore, Eq. (24) is insufficient to justify constants $C_1, C_2 < \infty$ satisfying

$$\|\nabla\mathcal{L}(\theta)\| \le C_1, \qquad \Delta\mathcal{L}(\theta) \le C_2 \qquad \text{for every } \theta \in \mathbb{R}^d. \tag{27}$$

2.7 A bilinear-logit counterexample

The following example demonstrates that positive quadratic regularization and bounded inputs do not guarantee differential confinement for every $s > 0$.

Proposition 2.4 (Failure of unconditional differential confinement) *Consider a two-class model with parameter vector $\theta = (a, b)^\top \in \mathbb{R}^2$ and logits*

$$z_{1,\theta} = ab, \qquad z_{2,\theta} = 0, \tag{28}$$

for a training sample whose target class is 1. Let

$$\mathcal{L}(a, b) = \log\left(1 + \exp(-ab)\right) \tag{29}$$

and

$$\mathcal{F}_\lambda(a, b) = \mathcal{L}(a, b) + \frac{\lambda}{2}(a^2 + b^2), \qquad \lambda > 0. \tag{30}$$

Then $\mathcal{F}_\lambda$ is smooth, coercive, and Gibbs-integrable. However, along the ray $(a, b) = (t, 0)$,

$$\Psi_{s,\lambda}(t, 0) = \left[\frac{\lambda^2 + \frac{1}{4}}{s} - \frac{1}{4}\right] t^2 - 2\lambda. \tag{31}$$

Consequently,

$$\Psi_{s,\lambda}(t, 0) \longrightarrow -\infty \qquad \text{as } t \longrightarrow \infty \tag{32}$$

whenever

$$s > 1 + 4\lambda^2. \tag{33}$$

Proof Define

$$q(a, b) = \frac{1}{1 + \exp(ab)}. \tag{34}$$

The gradient of the data-fitting loss is

$$\nabla\mathcal{L}(a, b) = \begin{pmatrix} -bq(a, b) \\ -aq(a, b) \end{pmatrix}. \tag{35}$$

The trace of its Hessian is

$$\Delta\mathcal{L}(a, b) = (a^2 + b^2)q(a, b)\left[1 - q(a, b)\right]. \tag{36}$$

Along $(a, b) = (t, 0)$, we have $q(t, 0) = 1/2$, and therefore

$$\nabla\mathcal{F}_\lambda(t, 0) = \begin{pmatrix} \lambda t \\ -\frac{t}{2} \end{pmatrix}, \tag{37}$$

while

$$\Delta\mathcal{F}_\lambda(t, 0) = 2\lambda + \frac{t^2}{4}. \tag{38}$$

Substituting these expressions into Eq. (20) gives

$$\Psi_{s,\lambda}(t, 0) = -\left(2\lambda + \frac{t^2}{4}\right) + \frac{1}{s}\left(\lambda^2 t^2 + \frac{t^2}{4}\right) = \left[\frac{\lambda^2 + \frac{1}{4}}{s} - \frac{1}{4}\right] t^2 - 2\lambda. \tag{39}$$

The coefficient of t^2 is negative precisely when $s > 1 + 4\lambda^2$, proving the result. □

Proposition 2.4 isolates the central difficulty. The regularizer makes the function values grow quadratically, but the positive quadratic growth of the data-loss Laplacian enters the diagnostic with a negative sign. For sufficiently large s, the regularization-induced squared-gradient contribution does not dominate this term.

The example is intentionally minimal, but its multiplicative logit structure is relevant to neural architectures containing products of trainable matrices. The result does not establish that every Transformer fails differential confinement. Instead, it shows that additional parameter-space derivative assumptions are necessary before such a conclusion can be reached.

2.8 Conditional criterion for differential confinement

We now state sufficient conditions under which quadratic regularization does imply differential confinement. The assumptions explicitly describe the permitted asymptotic growth of the data-loss gradient and Laplacian.

Assumption 2.1 (Gradient-growth control) *There exist constants $A_0 \geq 0$ and $A_1 \geq 0$ such that*

$$\|\nabla \mathcal{L}(\theta)\| \leq A_0 + A_1 \|\theta\| \qquad \text{for every } \theta \in \mathbb{R}^d. \tag{40}$$

Assumption 2.2 (Upper control of the Hessian trace) *There exist constants $B_0 \geq 0$ and $B_1 \geq 0$ such that*

$$\Delta \mathcal{L}(\theta) \leq B_0 + B_1 \|\theta\|^2 \qquad \text{for every } \theta \in \mathbb{R}^d. \tag{41}$$

Only an upper bound on $\Delta \mathcal{L}$ is required because the diagnostic contains $-\Delta \mathcal{L}$. Large negative values of $\Delta \mathcal{L}$ increase $\Psi_{s,\lambda}$ and therefore do not obstruct the lower bound.

Theorem 1 (Sufficient condition for differential confinement) *Let $\mathcal{F}_\lambda$ be defined by Eq. (12), where $\mathcal{L} \in C^2(\mathbb{R}^d)$ satisfies Assumptions 2.1 and 2.2. Suppose that $\lambda > A_1$ and define*

$$\gamma_{s,\lambda} = \frac{(\lambda - A_1)^2}{s} - B_1. \tag{42}$$

If $\gamma_{s,\lambda} > 0$, then there exist finite constants $R_{s,\lambda} > 0$ and $C_{s,\lambda} > 0$ such that

$$\Psi_{s,\lambda}(\theta) \geq \frac{\gamma_{s,\lambda}}{2} \|\theta\|^2 - C_{s,\lambda} \qquad \text{for all } \|\theta\| \geq R_{s,\lambda}. \tag{43}$$

Consequently,

$$\Psi_{s,\lambda}(\theta) \longrightarrow +\infty \qquad \text{as} \qquad \|\theta\| \longrightarrow \infty.$$

Proof By the reverse triangle inequality and Assumption 2.1,

$$\|\nabla \mathcal{F}_\lambda(\theta)\| = \|\nabla \mathcal{L}(\theta) + \lambda\theta\| \geq \lambda\|\theta\| - \|\nabla \mathcal{L}(\theta)\| \geq (\lambda - A_1)\|\theta\| - A_0. \tag{44}$$

Because $\lambda > A_1$, the right-hand side is non-negative whenever

$$\|\theta\| \geq \frac{A_0}{\lambda - A_1}. \tag{45}$$

For such parameter vectors,

$$\|\nabla \mathcal{F}_\lambda(\theta)\|^2 \geq [(\lambda - A_1)\|\theta\| - A_0]^2 . \tag{46}$$

Furthermore, by Assumption 2.2,

$$\Delta \mathcal{F}_\lambda(\theta) = \Delta \mathcal{L}(\theta) + \lambda d \leq B_0 + B_1 \|\theta\|^2 + \lambda d. \tag{47}$$

Combining Eqs. (46) and (47) yields

$$\begin{aligned} \Psi_{s,\lambda}(\theta) \geq & -B_0 - B_1\|\theta\|^2 - \lambda d + \frac{1}{s}[(\lambda - A_1)\|\theta\| - A_0]^2 \\ = & \gamma_{s,\lambda}\|\theta\|^2 - \frac{2A_0(\lambda - A_1)}{s}\|\theta\| + \frac{A_0^2}{s} - B_0 - \lambda d. \end{aligned} \tag{48}$$

Let $c_{s,\lambda} = 2A_0(\lambda - A_1)/s$. Since $\gamma_{s,\lambda} > 0$, Young's inequality gives

$$c_{s,\lambda}\|\theta\| \leq \frac{\gamma_{s,\lambda}}{2}\|\theta\|^2 + \frac{c_{s,\lambda}^2}{2\gamma_{s,\lambda}}. \tag{49}$$

Substituting Eq. (49) into Eq. (48) gives

$$\Psi_{s,\lambda}(\theta) \geq \frac{\gamma_{s,\lambda}}{2}\|\theta\|^2 - C_{s,\lambda}, \tag{50}$$

where, after harmlessly enlarging the finite remainder, one admissible positive constant is

$$C_{s,\lambda} = B_0 + \lambda d + \frac{c_{s,\lambda}^2}{2\gamma_{s,\lambda}} + \frac{A_0^2}{s}. \tag{51}$$

The precise finite constant is not essential for the asymptotic result. Because the coefficient $\gamma_{s,\lambda}/2$ is positive, Eq. (50) diverges to $+\infty$ as $\|\theta\| \to \infty$. □

Corollary 2.1 (Admissible range of s) *Under the assumptions of Theorem 1, if $B_1 > 0$, differential confinement is guaranteed for*

$$0 < s < \frac{(\lambda - A_1)^2}{B_1}. \tag{52}$$

If $B_1 = 0$, the criterion holds for every finite $s > 0$ provided that $\lambda > A_1$.

Proof The statement follows by solving $\gamma_{s,\lambda} > 0$ using Eq. (42). □

The conditions $\lambda > A_1$ and $\gamma_{s,\lambda} > 0$ are sufficient rather than necessary. In particular, the reverse-triangle estimate in Eq. (44) discards directional information about the alignment between $\nabla \mathcal{L}(\theta)$ and θ. Less conservative sufficient criteria may therefore be obtained when this alignment can be controlled directly.

The theorem exposes the interaction among the relevant quantities. Within the admissible regime $\lambda > A_1$, increasing λ enlarges the favorable coefficient $(\lambda - A_1)^2/s$, whereas larger values of A_1, B_1, or s make the sufficient condition more difficult to satisfy. Thus, even under the stated derivative-growth bounds, differential confinement may be guaranteed only over a restricted range of s.

2.9 Special case of uniformly bounded derivatives

A simpler result follows when the data-loss gradient and Hessian trace are globally bounded. This is a mathematically valid special case, but such bounds must be assumed or proved for the model under consideration; they do not follow from bounded inputs alone.

Corollary 2.2 (Uniform derivative bounds) *Suppose that there exist finite constants* $C_1, C_2 \geq 0$ *such that*

$$\|\nabla \mathcal{L}(\theta)\| \leq C_1, \qquad \Delta \mathcal{L}(\theta) \leq C_2 \qquad \text{for every } \theta \in \mathbb{R}^d. \tag{53}$$

Then, for every $\lambda > 0$ *and every* $s > 0$,

$$\Psi_{s,\lambda}(\theta) \longrightarrow +\infty \qquad \text{as } \|\theta\| \longrightarrow \infty. \tag{54}$$

Proof Set $A_0 = C_1$, $A_1 = 0$, $B_0 = C_2$, and $B_1 = 0$ in Theorem 1. Then

$$\gamma_{s,\lambda} = \frac{\lambda^2}{s} > 0$$

for every finite $s > 0$. □

This corollary recovers the intuition underlying the original quadratic domination argument. Its conclusion is correct when the uniform derivative assumptions in Eq. (53) are valid. The counterexample in Proposition 2.4 demonstrates why these assumptions cannot be inferred solely from smooth activations and bounded input embeddings.

2.10 Scope of the theoretical result

The analysis establishes an unconditional function-level result and a conditional derivative-level result. Positive quadratic regularization guarantees coercivity and Gibbs integrability, but differential confinement additionally requires the global derivative-growth conditions of Theorem 1. The theorem is architecture-agnostic; applying it to a particular Transformer requires analytical verification of those conditions.

Finite measurements of $\Psi_{s,\lambda}$ cannot prove the global limit in Eq. (21). They can, however, identify directions and parameter ranges in which the gradient and Hessian-trace terms support or obstruct confinement-compatible growth. This motivates the geometric and computational framework developed next.

3 Geometric Interpretation and Computational Estimation

Section 2 distinguishes coercivity of the regularized objective from differential confinement of its parameter-space derivatives. This section develops a computational interpretation of that distinction. The goal is to resolve the terms governing $\Psi_{s,\lambda}$, evaluate their behavior along selected parameter-space paths, and estimate the data-loss Hessian trace without forming the Hessian explicitly.

The resulting framework is a finite geometric diagnostic rather than a global certificate. Measurements over finitely many parameter points, directions, and radii can identify behavior compatible with or adverse to differential confinement, but they cannot establish the global limit in Eq. (21).

3.1 Geometry of quadratic regularization

The regularized objective and its derivatives are defined in Eqs. (12) and (13). The quadratic penalty affects three geometric quantities. It adds radial growth $\lambda\|\theta\|^2/2$ to the function value, contributes the radial gradient $\lambda\theta$, and shifts the Hessian by the isotropic term λI_d. Consequently, for every unit direction $u \in \mathbb{R}^d$,

$$u^\top \nabla^2 \mathcal{F}_\lambda(\theta) u = u^\top \nabla^2 \mathcal{L}(\theta) u + \lambda \|u\|^2 = u^\top \nabla^2 \mathcal{L}(\theta) u + \lambda. \tag{55}$$

Thus, at a fixed parameter vector, coupled L^2 regularization adds the same curvature increment in every Euclidean direction. Any additional anisotropy observed between independently trained conditions must arise from changes in the learned parameter vector and the corresponding data-loss Hessian.

To illustrate the resulting geometry, consider a two-dimensional affine slice through a reference vector $\theta^{(0)}$. For orthonormal directions $u_1, u_2 \in \mathbb{R}^d$, define

$$\theta(\alpha, \beta) = \theta^{(0)} + \alpha u_1 + \beta u_2. \tag{56}$$

On this slice, the regularizer contributes

$$\begin{aligned} \frac{\lambda}{2} \left\| \theta^{(0)} + \alpha u_1 + \beta u_2 \right\|^2 = \frac{\lambda}{2} \|\theta^{(0)}\|^2 + \lambda\alpha \langle \theta^{(0)}, u_1 \rangle \\ + \lambda\beta \langle \theta^{(0)}, u_2 \rangle + \frac{\lambda}{2} \left(\alpha^2 + \beta^2 \right). \end{aligned} \tag{57}$$

The final term is a rotationally symmetric paraboloid in the selected plane. The complete sliced objective also contains the potentially non-convex data-fitting surface, so the quadratic contribution does not by itself establish global convexity, uniqueness of the minimizer, or differential confinement.

Fig. 1 visualizes this distinction. The quadratic term adds increasing radial function values and an isotropic curvature increment to the displayed data-fitting landscape. This illustrates objective-level coercivity and the curvature shift in Eq. (55). The schematic does not imply that the data-loss derivatives are uniformly controlled or that $\Psi_{s,\lambda}$ diverges along every parameter-space direction.

3.2 Directional confinement profiles

The global condition in Eq. (21) must hold along every sequence whose parameter norm diverges. Direct verification over the full space is infeasible for a large Transformer, so the diagnostic is evaluated along the paths defined in Eq. (7). For a unit

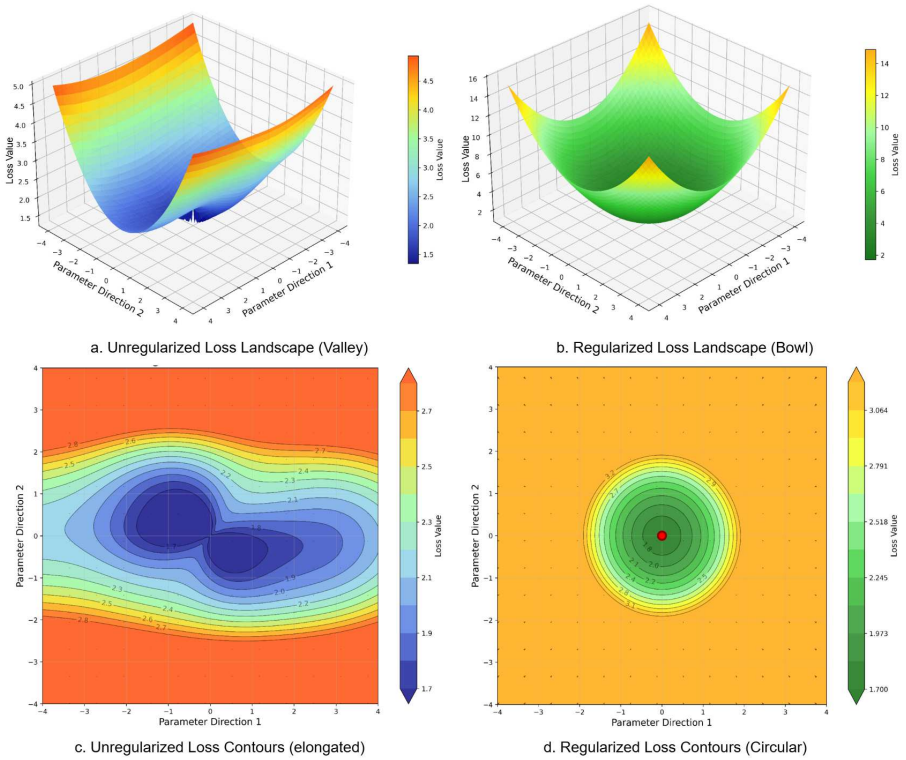


Fig. 1 Schematic effect of quadratic regularization on a two-dimensional parameter-space slice. The left and right panels illustrate the same data-fitting surface without and with the contribution $\lambda\|\theta\|^2/2$, respectively. The quadratic term adds radial growth to the displayed slice and shifts its directional curvature by λ. The illustration conveys the coercive tendency of the regularizer but does not establish global convexity, uniqueness of the minimizer, or differential confinement in every direction of the full parameter space.

direction u, define the directional profile

$$\psi_{s,\lambda}\left(r; u, \theta^{(0)}\right) = \Psi_{s,\lambda}\left(\theta^{(0)} + ru\right). \tag{58}$$

When the reference vector is clear, we write $\psi_{s,\lambda}(r; u)$.

The displacement r is measured from $\theta^{(0)}$, whereas the norm appearing in the confinement condition is measured from the origin of the selected parameterization. Their relationship is

$$\left\|\theta^{(0)} + ru\right\|^2 = \|\theta^{(0)}\|^2 + 2r\langle\theta^{(0)}, u\rangle + r^2. \tag{59}$$

Therefore, r^2 and $\|\theta^{(0)} + ru\|^2$ coincide only under additional orthogonality or origin assumptions. The experiments consequently retain both the radial displacement and the actual parameter norm.

The sampled directions include the following complementary classes.

Random directions:

An isotropic random direction is obtained from a Gaussian or Rademacher vector $\widetilde{u}$ by setting

$$u_{\mathrm{rand}} = \frac{\widetilde{u}}{\|\widetilde{u}\|}. \tag{60}$$

Random directions provide broad but necessarily incomplete coverage.

Gradient-aligned directions:

When $\nabla \mathcal{F}_\lambda(\theta^{(0)}) \neq 0$, the normalized positive and negative gradient directions are

$$u^{\pm}_{\mathrm{grad}} = \pm \frac{\nabla \mathcal{F}_\lambda(\theta^{(0)})}{\|\nabla \mathcal{F}_\lambda(\theta^{(0)})\|}. \tag{61}$$

These directions probe paths locally aligned with and opposed to the regularized gradient.

Hessian-aligned directions:

Approximate eigenvectors associated with the largest and smallest algebraic eigenvalues of $\nabla^2 \mathcal{F}_\lambda(\theta^{(0)})$ probe locally extremal curvature modes. The smallest algebraic eigenvalue corresponds to a negative-curvature direction only when it is negative. Because these directions are calculated at the reference vector, they need not remain Hessian eigenvectors along the complete radial path.

Parameter-group directions:

Let $\mathcal{I}_g \subseteq \{1, \ldots, d\}$ be the coordinate set associated with parameter group g. A group-restricted direction satisfies

$$[u_g]_j = 0 \qquad \text{for all } j \notin \mathcal{I}_g, \qquad \|u_g\| = 1. \tag{62}$$

Such directions isolate attention, feed-forward, embedding, normalization, or selected layer parameters.

Low-diagnostic directions:

Directions producing small diagnostic values are especially relevant because differential confinement is a worst-direction property. At a fixed radius r, the idealized search is

$$u_{\mathrm{low}}(r) \in \operatorname*{arg\,min}_{\|u\|=1} \Psi_{s,\lambda}\left(\theta^{(0)} + ru\right). \tag{63}$$

The global problem is intractable at the investigated scale, but candidate screening or constrained optimization on the unit sphere can provide approximate low-diagnostic directions.

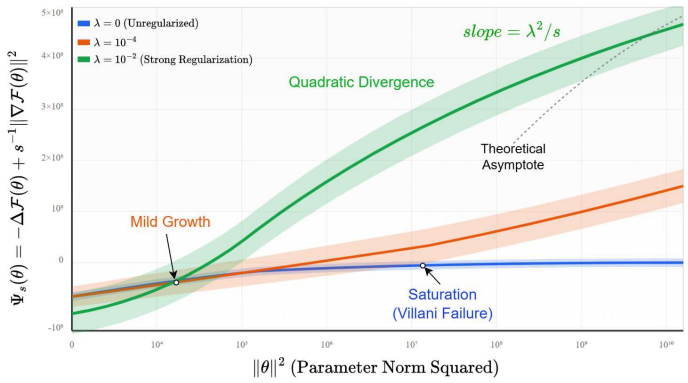


Fig. 2 Schematic directional profiles of the confinement diagnostic. The diagnostic is represented along a finite set of normalized parameter-space directions and plotted against the actual squared parameter norm. The variation among the profiles illustrates that finite-radius behavior can depend strongly on direction. The figure provides a geometric interpretation of directional growth and does not represent a global certificate over all directions or arbitrarily large parameter norms.

For a finite direction set $\mathcal{U} = \{u_1, \dots, u_K\}$, define the sampled lower envelope

$$\underline{\psi}^{(K)}_{s,\lambda}(r) = \min_{1 \le k \le K} \psi_{s,\lambda}(r; u_k). \tag{64}$$

The lower envelope is more informative than an average for detecting a possible adverse direction, but it is not the infimum over the unit sphere. Directional medians and lower and upper quantiles are therefore reported together with the sampled minimum to distinguish systematic behavior from a small number of atypical profiles.

Fig. 2 illustrates this directional interpretation. Different normalized paths can produce substantially different finite-radius profiles even when they originate from the same checkpoint. An increasing profile is compatible with confinement over the displayed interval, while a flat or decreasing profile identifies a direction in which such growth is not observed. Neither behavior alone determines the asymptotic limit because the figure represents only finitely many directions and a bounded radial range.

3.3 Component-wise interpretation of the diagnostic

The decomposition in Eq. (23) separates five contributions:

$$\begin{aligned} D_{\mathrm{tr}}(\theta) &= -\Delta\mathcal{L}(\theta), & R_{\mathrm{tr}} &= -\lambda d, \\ D_{\mathrm{grad}}(\theta) &= \frac{1}{s}\|\nabla\mathcal{L}(\theta)\|^2, & I_{\mathrm{grad}}(\theta) &= \frac{2\lambda}{s}\langle\nabla\mathcal{L}(\theta), \theta\rangle, \\ R_{\mathrm{grad}}(\theta) &= \frac{\lambda^2}{s}\|\theta\|^2. \end{aligned} \tag{65}$$

Consequently,

$$\Psi_{s,\lambda}(\theta) = D_{\mathrm{tr}}(\theta) + R_{\mathrm{tr}} + D_{\mathrm{grad}}(\theta) + I_{\mathrm{grad}}(\theta) + R_{\mathrm{grad}}(\theta). \tag{66}$$

The regularizer-trace contribution $R_{\mathrm{tr}} = -\lambda d$ is strictly negative for $\lambda > 0$. The sign of $D_{\mathrm{tr}} = -\Delta\mathcal{L}(\theta)$ depends on the data-loss Hessian trace: a positive trace reduces the diagnostic, whereas a negative trace increases it. The data-gradient and radial regularization terms are non-negative, while the interaction term can have either sign. Its sign is determined by the alignment between the data-loss gradient and the parameter vector. When both norms are non-zero, define

$$\cos\phi(\theta) = \frac{\langle \nabla\mathcal{L}(\theta), \theta \rangle}{\|\nabla\mathcal{L}(\theta)\| \, \|\theta\|}. \tag{67}$$

Then

$$I_{\mathrm{grad}}(\theta) = \frac{2\lambda}{s} \|\nabla\mathcal{L}(\theta)\| \|\theta\| \cos\phi(\theta). \tag{68}$$

Positive alignment reinforces the regularization gradient, whereas negative alignment reduces the diagnostic. Evaluating the five terms along each sampled path identifies whether an observed profile is controlled primarily by curvature, the radial penalty, the data gradient, or gradient–parameter alignment.

3.4 Stochastic estimation of the Hessian trace

The data-loss Laplacian is the trace of a $d \times d$ Hessian, whose explicit construction is impractical at the investigated model scale. Let

$$H_{\mathcal{L}}(\theta) = \nabla^2\mathcal{L}(\theta),$$

and let $v^{(1)}, \ldots, v^{(M)} \in \mathbb{R}^d$ be independent probes satisfying

$$\mathbb{E}[v^{(m)}] = 0, \qquad \mathbb{E}\left[v^{(m)}(v^{(m)})^\top\right] = I_d. \tag{69}$$

The Hutchinson estimator of the data-loss Hessian trace is

$$\widehat{\tau}_{\mathcal{L},M}(\theta) = \frac{1}{M} \sum_{m=1}^{M} (v^{(m)})^\top H_{\mathcal{L}}(\theta) v^{(m)}. \tag{70}$$

Using the probe moment condition and cyclicity of the trace gives

$$\mathbb{E}\left[\widehat{\tau}_{\mathcal{L},M}(\theta)\right] = \mathrm{tr}\left[\nabla^2\mathcal{L}(\theta)\right] = \Delta\mathcal{L}(\theta). \tag{71}$$

For standard Gaussian probes, $v^{(m)} \sim \mathcal{N}(0, I_d)$, and a symmetric Hessian,

$$\mathrm{Var}\left[\widehat{\tau}_{\mathcal{L},M}\right] = \frac{2}{M} \|H_{\mathcal{L}}\|_F^2. \tag{72}$$

For independent Rademacher probes,

$$\mathrm{Var}\left[\widehat{\tau}_{\mathcal{L},M}\right] = \frac{2}{M} \sum_{j \neq k} [H_{\mathcal{L}}]_{jk}^2. \tag{73}$$

Thus, estimator variability depends on the Hessian geometry and not only on the parameter dimension.

Let

$$h_m(\theta) = (v^{(m)})^\top H_{\mathcal{L}}(\theta) v^{(m)}$$

denote the individual quadratic forms, and define

$$\overline{h}_M(\theta) = \frac{1}{M} \sum_{m=1}^{M} h_m(\theta).$$

Their sample variance is

$$\widehat{\sigma}_h^2(\theta) = \frac{1}{M-1} \sum_{m=1}^{M} \left[h_m(\theta) - \overline{h}_M(\theta) \right]^2. \tag{74}$$

The estimated standard error of the trace mean is therefore

$$\widehat{\mathrm{SE}}\left[\widehat{\tau}_{\mathcal{L},M}(\theta)\right] = \frac{\widehat{\sigma}_h(\theta)}{\sqrt{M}}. \tag{75}$$

The exact regularizer trace does not require stochastic estimation. Using Eq. (13), the complete regularized trace is reconstructed as

$$\widehat{\tau}_{\lambda,M}(\theta) = \widehat{\tau}_{\mathcal{L},M}(\theta) + \lambda d. \tag{76}$$

Each quadratic form in Eq. (70) requires only the Hessian–vector product

$$w^{(m)} = \nabla^2 \mathcal{L}(\theta) v^{(m)}, \qquad h_m(\theta) = (v^{(m)})^\top w^{(m)}. \tag{77}$$

These products are obtained through automatic differentiation without forming or storing the full Hessian [1, 5, 10].

3.5 Estimated confinement diagnostic

Combining the exact gradient quantities with the stochastic data-loss trace gives

$$\widehat{\Psi}_{s,\lambda,M}(\theta) = -\widehat{\tau}_{\mathcal{L},M}(\theta) - \lambda d + \frac{1}{s} \left\| \nabla \mathcal{L}(\theta) + \lambda \theta \right\|^2. \tag{78}$$

Conditional on the parameter vector and evaluation batch, the trace is the only stochastic term when the gradient is evaluated exactly on the same fixed loss. Consequently,

$$\mathbb{E}\left[\widehat{\Psi}_{s,\lambda,M}(\theta)\right] = \Psi_{s,\lambda}(\theta), \qquad \mathrm{Var}\left[\widehat{\Psi}_{s,\lambda,M}(\theta)\right] = \mathrm{Var}\left[\widehat{\tau}_{\mathcal{L},M}(\theta)\right]. \tag{79}$$

The probe-based standard error of the diagnostic is therefore given by Eq. (75).

The estimated directional profile is

$$\widehat{\psi}_{s,\lambda,M}\left(r; u, \theta^{(0)}\right) = \widehat{\Psi}_{s,\lambda,M}\left(\theta^{(0)} + ru\right). \tag{80}$$

At each evaluated parameter point, the experimental record includes the trace estimate, probe uncertainty, parameter norm, data-gradient term, interaction term, radial term, and complete diagnostic.

Probe, data, and training variations represent distinct uncertainty sources. Probe variation arises from the stochastic trace vectors, data variation from the selected evaluation batch, and training variation from independently trained runs. These sources are reported separately rather than being combined into a single error bar.

3.6 Layerwise and parameter-group decomposition

To resolve the diagnostic across architectural components, partition the parameter vector into disjoint groups:

$$\theta = \left(\theta_1^\top, \theta_2^\top, \ldots, \theta_G^\top\right)^\top, \qquad \theta_g \in \mathbb{R}^{d_g}, \qquad \sum_{g=1}^{G} d_g = d. \tag{81}$$

A group may represent an attention projection, feed-forward block, embedding matrix, normalization module, complete layer, or another disjoint coordinate subset.

Let $g_g(\theta) = \nabla_{\theta_g} \mathcal{F}_\lambda(\theta)$ denote the group-gradient vector, and let $H_{gg}(\theta) = \nabla^2_{\theta_g} \mathcal{F}_\lambda(\theta)$ denote the corresponding diagonal Hessian block associated with θ_g. The groupwise diagnostic is

$$\Psi^{(g)}_{s,\lambda}(\theta) = -\operatorname{tr}\left[H_{gg}(\theta)\right] + \frac{1}{s}\|g_g(\theta)\|^2. \tag{82}$$

Because the groups form a disjoint partition,

$$\Delta\mathcal{F}_\lambda(\theta) = \sum_{g=1}^{G} \operatorname{tr}\left[H_{gg}(\theta)\right], \qquad \text{and} \qquad \|\nabla\mathcal{F}_\lambda(\theta)\|^2 = \sum_{g=1}^{G} \|g_g(\theta)\|^2$$

. Therefore,

$$\Psi_{s,\lambda}(\theta) = \sum_{g=1}^{G} \Psi^{(g)}_{s,\lambda}(\theta). \tag{83}$$

For stochastic estimation, let $v_g^{(m)} \in \mathbb{R}^{d_g}$ be an independent probe for group g. The block-trace estimator is

$$\widehat{\tau}_{g,M}(\theta) = \frac{1}{M}\sum_{m=1}^{M} (v_g^{(m)})^\top H_{gg}(\theta) v_g^{(m)}. \tag{84}$$

The corresponding estimated groupwise diagnostic is

$$\widehat{\Psi}^{(g)}_{s,\lambda,M}(\theta) = -\widehat{\tau}_{g,M}(\theta) + \frac{1}{s}\|g_g(\theta)\|^2. \tag{85}$$

The exact groupwise quantities are additive, as shown in Eq. (83). If the blockwise probes are sampled independently, the sum of the blockwise trace estimators is an unbiased estimator of the full Hessian trace. However, a particular blockwise stochastic realization need not coincide with a separately computed full-space Hutchinson realization because the latter also contains stochastic cross-block quadratic terms. The two estimators agree in expectation and converge to the same exact trace as the number of probes increases.

Because groupwise values can be negative, their normalized interpretation requires care. The signed share

$$\eta_g^{\text{signed}}(\theta) = \frac{\Psi^{(g)}_{s,\lambda}(\theta)}{\Psi_{s,\lambda}(\theta)} \tag{86}$$

sums to one when the denominator is non-zero, but may lie outside $[0, 1]$. The absolute-magnitude share

$$\eta_g^{\text{abs}}(\theta) = \frac{\left|\Psi^{(g)}_{s,\lambda}(\theta)\right|}{\sum_{h=1}^{G}\left|\Psi^{(h)}_{s,\lambda}(\theta)\right|} \tag{87}$$

lies in $[0, 1]$ and sums to one, but discards sign. These quantities are therefore descriptive attributions rather than probabilities or causal effects.

3.7 Finite-radius growth and temperature sensitivity

The theoretical condition concerns an infinite-radius limit, whereas numerical experiments access only finite intervals. For a direction u and radii $r_a < r_b$, define the finite-radius growth summary

$$G_{s,\lambda}(r_a, r_b; u) = \frac{\psi_{s,\lambda}(r_b; u) - \psi_{s,\lambda}(r_a; u)}{\|\theta(r_b; u)\|^2 - \|\theta(r_a; u)\|^2}, \tag{88}$$

provided the denominator is non-zero. Positive and negative values indicate growth and decrease, respectively, over the selected interval. They do not determine the asymptotic limit.

The temperature parameter controls only the squared-gradient contribution. At a fixed parameter vector,

$$\frac{\partial}{\partial s}\Psi_{s,\lambda}(\theta) = -\frac{1}{s^2}\left\|\nabla\mathcal{F}_\lambda(\theta)\right\|^2 \le 0. \tag{89}$$

Hence, the diagnostic is non-increasing in s. More precisely, for $0 < s_1 < s_2$,

$$\Psi_{s_1,\lambda}(\theta) - \Psi_{s_2,\lambda}(\theta) = \left(\frac{1}{s_1} - \frac{1}{s_2}\right)\left\|\nabla\mathcal{F}_\lambda(\theta)\right\|^2 \ge 0. \tag{90}$$

This exact identity provides an implementation check when multiple temperatures are evaluated from the same gradient and trace measurements.

3.8 Interpretive scope

The proposed framework measures finite, parameterization-dependent geometry. A sampled profile that increases over the evaluated range may saturate or decrease at larger radii, while an initially decreasing profile may eventually be dominated by the positive radial term. Similarly, positivity of $\widehat{\Psi}_{s,\lambda,M}(\theta)$ at one parameter point does not establish divergence to $+\infty$.

The diagnostic also depends on the Euclidean parameterization; it is not generally invariant under rescaling or other function-preserving reparameterizations. Comparisons therefore use matched architectures and parameterization conventions. Finally, finite diagnostic values are not converted into log-Sobolev constants, mixing rates, or causal claims about validation performance without additional theoretical or statistical justification.

Subject to these limitations, the framework provides a scalable method for identifying the gradient, curvature, regularization, direction, and temperature effects that support or obstruct confinement-compatible behavior. Section 4 applies this framework to the directional, checkpoint-level, spectral, estimator-reliability, and parameter-group analyses.

4 Experimental Results

This section evaluates the geometric diagnostic developed in Sections 2 and 3. The experiments examine its variation across parameter-space directions, radial displacements, temperatures, regularization strengths, optimizer implementations, and training checkpoints. Additional analyses assess stochastic trace-estimation reliability, Hessian spectral statistics, validation association, and parameter-group attribution.

The principal experiments use GPT-Neo-125M on Penn Treebank and WikiText-103. Uncertainty across training configurations is estimated from independent training seeds, while stochastic trace-estimation uncertainty is estimated from independent probe vectors. Complete model, dataset, optimizer, measurement, and statistical settings are provided in Section 7.

All reported results concern finite checkpoints, sampled directions, and bounded radial ranges. Accordingly, the terms “growth,” “saturation,” and “decrease” describe only the measured regime and do not establish the global limit in Eq. (21).

4.1 Experimental overview

Experiments used the regularization grid $\lambda \in \{0, 10^{-4}, 10^{-3}, 10^{-2}\}$ and the temperature grid $s \in \{0.25, 0.5, 1, 2, 4\}$. The value $\lambda = 0$ provides an unregularized reference, and $s = 1$ is used for the principal visualizations. At each selected parameter point, the analysis records the data-loss Hessian trace, the five diagnostic components in Eq. (66), and the complete estimate in Eq. (78).

Table 1 Overview of the experimental analyses. The model-specific settings and exact sample counts are provided in Section 7.

Analysis	Primary variables	Reported quantities
Directional profiles	Direction, radius, s, λ	$\widehat{\Psi}_{s,\lambda,M}$, diagnostic components, directional quantiles, lower envelope
Regularization comparison	$\lambda \in \{0, 10^{-4}, 10^{-3}, 10^{-2}\}$	Gradient contribution, Hessian trace, interaction term, finite-radius growth
Optimizer comparison	Coupled L^2, decoupled weight decay, no decay	Loss, perplexity, parameter norm, diagnostic components
Trace reliability	Probe count and probe distribution	Probe-set stability, empirical dispersion, standard error, confidence interval
Training evolution	Checkpoint and random seed	Training loss, validation perplexity, $\widehat{\Psi}_{s,\lambda,M}$
Groupwise extension	Architecture block, layer, regularization	Signed and absolute-magnitude group contributions

Three uncertainty sources are retained separately: variation across training seeds, variation across evaluation batches, and variation across stochastic probe sets. Table 1 summarizes the experimental comparisons.

4.2 Directional profiles of the confinement diagnostic

We evaluated the estimated diagnostic in Eq. (78) along the directional paths defined in Eq. (7). The sampled set contained $K = 42$ unit directions, and profiles were reported against the actual squared parameter norm rather than r^2, following Eq. (59). At each radial position, we summarized the distribution across directions using its median, interquartile range, 10th–90th percentile interval, and sampled lower envelope from Eq. (64).

The analysis distinguishes the complete radial sweep from an operationally defined usable regime. Directional summaries include only points classified as `near_reference` or `perturbed_below_uniform`; all finite points remain visible in the complete-sweep plots. The classification criteria are specified in Section 7.4.

Figs. 3a–c show the complete radial sweep for the post-hoc setting $\lambda = 10^{-3}$ and $s = 1$. Figs. 3d and 3e show the usable-regime directional summaries, while Fig. 3f reports the temperature-dependent squared-gradient contribution.

Fig. 3a shows substantial directional variation over the complete radial sweep. The extreme positive and negative diagnostic values at large displacements coincide with pronounced cross-entropy increases in Fig. 3b. These points lie outside the operationally defined usable regime and are excluded from the directional summaries. Within the retained regime, Figs. 3d and 3e show negative diagnostic values and substantial variation across directions and post-hoc regularization settings. The sampled minimum remains below the median, and none of the displayed summaries exhibits

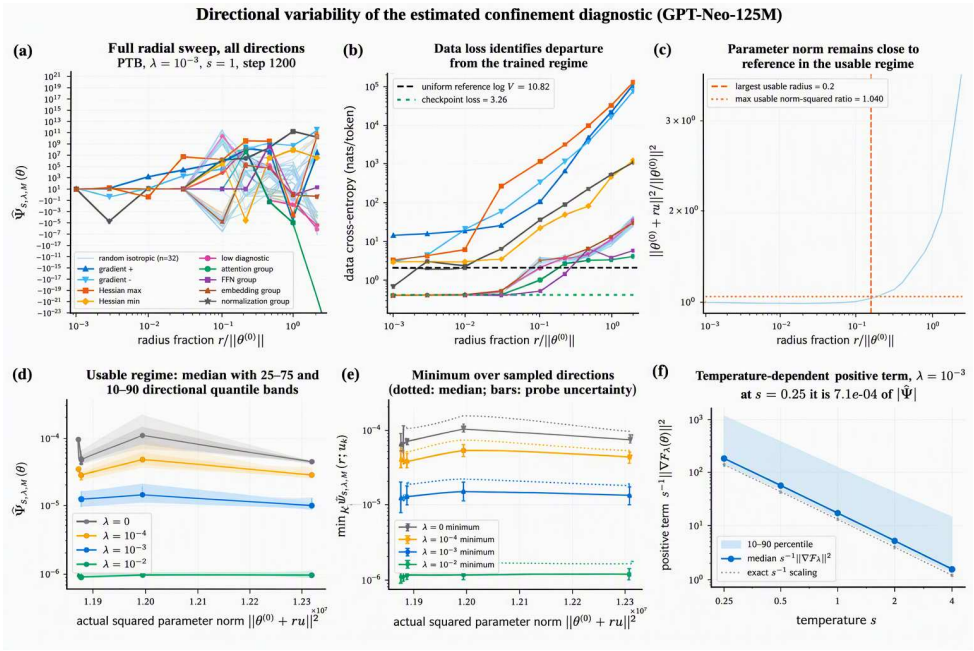


Fig. 3 Directional variability and temperature sensitivity of the estimated confinement diagnostic. GPT-Neo-125M is evaluated on a fixed PTB validation batch at training step 1200 of the coupled-L^2 run. The analysis uses $K = 42$ unit directions, comprising 32 isotropic random directions and 10 gradient-aligned, Hessian-aligned, low-diagnostic, and parameter-group directions. Each trace estimate uses $M = 64$ Rademacher probes. (a) Complete directional profiles for $\lambda = 10^{-3}$ and $s = 1$ over the full radial sweep. (b) Data cross-entropy along the same paths. The uniform-prediction reference $\log V$ separates parameter points below that reference from large perturbations that move the model outside its trained operating regime. (c) Change in the actual squared parameter norm relative to the reference checkpoint. The largest retained radial fraction is 0.2, at which the maximum squared-norm ratio is approximately 1.040. (d) Directional medians with 25th–75th and 10th–90th percentile bands within the usable regime. The four λ curves are post-hoc evaluations at the same reference checkpoint rather than independently trained model trajectories. (e) Minimum over the sampled directions, with dotted curves showing the corresponding medians. Error bars represent probe uncertainty for the specific direction attaining the displayed minimum and do not represent directional or training-run variability. (f) Temperature-dependent positive term $s^{-1}\|\nabla\mathcal{F}_\lambda(\theta)\|^2$ for $s \in \{0.25, 0.5, 1, 2, 4\}$, together with its exact s^{-1} reference scaling. At $s = 1$, its median magnitude is approximately 1.76×10^{-4} of the absolute total diagnostic over the retained points. The sampled minimum in (e) is not the infimum over the unit sphere, and the measurements provide finite directional evidence rather than a certificate of the global asymptotic condition.

uniform quadratic growth over the measured interval. Fig. 3f confirms the expected s^{-1} scaling of the positive squared-gradient contribution. At $s = 1$, its median magnitude is approximately 1.76×10^{-4} of the absolute diagnostic magnitude over the retained points. The measured diagnostic is therefore dominated by other components in this finite regime, motivating the decomposition below. These finite directional results neither verify nor contradict Theorem 1, because the global derivative-growth assumptions and the positivity of $\gamma_{s,\lambda}$ have not been established for the evaluated checkpoint.

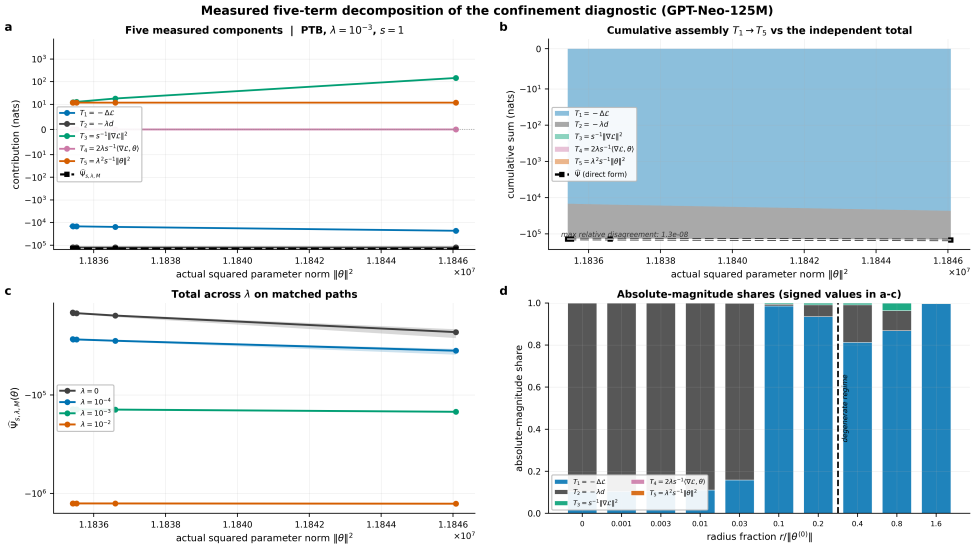


Fig. 4 Measured five-term decomposition of the confinement diagnostic. GPT-Neo-125M is evaluated on a fixed Penn Treebank validation batch at training step 1200 of the coupled-L^2 run. Results are restricted to parameter points classified as `near_reference` or `perturbed_below_uniform`. (a) Median signed contributions of the negative data-loss Hessian trace $T_1 = -\widehat{\Delta\mathcal{L}}$, regularizer trace $T_2 = -\lambda d$, data-gradient term $T_3 = s^{-1}\|\nabla\mathcal{L}\|^2$, gradient-interaction term $T_4 = 2\lambda s^{-1}\langle\nabla\mathcal{L}, \theta\rangle$, and radial regularization term $T_5 = \lambda^2 s^{-1}\|\theta\|^2$, together with the directly evaluated total diagnostic for $\lambda = 10^{-3}$ and $s = 1$. (b) Cumulative assembly of T_1 through T_5 compared with the independently evaluated total. The maximum decomposition discrepancy relative to the component scale is approximately 1.8×10^{-8}. (c) Median diagnostic with 10th–90th directional percentile bands for post-hoc values $\lambda \in \{0, 10^{-4}, 10^{-3}, 10^{-2}\}$ evaluated along matched paths at the same reference checkpoint. These curves are not independently trained λ-specific trajectories. (d) Median absolute-magnitude shares of the five terms across directions within the usable regime. Absolute shares discard component signs and are descriptive normalized magnitudes rather than probabilities. The measured decomposition shows that the data-loss Hessian trace and the constant regularizer-trace term dominate the total diagnostic over the retained finite-radius range.

4.3 Contribution of the gradient and Hessian-trace terms

We next evaluated the five contributions in Eq. (66) at matched parameter points using the same evaluation batch and trace estimates. Figs. 4a–c report signed components, cumulative reconstruction, and post-hoc dependence on λ within the usable regime. Fig. 4d reports normalized absolute component magnitudes while separating the large-radius measurements.

Fig. 4a shows that the negative data-loss trace and the constant regularizer trace $-\lambda d$ dominate the diagnostic within the retained regime. The positive data-gradient and radial terms are substantially smaller, while the median interaction term remains near zero. Fig. 4b confirms numerical reconstruction of the compact diagnostic from its five components. The maximum relative discrepancy is approximately 1.8×10^{-8} over the displayed subset. Fig. 4c shows that increasing the post-hoc value of λ shifts the diagnostic toward more negative values at the same checkpoint. This shift is primarily associated with the exact term $-\lambda d$. Because these curves use one checkpoint, they do

not represent independently trained λ-specific trajectories. Fig. 4d shows that the regularizer trace accounts for most of the absolute component magnitude near the reference point, while the data-loss trace becomes more prominent at larger radial fractions. The normalized magnitudes discard sign and must therefore be interpreted together with the signed components in Figs. 4a–c. Over the measured near-reference interval, the positive radial term does not dominate the trace contributions, and the total diagnostic does not exhibit uniform quadratic growth across the sampled directions.

4.4 Post-hoc dependence on regularization strength

Fig. 4c compares post-hoc values $\lambda \in \{0, 10^{-4}, 10^{-3}, 10^{-2}\}$ along matched paths from the same reference checkpoint. Increasing λ introduces the competing exact terms $-\lambda d$ and $\lambda^2\|\theta\|^2/s$. Over the measured range, the negative constant term dominates, shifting the diagnostic toward more negative values. Because the underlying parameter vectors are shared, this comparison isolates post-hoc diagnostic dependence on λ but does not represent separately trained model trajectories.

4.5 Coupled L^2 regularization and decoupled weight decay

The theoretical objective in Eq. (12) corresponds to coupled L^2 regularization. AdamW instead applies a decoupled parameter-decay step and does not generally follow the gradient of the same Euclidean regularized objective under adaptive preconditioning. We therefore treat the two implementations as distinct experimental conditions.

The comparison used three independent seeds per condition and a common nominal coefficient $\lambda = 10^{-3}$ for the two regularized conditions. Training and validation metrics were recorded along the complete trajectories. The data-loss Hessian trace was evaluated using a common stochastic-estimation protocol. For the AdamW checkpoints, $\widehat{\Psi}_{s,\lambda,M}$ was calculated post hoc using $\lambda = 10^{-3}$ and should not be interpreted as the objective minimized during AdamW training.

Figs. 5a–c show different trajectories for coupled L^2 regularization and AdamW at the common nominal coefficient $\lambda = 10^{-3}$. Coupled L^2 regularization produces higher training loss and validation perplexity together with a smaller parameter norm, whereas AdamW remains close to unregularized Adam in these measurements. At the final checkpoint, mean validation perplexity is approximately 21.600, 25.355, and 21.600 for unregularized Adam, coupled L^2, and AdamW, respectively. Corresponding mean parameter norms are approximately 3691.83, 3440.27, and 3691.66. Fig. 5d shows larger data-loss Hessian-trace estimates for the coupled-L^2 checkpoints than for the unregularized and AdamW checkpoints. Fig. 5e shows that the diagnostics evaluated at the common post-hoc value $\lambda = 10^{-3}$ remain negative for both regularized conditions. Only the coupled-L^2 condition directly optimizes Eq. (12). The AdamW diagnostic is a post-hoc evaluation and is not the objective minimized during AdamW training.

4.6 Hessian spectral statistics

The Hessian trace aggregates curvature across coordinates but does not describe its extremal directions. We therefore estimated selected largest and smallest algebraic

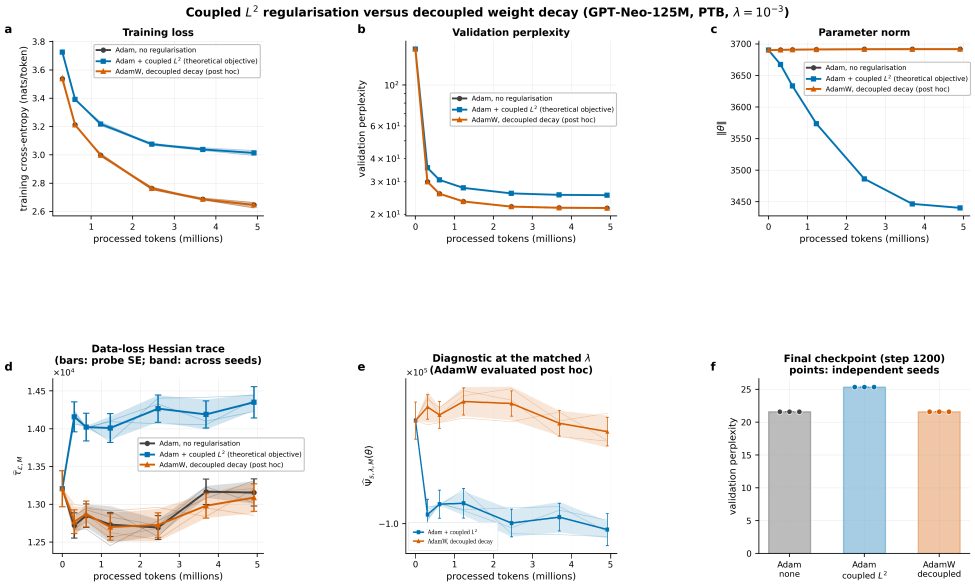


Fig. 5 Comparison of coupled L^2 regularization and decoupled weight decay. Three matched conditions are evaluated for GPT-Neo-125M on Penn Treebank: Adam without a penalty, Adam minimizing $\mathcal{L}(\theta) + \lambda\|\theta\|^2/2$, and AdamW with decoupled weight decay. The two regularized conditions use the common nominal coefficient $\lambda = 10^{-3}$, and each condition contains three independent training seeds. (a) Training cross-entropy. (b) Validation perplexity. (c) Parameter norm. (d) Data-loss Hessian-trace estimate. Faint curves denote independent seeds, shaded bands denote across-seed standard deviations, and error bars denote stochastic-probe standard errors. (e) Confinement diagnostic evaluated at the common value $\lambda = 10^{-3}$ for the coupled-L^2 and AdamW checkpoints. The coupled-L^2 condition directly corresponds to the theoretical objective, whereas the AdamW diagnostic is evaluated post hoc and is not the objective minimized by AdamW. (f) Final-checkpoint validation perplexity, with individual seeds shown as points and error bars representing across-seed standard deviations. Run variability and stochastic-probe uncertainty are displayed separately. The comparison is specific to the tested coefficient, dataset, and training configuration and does not establish a universally preferred regularization method.

eigenvalues, together with descriptive spectral ratios, using matrix-free Hessian–vector products.

Let $\widehat{\rho}_{\max}(\theta)$ and $\widehat{\rho}_{\min}(\theta)$ denote the largest and smallest algebraic retained Ritz values, respectively. The spectral anisotropy ratio reported below is defined as

$$\mathcal{A}(\theta) = \frac{|\widehat{\rho}_{\max}(\theta)|}{|\widehat{\rho}_{\min}(\theta)| + \varepsilon_{\text{spec}}}, \tag{91}$$

where $\varepsilon_{\text{spec}} > 0$ is the numerical stabilization constant used in the computation. Because the evaluated Hessians can be indefinite, $\mathcal{A}(\theta)$ is interpreted only as a descriptive extremal spectral ratio and not as a condition number.

At a fixed parameter vector, coupled L^2 regularization shifts every Hessian eigenvalue by exactly λ, as shown in Eq. (14). The checkpoints compared in Fig. 6 were obtained from independently trained models. Their spectral differences therefore

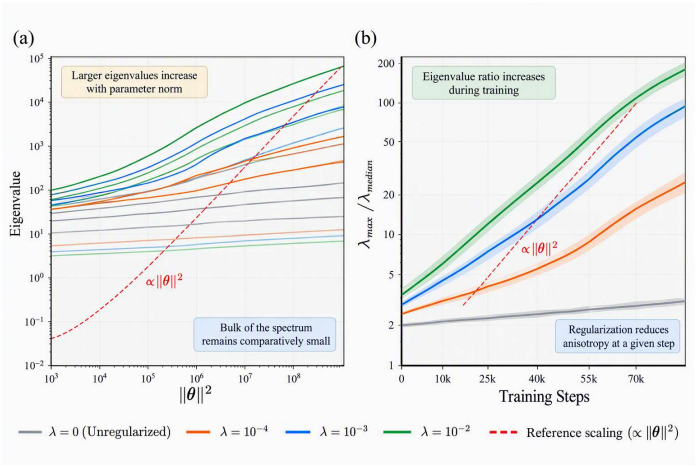


Fig. 6 Descriptive Hessian spectral statistics across regularization conditions and checkpoints. The panels report selected extremal eigenvalue summaries and a spectral anisotropy ratio obtained using Hessian–vector products. Curves corresponding to separately trained models reflect both the exact isotropic contribution λI_d and condition-dependent changes in the data-loss Hessian at the learned parameter vectors. At a fixed checkpoint, adding coupled L^2 regularization shifts every eigenvalue by exactly λ and does not selectively enlarge only the spectral tail. The displayed spectral ratios are descriptive summaries of the evaluated checkpoints; they do not establish global strong convexity, improved conditioning, a log-Sobolev inequality, or an accelerated Langevin mixing rate.

reflect both the known isotropic shift and condition-dependent changes in the data-loss Hessian.

Fig. 6 shows checkpoint- and condition-dependent differences in the selected extremal eigenvalues and spectral ratios. Because the models were trained independently, these differences cannot be attributed solely to the instantaneous addition of λI_d. The trace and extremal spectral statistics provide complementary curvature summaries. The former aggregates all coordinate contributions, whereas the latter emphasizes selected directions. Neither quantity alone determines differential confinement, which also depends on the squared gradient and its asymptotic behavior.

4.7 Reliability of stochastic trace estimation

The preceding analyses depend on stochastic estimates of the data-loss Hessian trace. We therefore examined the stability of the Hutchinson estimator as a function of probe count. All comparisons were performed at fixed model checkpoints and fixed evaluation batches so that the reported dispersion reflects stochastic-probe variation rather than changes in the model parameters or sampled data.

For each condition, the data-loss Hessian trace was estimated from independent probe sets using $M \in \{8, 16, 32, 64, 128\}$. Rademacher probes were used for the primary analysis, while Gaussian probes were included as a secondary comparison. The trace contribution of the quadratic regularizer was not estimated stochastically. When the complete regularized Hessian trace was required, the exact contribution λd was added analytically.

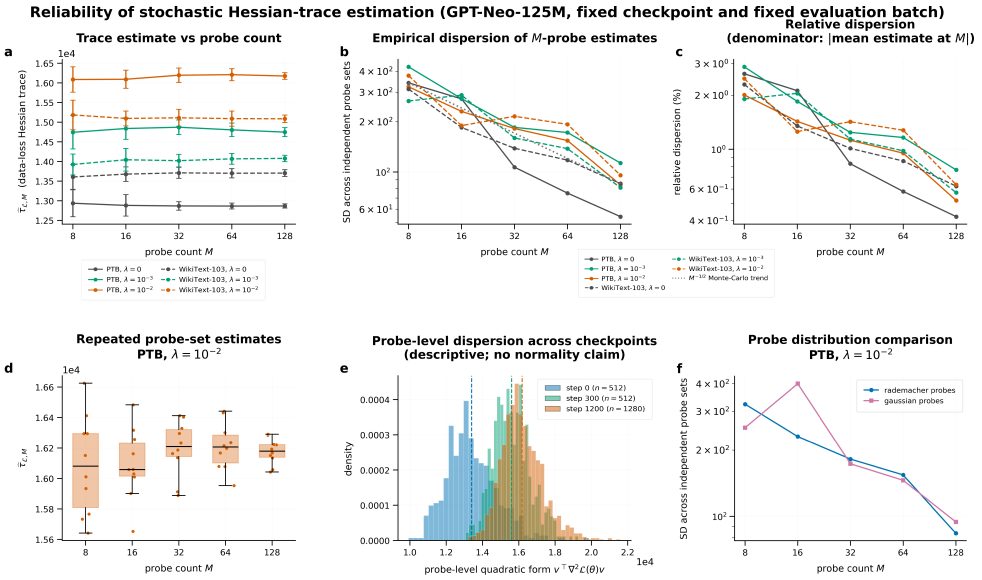


Fig. 7 Probe-count convergence and reliability of stochastic data-loss Hessian-trace estimation. GPT-Neo-125M is evaluated at fixed checkpoints and fixed evaluation batches on Penn Treebank and WikiText-103. (a) Mean data-loss Hessian-trace estimate as a function of probe count M for $\lambda \in \{0, 10^{-3}, 10^{-2}\}$. Solid and dashed curves distinguish Penn Treebank and WikiText-103, respectively. Error bars denote the standard deviation across independent Rademacher probe sets. (b) Empirical standard deviation of the M-probe estimates. The $M^{-1/2}$ line is shown only as a Monte Carlo reference slope and is not used as a dimension-dependent variance formula. (c) Relative dispersion, defined as the standard deviation across independent probe-set estimates divided by the absolute mean estimate. (d) Repeated Rademacher probe-set estimates for Penn Treebank at $\lambda = 10^{-2}$ and training step 1200. (e) Probe-level quadratic-form distributions at training steps 0, 300, and 1200. Dashed vertical lines mark the corresponding sample means. The histograms are descriptive and are not used to impose or validate a normality assumption. (f) Empirical dispersion obtained with Rademacher and Gaussian probes for Penn Treebank at $\lambda = 10^{-2}$ and step 1200. This panel compares probe-set variability only. The regularizer trace λd is added analytically when constructing the regularized trace and is not estimated by the Hutchinson procedure.

Fig. 7a shows stable mean data-loss trace estimates as the probe count increases. At the final checkpoints, the $M = 64$ Rademacher estimates differ from the corresponding $M = 128$ estimates by less than approximately 0.4% across the displayed conditions. Figs. 7b and 7c show decreasing empirical and relative dispersion with increasing M. At $M = 64$, relative dispersion across independent Rademacher probe sets ranges from approximately 0.6% to 1.3% over the six final-checkpoint conditions. Fig. 7d shows contraction of the independent probe-set estimates with increasing M. Fig. 7e shows checkpoint-dependent quadratic-form distributions, and Fig. 7f compares Rademacher and Gaussian probes under one matched condition. Both probe distributions show decreasing dispersion, without a consistent ordering across the evaluated probe counts. Table 2 reports the $M = 64$ Rademacher estimates at the final checkpoint. The standard errors and confidence intervals quantify uncertainty in the mean across independent probe sets at fixed parameters and a fixed evaluation batch.

Table 2 Reliability of stochastic data-loss Hessian-trace estimation at $M = 64$. The quantity reported is $\mathrm{tr}[\nabla^2 \mathcal{L}(\theta)]$, estimated with Rademacher probes and Hessian–vector products at training step 1200 using a fixed evaluation batch. The trace estimate is the mean across $R = 10$ independent probe-set estimates. The estimated standard error and 95% confidence interval refer to that mean. The exact regularizer contribution λd is not included in the displayed trace values.

Dataset	λ	Checkpoint	M	Trace estimate	Estimated SE	95% CI
PTB	0	1200	64	1.287×10^4	23.7	$[1.281, 1.292] \times 10^4$
PTB	10^{-3}	1200	64	1.481×10^4	54.5	$[1.468, 1.493] \times 10^4$
PTB	10^{-2}	1200	64	1.621×10^4	48.8	$[1.610, 1.632] \times 10^4$
WikiText-103	0	1200	64	1.370×10^4	37.3	$[1.362, 1.379] \times 10^4$
WikiText-103	10^{-3}	1200	64	1.407×10^4	43.6	$[1.397, 1.417] \times 10^4$
WikiText-103	10^{-2}	1200	64	1.509×10^4	61.0	$[1.495, 1.523] \times 10^4$

Variability across evaluation batches and training seeds is reported separately in the corresponding analyses.

4.8 Evolution of the diagnostic during training

Having examined the reliability of the stochastic trace estimator, we next evaluated the diagnostic over the training trajectory. Checkpoints were selected according to a prespecified schedule rather than according to their diagnostic values. The analysis used coupled-L^2 runs with $\lambda \in \{10^{-4}, 10^{-3}, 10^{-2}\}$, together with an unregularized reference condition.

At every checkpoint, the diagnostic and its components were evaluated at $s = 1$ using the value of λ employed during training. Measurements were averaged over two fixed evaluation batches within each run and then summarized across three independent training seeds. Stochastic-probe uncertainty, evaluation-batch variability, and training-seed variability were retained as distinct quantities. The longitudinal analysis used $M = 32$ Rademacher probes at each checkpoint, with the resulting probe uncertainty shown explicitly. This probe count is lower than the $M = 64$ setting used for the principal directional analyses and should therefore be interpreted together with the displayed trace-estimation error bars.

Figs. 8a and 8b show decreasing training cross-entropy and validation perplexity for every condition. Larger coupled-L^2 coefficients produce higher final values for both quantities in the tested PTB configuration. Fig. 8c shows a rapid early change in the data-loss Hessian trace followed by a narrower condition-dependent range. At later checkpoints, the $\lambda = 10^{-3}$ and $\lambda = 10^{-2}$ conditions generally exhibit larger data-loss trace estimates than the unregularized reference. The displayed quantity does not include the exact regularizer trace λd. Fig. 8d shows a sharp early reduction in the data-gradient contribution. Fig. 8e shows that the interaction term approaches values near zero, while the positive radial term remains approximately condition-specific and varies only modestly after the initial checkpoint. The estimated diagnostic is negative at every displayed checkpoint and training seed. Accordingly, Fig. 8f reports $-\widehat{\Psi}_{s,\lambda,M}$ on a logarithmic scale. Larger displayed values therefore correspond to more negative diagnostic values. The condition-dependent separation is associated primarily with $-\lambda d$ and differences in the learned data-loss Hessian trace. No measured trajectory exhibits a transition to a positive diagnostic during the observed training interval.

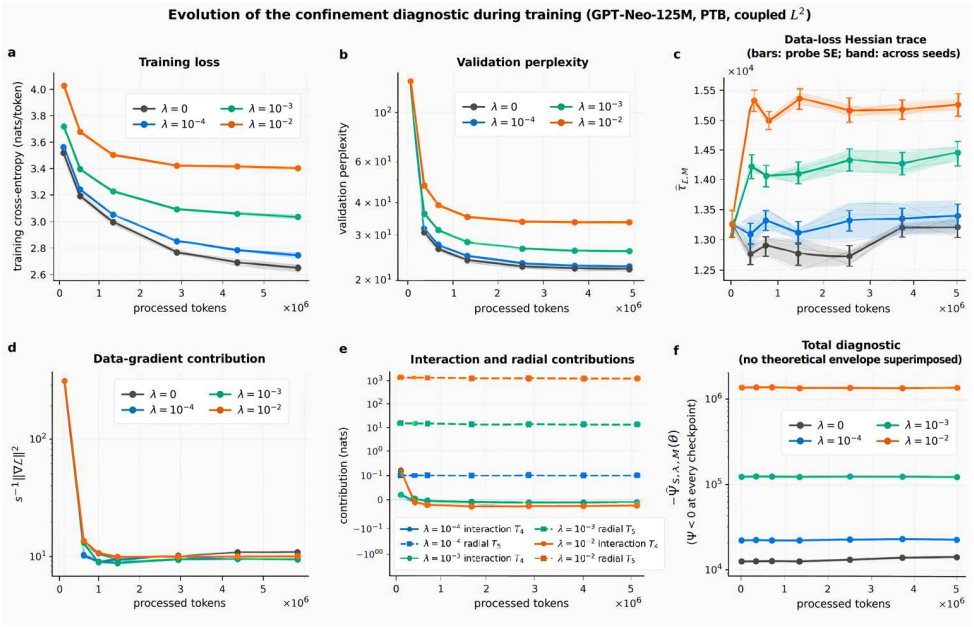


Fig. 8 Evolution of the confinement diagnostic and model performance during training. GPT-Neo-125M is evaluated on Penn Treebank for coupled-L^2 regularization coefficients $\lambda \in \{10^{-4}, 10^{-3}, 10^{-2}\}$ and an unregularized $\lambda = 0$ reference. Each condition contains three independent training seeds, and the diagnostic is evaluated at $s = 1$. Faint curves denote individual training seeds, prominent curves denote across-seed means, and shaded bands denote across-seed standard deviations. In panels containing trace-dependent quantities, error bars represent stochastic-probe standard errors. (a) Training cross-entropy. (b) Validation perplexity. (c) Estimated data-loss Hessian trace. (d) Data-gradient contribution $s^{-1}\|\nabla\mathcal{L}\|^2$. (e) Gradient-interaction contribution $2\lambda s^{-1}\langle\nabla\mathcal{L}, \theta\rangle$ and radial contribution $\lambda^2 s^{-1}\|\theta\|^2$. (f) Magnitude of the negative total diagnostic, $-\widehat{\Psi}_{s,\lambda,M}$, shown on a logarithmic scale because $\widehat{\Psi}_{s,\lambda,M} < 0$ at every displayed checkpoint and training seed. Each Hessian-trace estimate uses $M = 32$ Rademacher probes over two fixed evaluation batches. No theoretical exponential optimization, log-Sobolev, or mixing envelope is superimposed.

4.9 Checkpoint-level association with validation performance

We next examined the checkpoint-level association between the estimated confinement diagnostic and validation perplexity. The analysis used the same PTB runs, checkpoints, training seeds, evaluation batches, and matched regularization coefficients as the longitudinal experiment in Fig. 8. Each plotted observation represents one independently trained run at one scheduled checkpoint after averaging the diagnostic over the two fixed evaluation batches.

Because multiple checkpoints originate from the same training run, the observations are not statistically independent. Moreover, both the diagnostic and validation perplexity vary with training stage and regularization condition. The figure is therefore intended as a descriptive visualization of the measured trajectories rather than as a test of whether the diagnostic predicts validation performance.

Fig. 9 shows condition-specific diagnostic ranges and within-run checkpoint progression. The separation across regularization conditions is dominated by the exact

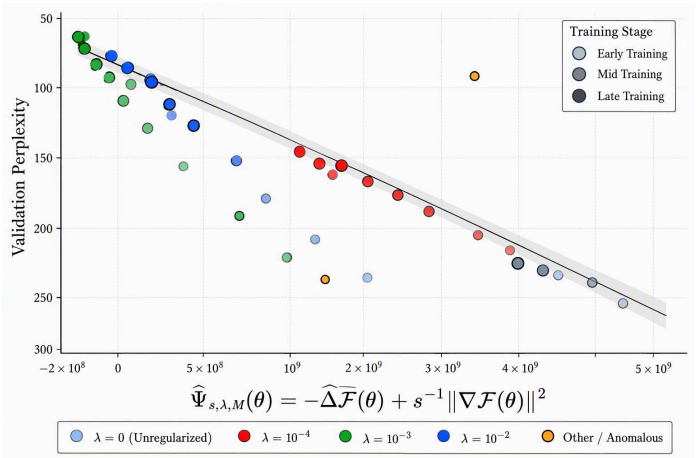


Fig. 9 Checkpoint-level association between the estimated confinement diagnostic and validation perplexity. Each point represents one scheduled checkpoint from an independently trained GPT-Neo-125M run on Penn Treebank. Colors denote regularization coefficients, marker size denotes checkpoint progression, and faint lines connect repeated checkpoints from the same training seed. The diagnostic is evaluated at $s = 1$ using the coefficient applied during training and is averaged over two fixed evaluation batches. The visualization is descriptive; no pooled regression is fitted across regularization conditions.

term $-\lambda d$, whereas validation perplexity decreases substantially during early training. Within individual conditions, the diagnostic occupies a comparatively narrow range relative to the change in perplexity. The measurements do not exhibit a common checkpoint-level relationship across conditions. Because checkpoints from the same run are repeated observations and both quantities vary with training stage, the figure is interpreted descriptively rather than as a predictive or causal analysis.

4.10 Layerwise and parameter-group extension

To examine architectural variation, we additionally report a ViT-B/16 extension using $\lambda \in \{0, 10^{-4}, 10^{-3}, 10^{-2}\}$. Parameters in each Transformer layer were partitioned into query, key, value, feed-forward, and LayerNorm groups. Groupwise diagnostics were computed using Eq. (85), and absolute-magnitude shares were used for the displayed normalized attributions.

Fig. 10 shows stronger normalized feed-forward attribution under zero and weak regularization. At $\lambda = 10^{-3}$, the attribution is more distributed across feed-forward and attention groups. At $\lambda = 10^{-2}$, query, key, and value groups become comparatively more prominent, while LayerNorm remains limited across the displayed layers. Table 3 summarizes these patterns qualitatively. The categories describe relative absolute-magnitude attribution and do not represent signed effect sizes or uncertainty estimates.

4.11 Summary of empirical findings

The experiments yield five principal findings. First, the estimated diagnostic varies substantially across parameter-space directions, and the sampled lower envelope

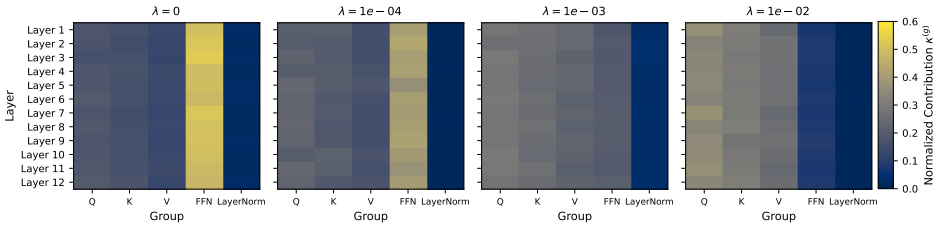


Fig. 10 Layerwise and parameter-group attribution of the confinement diagnostic in a ViT-B/16 extension. Rows correspond to Transformer layers and columns correspond to query (Q), key (K), value (V), feed-forward (FFN), and LayerNorm parameter groups. The four panels show the normalized groupwise attribution scores obtained under $\lambda \in \{0, 10^{-4}, 10^{-3}, 10^{-2}\}$. The unregularized and weakly regularized settings show comparatively strong FFN attribution, whereas the attention projections become more prominent as the regularization coefficient increases. LayerNorm remains comparatively limited across the displayed settings. The heatmaps describe relative component-level attribution within this experiment and do not establish a causal relationship between an architectural group and downstream model performance.

Table 3 Qualitative summary of normalized groupwise diagnostic attribution in the ViT-B/16 extension. The entries summarize the relative patterns visible in Fig. 10. They are categorical descriptions, not numerical estimates of signed contributions.

λ	**Query/key attribution**	**Value attribution**	**Feed-forward attribution**	**LayerNorm attribution**	**Principal observed pattern**
0	Low	Low	High	Low	Attribution is concentrated primarily in the feed-forward groups.
10^{-4}	Low to moderate	Low to moderate	Moderate to high	Low	Feed-forward groups remain prominent, with an emerging contribution from attention projections.
10^{-3}	Moderate	Moderate	Moderate	Low	Attribution becomes more distributed between attention and feed-forward groups.
10^{-2}	Moderate to high	Moderate to high	Low to moderate	Low	The relative attribution shifts toward attention projections across multiple layers.

remains below the directional median over the retained regime. Second, the negative data-loss Hessian trace and the constant regularizer trace dominate the diagnostic near the reference checkpoint, while the positive radial term remains smaller over the measured range. Third, the diagnostic remains negative at all displayed training checkpoints and does not exhibit uniform quadratic growth across the sampled finite-radius profiles. Fourth, coupled L^2 regularization and decoupled AdamW weight decay produce distinct training trajectories and geometric measurements at the same nominal coefficient. Fifth, stochastic trace estimation is stable at the selected probe counts, and the groupwise extension reveals substantial variation across architectural components. These results characterize finite measured geometry. They do not establish global differential confinement, a log-Sobolev constant, a mixing rate, or a causal relationship with validation performance.

5 Discussion

This study examined whether coupled quadratic regularization is sufficient to make a Transformer objective differentially confining and whether the relevant derivative-level behavior can be investigated computationally at realistic model scale. The theoretical and empirical results provide a consistent answer. Positive coupled L^2 regularization guarantees coercivity and Gibbs integrability, but it does not guarantee differential confinement without additional control of the data-loss gradient and Hessian trace. The proposed directional and stochastic-trace framework makes this distinction computationally observable, while preserving the separation between finite measured geometry and a global asymptotic statement.

The first part of this conclusion follows from the different mathematical levels at which the relevant properties are defined. Because the cross-entropy loss is non-negative, the quadratic lower bound in Eq. (15) ensures that $\mathcal{F}_\lambda(\theta) \to +\infty$ as $\|\theta\| \to \infty$. Proposition 2.2 therefore establishes coercivity, while Proposition 2.3 establishes Gaussian-dominated integrability of the corresponding Gibbs density. These conclusions depend on function values and require no global bounds on $\nabla\mathcal{L}$ or $\Delta\mathcal{L}$.

Differential confinement is stronger because it depends on the derivative-level balance in Eq. (23). The regularizer contributes the favorable quadratic term $\lambda^2\|\theta\|^2/s$, but it also contributes the negative constant $-\lambda d$. In addition, the data-loss Hessian trace enters with a negative sign when it is positive, and the interaction $\langle\nabla\mathcal{L}(\theta), \theta\rangle$ may either increase or decrease the diagnostic. Fig. 1 illustrates why radial growth of the objective should not be conflated with control of these derivative terms. The quadratic penalty adds an isotropic bowl-like contribution to a displayed parameter-space slice, but the complete data-fitting geometry can remain non-convex and direction dependent.

The counterexample in Proposition 2.4 establishes that this distinction is not merely technical. Along the selected ray, the data-loss Laplacian grows quadratically and enters $\Psi_{s,\lambda}$ with a negative sign. When $s > 1 + 4\lambda^2$, Eq. (31) consequently diverges to $-\infty$, even though the same objective remains smooth, coercive, and Gibbs-integrable. Thus, bounded embedded inputs, smooth model operations, and positive quadratic regularization do not imply global differential confinement. The multiplicative structure of this minimal example is also relevant to attention and feed-forward parameterizations, in which derivatives with respect to one parameter matrix can depend on the magnitudes of other trainable matrices.

Theorem 1 supplies the corresponding positive result. Under the gradient-growth and Hessian-trace-growth conditions in Eqs. (40) and (41), differential confinement follows when $\lambda > A_1$ and the coefficient $\gamma_{s,\lambda}$ in Eq. (42) is positive. The theorem therefore identifies the precise competition omitted by an argument based on coercivity alone. The regularized gradient must outgrow the permitted linear growth of the data-loss gradient, while its squared contribution must dominate the allowed quadratic growth of the data-loss Hessian trace. Corollary 2.1 further shows that this balance can hold only over a restricted temperature range when $B_1 > 0$. The theoretical contribution is therefore a conditional confinement criterion rather than an unconditional claim about all regularized Transformer objectives.

The directional framework in Section 3 translates this theoretical balance into finite measurements. Fig. 2 illustrates the central geometric issue: profiles originating from the same checkpoint can vary substantially across parameter-space directions. The measured experiment in Fig. 3 confirms that this directional variation persists in GPT-Neo-125M. Fig. 3a shows large positive and negative diagnostic values over the complete radial sweep, while Fig. 3b shows that the largest perturbations also produce pronounced increases in cross-entropy. These extreme points document the behavior of the derivative-level quantities away from the trained checkpoint, but they do not represent the model's near-reference operating regime.

The usable-regime summaries provide a more controlled comparison. Figs. 3d and 3e show that the diagnostic remains negative and varies substantially across directions and post-hoc regularization settings. The sampled lower envelope remains below the directional median, demonstrating why average behavior alone is insufficient for studying a global condition that can fail along a single parameter-space sequence. At the same time, the sampled envelope is not the infimum over the full unit sphere. These measurements therefore provide finite evidence of directional heterogeneity, not a global failure certificate.

The component decomposition explains the observed profiles. Figs. 4a and 4b show that, within the retained regime, the diagnostic is governed primarily by the negative data-loss Hessian-trace contribution and the constant regularizer trace $-\lambda d$. The positive data-gradient and radial terms are substantially smaller, and the median interaction term remains near zero. The numerical agreement between the compact diagnostic and the sum of its five components also supports the internal consistency of the implementation.

Fig. 4c shows that increasing λ post hoc at the same checkpoint shifts the diagnostic toward more negative values over the measured range. This finite-range behavior does not imply that stronger regularization generally obstructs asymptotic confinement. At fixed parameters, increasing λ simultaneously changes the negative term $-\lambda d$, the interaction term, and the positive term $\lambda^2\|\theta\|^2/s$. In the investigated regime, the negative constant is larger than the positive radial contribution. At sufficiently large norms, or under different derivative-growth conditions, their relative importance may change. The post-hoc result therefore characterizes the measured checkpoint and radial interval, whereas Theorem 1 concerns an infinite-norm limit under global assumptions.

The absolute component shares in Fig. 4d support the same interpretation. The regularizer trace accounts for most of the near-reference component magnitude, while the data-loss trace becomes more prominent at larger radial fractions. Because these shares discard sign, they quantify relative magnitude rather than favorable or adverse contributions. Their interpretation consequently requires the signed components in Figs. 4a–c.

Temperature sensitivity provides an additional link between the theory and the measurements. Equation (90) shows that $\Psi_{s,\lambda}$ is non-increasing in s at a fixed parameter vector. Fig. 3f reproduces the expected s^{-1} scaling of the positive squared-gradient term. Its small magnitude relative to the absolute diagnostic in the retained regime is consistent with the trace-dominated component decomposition. This result also clarifies why temperature cannot be treated as a passive scaling parameter: larger

s weakens the favorable squared-gradient contribution without changing the Hessian trace, in agreement with the admissible range identified by Corollary 2.1.

The training-evolution analysis extends the finite-radius findings across optimization checkpoints. Figs. 8a and 8b show decreasing training cross-entropy and validation perplexity for every tested condition, while Fig. 8c shows that the data-loss Hessian trace changes rapidly early in training and then enters a narrower, condition-dependent range. Fig. 8d shows a corresponding early reduction in the data-gradient contribution, and Fig. 8e shows that the interaction term approaches values near zero while the positive radial term remains approximately condition specific.

Despite these changes, the complete diagnostic remains negative at every displayed checkpoint and training seed. Fig. 8f therefore reports $-\widehat{\Psi}_{s,\lambda,M}$, for which larger displayed values mean a more negative diagnostic. The late-training trajectories indicate a comparatively stable finite-checkpoint balance among the trace, gradient, interaction, and regularization components. They do not show a transition to positive differential confinement during the measured training interval.

This observation does not contradict Theorem 1. The theorem concerns the limit $\|\theta\| \to \infty$ and assumes global constants A_0, A_1, B_0, and B_1. The checkpoint analysis neither estimates these constants nor establishes $\gamma_{s,\lambda} > 0$. Moreover, a negative diagnostic at finite checkpoints does not logically preclude divergence to $+\infty$ at larger parameter norms. The theoretical and empirical results therefore address complementary questions: the theorem identifies sufficient global conditions, while the experiments determine which terms control the diagnostic at measured checkpoints and along sampled finite paths.

The optimizer comparison further clarifies the scope of the theoretical objective. Figs. 5a–c show that coupled L^2 regularization and AdamW produce different loss, perplexity, and parameter-norm trajectories at the same nominal coefficient. Coupled L^2 regularization produces stronger norm contraction but higher validation perplexity in the tested PTB configuration, whereas AdamW remains close to unregularized Adam in these measurements. Fig. 5d additionally shows differences in the learned data-loss Hessian trace.

Only coupled L^2 regularization directly optimizes Eq. (12). The AdamW values in Fig. 5e are post-hoc evaluations of that diagnostic and should not be interpreted as the objective minimized by AdamW. The comparison therefore supports the distinction made in the Introduction: equal numerical coefficients do not define equivalent regularized objectives under adaptive optimization. Parameter norm, validation performance, and diagnostic value also do not follow a common ordering, indicating that these quantities describe different properties of the trained model.

The spectral measurements provide a complementary view of curvature. At a fixed parameter vector, Eq. (14) shows that coupled L^2 regularization shifts every Hessian eigenvalue by exactly λ. The condition-dependent differences in Fig. 6 cannot therefore be attributed only to this instantaneous isotropic shift because the displayed checkpoints were obtained from independently trained models. The differences also reflect changes in the learned data-loss Hessian.

The trace and extremal eigenvalue analyses address different aspects of the same geometry. The Hessian trace aggregates curvature across all coordinates, while the

spectral summaries emphasize selected extremal directions and anisotropy. Neither quantity independently establishes differential confinement, which also depends on the squared-gradient term and requires a global asymptotic limit. Likewise, the displayed spectral ratios do not establish global strong convexity, a uniform Hessian lower bound, or a Langevin mixing rate.

The reliability analysis supports the numerical use of stochastic trace estimation in this framework. Fig. 7a shows that the $M = 64$ Rademacher estimates differ from their $M = 128$ counterparts by less than approximately 0.4% across the displayed final-checkpoint conditions. Figs. 7b and 7c show decreasing dispersion with increasing probe count, and Fig. 7d shows contraction across independent probe-set estimates. The checkpoint-dependent quadratic-form distributions in Fig. 7e also demonstrate why trace uncertainty cannot be specified solely from model dimension.

Fig. 7f does not identify a universally superior probe distribution, but it confirms decreasing dispersion for both Rademacher and Gaussian probes under the matched condition. The trace estimates and confidence intervals in Table 2 further show that probe uncertainty is small relative to the reported trace magnitudes for the selected checkpoints. These results support the chosen probe counts for the present experiments, while leaving batch and training-run variation as separate uncertainty sources.

The checkpoint-level association analysis demonstrates why the diagnostic should not be interpreted as a direct surrogate for validation performance. Fig. 9 shows widely separated diagnostic ranges across regularization conditions, largely because of the exact contribution $-\lambda d$. Validation perplexity, by contrast, changes primarily with training progression within each run. A pooled relationship would therefore mix between-condition offsets with repeated within-run observations. The diagnostic is best interpreted as a geometric property of the evaluated objective, not as a validated predictor of perplexity or generalization.

The parameter-group extension shows that the global diagnostic can also conceal substantial architectural heterogeneity. Fig. 10 and Table 3 show comparatively strong feed-forward attribution under zero and weak regularization, a broader distribution across feed-forward and attention groups at intermediate regularization, and increasingly prominent query, key, and value attributions at the strongest tested setting. LayerNorm remains limited in the displayed normalized attribution.

These groupwise results complement the global directional analysis by showing that the balance between gradient and curvature contributions is not distributed uniformly across Transformer components. However, the displayed absolute-magnitude shares discard sign, and the categorical summaries in Table 3 do not provide signed effect sizes or causal evidence. A prominent groupwise attribution identifies where the measured diagnostic magnitude is concentrated, not which component determines optimization or validation performance.

Taken together, the results support a conditional rather than universal account of differential confinement in regularized Transformer objectives. Quadratic regularization supplies objective-level coercivity, Gibbs integrability, a radial gradient, and an isotropic Hessian shift. Whether these contributions produce differential confinement depends on their competition with the asymptotic growth of the data-loss gradient and

Hessian trace. The bilinear counterexample proves that positive regularization alone is insufficient, while Theorem 1 identifies explicit sufficient conditions under which it becomes sufficient.

The experiments complement this theory by resolving the same competing terms at realistic model scale. Across the investigated checkpoints and finite directional ranges, the data-loss Hessian trace and constant regularizer trace dominate the positive radial contribution, the diagnostic remains negative, and uniform quadratic growth is not observed across the sampled directions. These findings do not decide the global asymptotic condition, but they identify the mechanisms that prevent the sufficient theoretical balance from being demonstrated in the measured regime.

The principal significance of the framework is therefore methodological as well as theoretical. It replaces an unconditional inference from quadratic regularization with a precise growth criterion and a scalable procedure for examining its constituent terms. The theory specifies what must be established for global differential confinement, while the computational analysis identifies the directions, temperatures, checkpoints, optimizer implementations, and architectural components that support or obstruct confinement-compatible behavior. This combination provides a more rigorous basis for evaluating the geometric assumptions used in functional-analytic and Langevin-based analyses of large neural networks.

6 Limitations and Future Directions

The conclusions of this study are subject to several limitations. First, Theorem 1 provides a sufficient, rather than necessary, condition for differential confinement. Its application requires global bounds on the growth of the data-loss gradient and Hessian trace, together with $\lambda > A_1$ and $\gamma_{s,\lambda} > 0$. These conditions were not established analytically for GPT-Neo-125M. The experiments resolve the corresponding gradient and trace contributions at finite parameter points, but they do not estimate the global constants A_1 and B_1 or verify the asymptotic lower bound in Eq. (43). Future theoretical work should derive architecture-aware growth bounds for attention, feed-forward, residual, embedding, and normalization operations and determine when these bounds satisfy the proposed criterion.

Second, the numerical analysis is restricted to finitely many directions and radial displacements. The sampled lower envelope in Eq. (64) is not the infimum over the unit sphere, and finite-radius behavior does not determine the global limit in Eq. (21). The substantial directional variation in Fig. 3 further shows that medians and averages can conceal adverse directions. More systematic searches could combine projected optimization on the unit sphere, curvature-guided initialization, parameter-group constraints, and adaptive screening for low-diagnostic paths. Such procedures would strengthen finite directional evidence without converting it into an asymptotic certificate.

Third, the diagnostic depends on the chosen Euclidean parameterization. Parameter norms, gradients, Hessian traces, and directional profiles are not invariant under general rescaling or other function-preserving reparameterizations of a neural network. The present comparisons therefore use matched architectures and parameterization

conventions. Future work could investigate block-normalized, layer-normalized, or metric-aware variants of the diagnostic and examine which conclusions remain stable under common Transformer reparameterizations.

Fourth, stochastic trace estimation and mini-batch evaluation introduce measurement uncertainty. The convergence analysis in Fig. 7 and Table 2 supports the selected probe counts for the evaluated checkpoints, but it does not establish a universal probe requirement. The longitudinal analysis also uses fewer probes than the principal directional analysis. Larger independent probe sets, additional evaluation batches, and hierarchical uncertainty models could improve the separation of probe, data-sampling, and training-run variability.

Fifth, the principal experiments are limited to GPT-Neo-125M, Penn Treebank, and WikiText-103, with a separate ViT-B/16 parameter-group extension. The observed geometric patterns should therefore not be assumed to generalize across model scales, architectures, datasets, or training objectives. Evaluation on larger autoregressive models, encoder-only Transformers, vision and multimodal architectures, and instruction-tuned checkpoints would clarify which findings persist across settings. The groupwise formulation in Eq. (82) also provides a basis for comparing the contributions of attention, feed-forward, embedding, and normalization blocks as model scale increases.

Sixth, the experiments cover finite grids of regularization coefficients, temperatures, optimizer conditions, and checkpoints. In particular, the coupled-L^2 and AdamW conditions in Fig. 5 use the same nominal coefficient, although the two implementations do not define equivalent optimization dynamics. Future comparisons could match effective parameter-norm contraction, validation performance, or update magnitude rather than the numerical coefficient alone. A broader joint analysis of s and λ would also help relate empirical temperature sensitivity to the admissible range in Corollary 2.1.

Finally, the diagnostic is not a validated surrogate for generalization, a log-Sobolev constant, or a finite-time Langevin mixing rate. The checkpoint-level relationship in Fig. 9 is descriptive and includes repeated observations from the same training runs. Establishing an association with validation performance would require an inferential model that accounts for training stage, regularization condition, dataset, and within-run dependence. Likewise, translating finite diagnostic values into functional inequalities or mixing guarantees would require additional results whose assumptions and normalizations match the regularized objective studied here.

These limitations motivate two complementary directions. The primary theoretical objective is to derive verifiable derivative-growth bounds for concrete Transformer parameterizations and determine whether the sufficient coefficient $\gamma_{s,\lambda}$ can be positive in relevant regimes. The primary computational objective is to develop adaptive low-diagnostic direction searches, parameterization-aware formulations, broader model and dataset evaluations, and more complete uncertainty quantification. Together, these developments could determine when finite confinement-compatible behavior extends across directions and when it can support a global differential-confinement result.

7 Methods and Reproducibility

This section describes the models, datasets, training conditions, geometric measurements, uncertainty analyses, and computational environment used in the experiments. All diagnostic quantities were computed from saved checkpoints using fixed evaluation batches and prespecified random seeds. Raw measurements were stored before aggregation into figures and tables.

7.1 Model and datasets

The principal experiments used the pretrained `EleutherAI/gpt-neo-125m` causal language model. The model was loaded using eager attention, and key–value caching was disabled during geometric evaluation. All unique trainable parameters were included in the parameter vector. The tied token-embedding and language-model-head weights were counted once.

Experiments were conducted on Penn Treebank and WikiText-103. Penn Treebank was loaded from the `ptb_text_only` repository using the `penn_treebank` configuration. WikiText-103 was loaded from the `Salesforce/wikitext` repository using the `wikitext-103-raw-v1` configuration, with the training corpus restricted to 120,000 source documents.

The GPT-Neo tokenizer was used for both datasets. When no padding token was defined, the end-of-sequence token was assigned as the padding token. Documents were concatenated with end-of-sequence separators and divided into non-overlapping sequences of 512 tokens. The training objective was the mean causal cross-entropy per predicted token. Training batches contained eight sequences, corresponding to 4096 tokens per update. Geometric evaluation used fixed batches of four sequences, or 2048 tokens. Validation blocks were sampled without replacement using deterministic seeds. Their block indices and identifiers were stored so that the same evaluation batches could be reconstructed across experimental modules.

7.2 Controlled training protocol

The controlled GPT-Neo experiments compared Adam without regularization, Adam with coupled L^2 regularization, and AdamW with decoupled weight decay. In the coupled condition, the optimized objective was Eq. (12), with the quadratic penalty applied to all unique trainable parameters. AdamW applied the same nominal coefficient through a decoupled parameter update and did not include the quadratic term in the differentiable loss.

The regularization grid was $\lambda \in \{0, 10^{-4}, 10^{-3}, 10^{-2}\}$. Each controlled condition used three independent training seeds. Matched conditions used the same pretrained initialization protocol, data-ordering procedure, batch construction, learning-rate schedule, numerical precision, and parameter-coverage policy. Training was performed for 1200 updates using a base learning rate of 5×10^{-5}. The first 60 updates used linear warm-up, followed by cosine learning-rate decay. Adam and AdamW used $(\beta_1, \beta_2) = (0.9, 0.999)$ and $\epsilon = 10^{-8}$. The gradient norm was clipped at 1.0 before each optimizer update. Checkpoints were saved at steps $\{0, 75, 150, 300, 600, 900, 1200\}$. With 4096 tokens per update, the final checkpoint corresponds to approximately

4.92×10^6 processed tokens. Validation cross-entropy was evaluated using 16 validation batches at every saved checkpoint. The checkpoint schedule was specified before the geometric measurements and did not depend on the observed diagnostic values.

7.3 Differential measurements

At each parameter vector and evaluation batch, the implementation computed the data-loss cross-entropy, $\|\nabla\mathcal{L}(\theta)\|^2$, $\langle\nabla\mathcal{L}(\theta), \theta\rangle$, and $\|\theta\|^2$. The data-loss Hessian trace was estimated using Hessian–vector products, after which the complete diagnostic and its five components were assembled using Eqs. (66) and (78). The data-loss gradient and Hessian trace were calculated once at each parameter point. Diagnostic values for different λ and s were then assembled analytically from the same measurements. The temperature grid was $s \in \{0.25, 0.5, 1, 2, 4\}$, with $s = 1$ used for the principal visualizations.

The regularized-gradient norm was evaluated directly as $\|\nabla\mathcal{L}(\theta) + \lambda\theta\|^2$ and independently reconstructed from the componentwise expansion. The two forms were required to agree within a relative tolerance of 10^{-6}, measured against the sum of the absolute component magnitudes. Additional checks verified the exact regularizer-trace and radial contributions, the sign of the gradient-interaction term, and the monotonic ordering with respect to s in Eq. (90). Model parameters and automatic-differentiation operations used 32-bit floating-point precision, while scalar accumulations used 64-bit precision. TensorFloat-32 matrix multiplication and cuDNN benchmarking were disabled to reduce numerical and execution variability.

7.4 Directional and radial analysis

The principal directional analysis used the coupled-L^2 checkpoint at training step 1200 as the reference vector $\theta^{(0)}$. The direction set contained 42 unit vectors: 32 isotropic Gaussian directions, two gradient-aligned directions, two Hessian-aligned directions, four parameter-group directions, and two screened low-diagnostic directions.

The parameter-group directions were restricted to attention, feed-forward, embedding, and normalization coordinates. The attention group included query, key, value, and output projections; the feed-forward group included both multilayer-perceptron projections; the embedding group included token and positional embeddings; and the normalization group included all LayerNorm parameters.

Hessian-aligned directions were obtained by matrix-free Lanczos iterations with full reorthogonalization. The low-diagnostic directions were selected from an independent pool of 24 random candidates. Screening used a separate validation batch, eight Rademacher probes, $s = 1$, $\lambda = 10^{-3}$, and displacement $r = 0.2\|\theta^{(0)}\|$. The two candidates with the smallest screening diagnostic were retained. Candidate seeds were disjoint from the seeds used for the displayed random directions, and screening did not use the batch employed for the final profiles.

The evaluated radial fractions were $\{0, 0.001, 0.003, 0.01, 0.03, 0.1, 0.2, 0.4, 0.8, 1.6\}$. For a fraction f, the displacement was $r = f\|\theta^{(0)}\|$. Directions were normalized in the full parameter space, and $\|\theta^{(0)} + ru\|^2$ was calculated explicitly at every point.

Finite parameter points were classified using their data cross-entropy. A point was labelled `near_reference` when its loss did not exceed 1.5 times the reference loss.

A larger finite perturbation with loss below 10.825, the uniform-prediction reference for the GPT-Neo vocabulary, was labelled `perturbed_below_uniform`. Values from 10.825 through 100 were classified as `above_uniform_plateau`, and values greater than 100 as `degenerate_large_radius`. Directional quantiles and the sampled lower envelope used the first two categories; all finite values were retained in the complete radial sweep. This classification is an operational convention for finite experimental interpretation, not a theoretical domain of validity for the diagnostic.

7.5 Hessian-trace and spectral estimation

The data-loss Hessian was not formed explicitly. For each probe vector v, the quadratic form $v^\top \nabla^2 \mathcal{L}(\theta) v$ was evaluated through automatic differentiation. The first-order gradient graph was constructed once at each parameter point and reused for the corresponding second-order evaluations. The exact regularizer contribution λd was added analytically and was not estimated stochastically.

Rademacher probes were used for the principal analyses. Directional measurements used $M = 64$ probes, while the checkpoint-evolution analysis used $M = 32$. Probe-count reliability was evaluated for $M \in \{8, 16, 32, 64, 128\}$. At the final checkpoint, ten independent Rademacher probe sets were evaluated for each displayed condition; four independent sets were used at non-final checkpoints. Gaussian probes were also evaluated at the final checkpoint as a secondary comparison.

Within each probe set, the trace estimate was the mean of the probe-level quadratic forms, and its standard error was calculated using Eq. (75). Dispersion across independent probe sets was calculated separately. The confidence intervals in Table 2 used the t distribution and the standard error of the mean across the ten independent probe-set estimates.

Selected data-loss Hessian eigenvalues were estimated using 40 matrix-free Lanczos iterations with full reorthogonalization at training steps 0, 300, and 1200. The spectral analysis used training seed 0 and retained the six largest and six smallest algebraic Ritz values. Regularized-objective eigenvalues were obtained using the exact shift in Eq. (14). Ritz residual norms were stored to assess the numerical quality of the approximations.

7.6 Training evolution and validation association

The longitudinal analysis evaluated every saved checkpoint from the controlled training runs. Two disjoint, fixed validation batches were used for each dataset and checkpoint. The diagnostic was computed over the complete λ and s grids. The principal training-evolution analysis retained the value of λ used to train the corresponding checkpoint and used $s = 1$.

Measurements were first averaged across the two evaluation batches within each training seed and checkpoint. Means and standard deviations were then calculated across the three independent seeds. Probe uncertainty, evaluation-batch variation, and training-seed variation were retained as separate quantities.

The association with validation perplexity used one observation per training run and scheduled checkpoint after averaging across the two fixed evaluation batches.

Checkpoints from the same run were treated as repeated observations. The visualization connected within-run trajectories and did not fit a pooled regression across regularization conditions. The analysis was interpreted descriptively because both the diagnostic and perplexity change with training stage and regularization condition.

7.7 Optimizer and parameter-group comparisons

The optimizer comparison used unregularized Adam, coupled-L^2 Adam, and AdamW conditions with three independent seeds per condition. The coupled-L^2 and AdamW conditions used the common nominal coefficient $\lambda = 10^{-3}$. For the AdamW checkpoints, the diagnostic was evaluated post hoc using this value of λ and was not interpreted as the objective minimized during training.

The parameter-group extension used a ViT-B/16 architecture trained on ImageNet-1K with AdamW and decoupled weight decay. Training used FP16 precision and a learning-rate schedule consisting of linear warm-up followed by cosine decay. The regularization coefficients were $\lambda \in \{0, 10^{-4}, 10^{-3}, 10^{-2}\}$, representing separately trained weight-decay conditions. Parameters within each Transformer layer were partitioned into query, key, value, feed-forward, and LayerNorm groups. Groupwise trace estimates were computed using $M = 64$ Rademacher probes and Hessian–vector products implemented with BackPACK. Groupwise trace and gradient contributions were combined according to Eq. (85).

The displayed heatmaps use the absolute-magnitude shares in Eq. (87) and report the resulting normalized attribution separately for each layer and parameter group. Because these shares discard sign, they are interpreted as descriptive relative magnitudes rather than probabilities, signed effects, or causal attributions. Because AdamW applies decoupled weight decay, the groupwise regularized-objective diagnostic is interpreted as a post-hoc geometric evaluation of the resulting checkpoints rather than as the objective directly minimized by AdamW.

7.8 Computational environment and provenance

Experiments were executed using Python 3.12.2 and PyTorch on a workstation with an Intel Xeon W5-3435X processor, 128 GB of system memory, and an NVIDIA RTX 6000 Ada GPU with 48 GB of video memory. GPU acceleration was used for training, gradient evaluation, Hessian–vector products, Lanczos iterations, and stochastic trace estimation.

A master seed of 20260101 was used to derive deterministic seeds for training, data ordering, evaluation-batch construction, direction generation, candidate screening, Lanczos initialization, probe sets, and individual probes. Seeds were applied to Python, NumPy, PyTorch, and all available CUDA devices.

Each checkpoint was stored with its training condition, regularization implementation, training seed, processed-token count, parameter dimension, validation loss, parameter norm, and SHA-256 hash. Direction catalogues, screening results, evaluation-batch identifiers, probe seeds, resolved configurations, and consistency-check outcomes were stored with the corresponding measurements.

Each experimental module wrote canonical raw measurements before generating summaries, tables, and figures. Long-running modules wrote results incrementally to disk. Environment manifests recorded the operating platform, Python and PyTorch versions, CUDA build, hardware properties, master seed, sequence length, batch sizes, loss-normalization convention, and installed package versions. Figures were exported as PDF and 400-dpi PNG files using a consistent colorblind-safe palette.

8 Conclusion

This study resolves a central ambiguity in the analysis of regularized Transformer objectives: quadratic coercivity does not imply differential confinement. Positive coupled L^2 regularization guarantees that a non-negative cross-entropy objective is coercive and that its associated Gibbs density is normalizable, but these function-level properties do not control the parameter-space derivatives required by the confinement condition.

The bilinear-logit counterexample establishes that bounded inputs, smooth model operations, and positive quadratic regularization are insufficient to guarantee differential confinement. In particular, the data-loss Hessian trace can outweigh the favorable squared-gradient contribution along selected parameter-space directions. Theorem 1 provides the corresponding positive result: confinement follows under explicit gradient- and Hessian-trace-growth bounds when the regularization strength and temperature satisfy the sufficient condition $\gamma_{s,\lambda} > 0$. The theory therefore replaces an unconditional claim with a precise criterion describing when quadratic regularization can dominate the asymptotic derivative geometry.

The computational framework makes this theoretical balance measurable in large models using exact gradients, Hessian–vector products, and stochastic trace estimation. Across GPT-Neo-125M checkpoints and sampled directions, the diagnostic exhibits substantial dependence on direction, radial displacement, temperature, and regularization implementation. Within the investigated near-reference regime, the negative data-loss Hessian-trace contribution and the constant regularizer trace $-\lambda d$ dominate the positive radial term. The diagnostic remains negative at the evaluated checkpoints and does not show uniform quadratic growth across the sampled finite-radius paths. These observations do not determine the global asymptotic limit, but they identify why the sufficient theoretical balance is not demonstrated in the measured regime.

The experiments further show that coupled L^2 regularization and decoupled AdamW weight decay are not interchangeable under adaptive optimization. At the same nominal coefficient, the two implementations produce different training trajectories, parameter norms, validation performance, curvature statistics, and diagnostic values. The spectral and parameter-group analyses additionally reveal geometric variation that is concealed by a single global scalar, while the probe-convergence analysis demonstrates that the required second-order quantities can be estimated reliably at the investigated model scale.

Taken together, the results support a conditional rather than universal account of differential confinement in Transformer parameter spaces. Quadratic regularization

supplies the radial mechanism needed for favorable asymptotic growth, but whether this mechanism dominates depends jointly on the data-loss gradient, Hessian trace, temperature, direction, and learned parameter geometry. Finite measurements cannot replace the global derivative-growth analysis required by the theorem, but they can identify the mechanisms and parameter-space directions that support or obstruct confinement-compatible behavior.

By separating objective-level coercivity from derivative-level confinement, this work provides both a rigorous theoretical criterion and a scalable computational methodology for studying confinement-related parameter-space geometry in large neural networks. The resulting framework offers a systematically documented foundation for determining when regularized Transformer objectives satisfy, or fail to support, the geometric assumptions used in functional-analytic and Langevin-based analyses.

Data and Code Availability

The processed experimental data, resolved configurations, supporting outputs, and implementation used in this study are available from the corresponding author upon reasonable request, subject to applicable GE HealthCare policies and licensing restrictions.

Funding

The authors received no specific funding for this work.

References

[1] H. Avron and S. Toledo. Randomized Algorithms for Estimating the Trace of an Implicit Symmetric Positive Semi-Definite Matrix. *Journal of the ACM*, 58(2): 8:1–8:34, 2011. doi: 10.1145/1944345.1944349.

[2] L. Chizat. Mean-Field Langevin Dynamics: Exponential Convergence and Annealing. *Transactions on Machine Learning Research*, 2022.

[3] A. Dosovitskiy et al. An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*, 2021.

[4] A. Gulati et al. Conformer: Convolution-Augmented Transformer for Speech Recognition. In *Proceedings of Interspeech 2020*, pages 5036–5040, 2020. doi: 10.21437/Interspeech.2020-3015.

[5] M. F. Hutchinson. A Stochastic Estimator of the Trace of the Influence Matrix for Laplacian Smoothing Splines. *Communications in Statistics—Simulation and Computation*, 19(2):433–450, 1990. doi: 10.1080/03610919008812866.

[6] S. Kobayashi, Y. Akram, and J. von Oswald. Weight Decay Induces Low-Rank Attention Layers. In *Advances in Neural Information Processing Systems*,

volume 37, pages 4481–4510. Curran Associates, Inc., 2024. doi: 10.52202/079017-0146.

[7] A. Krogh and J. A. Hertz. A Simple Weight Decay Can Improve Generalization. In *Advances in Neural Information Processing Systems*, volume 4, pages 950–957. Morgan Kaufmann, 1991.

[8] I. Loshchilov and F. Hutter. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*, 2019.

[9] A. Nitanda et al. Propagation of Chaos for Mean-Field Langevin Dynamics and Its Application to Model Ensemble. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 46586–46610. PMLR, 2025.

[10] B. A. Pearlmutter. Fast Exact Multiplication by the Hessian. *Neural Computation*, 6(1):147–160, 1994. doi: 10.1162/neco.1994.6.1.147.

[11] M. Raginsky, A. Rakhlin, and M. Telgarsky. Non-Convex Learning via Stochastic Gradient Langevin Dynamics: A Nonasymptotic Analysis. In *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pages 1674–1703. PMLR, 2017.

[12] A. Vaswani et al. Attention Is All You Need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008. Curran Associates, Inc., 2017.

[13] C. Villani. *Optimal Transport: Old and New*, volume 338 of *Grundlehren der Mathematischen Wissenschaften*. Springer, Berlin, Heidelberg, 2009. doi: 10.1007/978-3-540-71050-9.

[14] M. Welling and Y. W. Teh. Bayesian Learning via Stochastic Gradient Langevin Dynamics. In *Proceedings of the 28th International Conference on Machine Learning*, pages 681–688, New York, NY, USA, 2011. ACM.