# INFERENCE FOR TRANSFORMER REPRESENTATION SHIFTS THROUGH METRIC FIELDS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

How can we determine whether a transformer has undergone a systematic shift in its internal representations? Existing methods typically summarize activations over a fixed collection of inputs, but do not directly quantify whether an observed difference persists across a population of input sequences. We formulate representation-shift detection as a population-level inference problem. Each randomly sampled sequence induces a hidden-state surface over layer depth and token position, which we treat as a random functional object and characterize through its induced metric tensor field. This representation is invariant to fixed global orthogonal transformations while preserving the spatial organization of model representations. Using paired inputs, we define the expected difference between metric fields over a prespecified task distribution as the inferential target and develop a global spectral test for detecting systematic shifts. In a controlled study of continued training, replacing only 1% of the training corpus with data from a different source produces a detection signal more than two orders of magnitude larger than that under the batch-order control. We further introduce a tolerance-based screening rule to distinguish small but detectable departures, such as those arising from uncontrollable noise, from substantively larger representation shifts. Together, these results provide a statistical framework for detecting, quantifying, and spatially characterizing transformer representation shifts at the population level.

## 1 INTRODUCTION

When comparing two models, or asking whether a model has changed after training, we first need to decide what it means for them to be different. A model may behave differently on some inputs but similarly on others, so a comparison based on a fixed collection of examples does not by itself describe a general model change. We therefore view model comparison relative to a population of inputs and ask whether a systematic difference persists across that population. We treat the models being compared as fixed and the input sequence as the source of variation. Each input induces a collection of hidden states across model depth and token position. As inputs vary over an evaluation population, these internal representations vary as well, providing a population of input-induced representations on which model comparison can be based.

Existing methods for comparing neural representations typically operate on activation matrices collected from a fixed set of inputs and return a layerwise or model-level similarity score. In this formulation, the input set is usually treated as fixed rather than as a sample from an explicit population, so the comparison does not directly specify a population estimand or a source of sampling variability. Consequently, it is difficult to distinguish a systematic representation shift that persists across inputs from differences driven by the particular examples being evaluated, or to attach population-level uncertainty to the comparison. Moreover, collapsing activations into layerwise or global similarity scores discards the prompt-level variation and joint organization of representation change across layer depth and token position. What is missing is therefore not another similarity measure, but a framework in which each input provides a statistical observation, model change is defined as a population quantity, and systematic representation drift can be estimated and tested.

We address this by treating the hidden states induced by each input as a functional data object over a continuous domain of layer depth and token position. For each input, a forward pass produces hidden states across layers and token positions. Therefore, each collection of hidden states can be treated as



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

a random observation of the model, while layers and token positions serve as ordered coordinates within that observation rather than as independent replicates. This distinction makes representation comparison a population-level problem over random inputs and provides a clear inferential target.

The layer and token indices also have genuine spatial organization. Adjacent layers represent successive stages of computation, while adjacent token positions represent neighboring locations in context. For that matter, our functional data approach naturally preserves this ordering and makes local variation across depth and context part of the representation itself. This framework also provides a common domain for different layer counts, sequence lengths, and observation resolutions. Directly comparing the resulting hidden-state surfaces remains difficult because raw activations are high-dimensional. We therefore represent each surface through its induced metric field, which provides effective dimension reduction and facilitates further statistical operations. This representation is compact and invariant to global orthogonal changes of hidden coordinates, making it more comparable across models while retaining the spatial structure needed.

Our work is to move representation comparison from fixed scalar measures or similarity comparisons to metric fields as random functional observations. From this formulation, three contributions follow:

- **Random metric-field representation.** We formalize hidden states as random surfaces over normalized layer depth and token position, and summarize their local geometry through a three-component metric field. The field records depthwise variation, tokenwise variation, and their coupling. It is invariant to fixed global orthogonal changes of hidden coordinates and retains the spatial organization of these quantities on a common layer–token domain, regardless of hidden-state width.
- **Inference for systematic shifts.** For two fixed models evaluated on paired inputs sampled from a prespecified task distribution, we define systematic geometric change through the expected metric-field difference. These paired random fields, or the difference between them, provide the replication needed to quantify input-to-input variability and to generalize the comparison beyond the particular task instances observed. We then develop a global testing procedure for determining whether this systematic shift in model representation is distinguishable from noninformative variability. To solidify the inference procedure, it is accompanied by a derived null distribution and asymptotic guarantees.
- **Controlled validation for training monitoring.** We evaluate the framework through a controlled dose–response study: starting from a common checkpoint, we vary the proportion of continuation-training data drawn from a second corpus and compare matched checkpoints on identical held-out inputs. Execution and batch-order controls provide reference levels for run-to-run variation. At the final checkpoint, the metric-field statistic increases monotonically across the six tested substitution doses and is $184$ times its batch-order-control value under $1\%$ substitution. Across these doses, its control-normalized ratios exceed those of the evaluated similarity-based diagnostics (Table 1).

## 2 Related work

**Model representation** A broad literature compares neural representations calculated by a shared collection of inputs. Canonical-correlation methods such as SVCCA and PWCCA identify correlated representational subspaces while reducing sensitivity to coordinate-level differences (Raghu et al., 2017; Morcos et al., 2018), whereas centered kernel alignment (CKA) compares centered Gram structure across activation matrices (Kornblith et al., 2019). Generalized shape metrics define distances after quotienting out specified transformation classes (Williams et al., 2021). Recent work makes the role of invariance and alignment increasingly explicit. Cannistraci et al. (2024) construct latent representations incorporating products of transformation invariances, while Insulla et al. (2025) connect metric, probabilistic, and spectral notions of representation alignment and relate stitching to kernel alignment. Moving beyond aggregate activation-set comparisons, Jiang et al. (2025) use sample-wise similarity to trace representation progression through transformer layers, and Luthra et al. (2026) study training-time alignment under shared initialization, batches, and data randomness. These methods characterize whether activations and the learned spaces are aligned under a selected transformation class. Our individual unit of analysis is instead an independent prompt. Any follow-up comparison is formulated through the population mean of paired field differences. The proposed framework is therefore not another scalar valued measure for similarity in representations.

**Geometry and dynamics of representations.** Hidden state geometry has been studied both to characterize what neural representations encode and to understand how computation evolves through model depth. Contextual representations exhibit anisotropy and organize semantic and syntactic information geometrically (Ethayarajh, 2019; Reif et al., 2019), while fine-tuning can substantially reorganize the geometry of intermediate representations (Merchant et al., 2020; Durrani et al., 2022). Layerwise methods provide a complementary view of transformer computation. For example, Jelenić et al. (2024) use the smoothness of transformations between successive transformer layers as an internal diagnostic for out-of-distribution inputs. The normalized transformer of Loshchilov et al. (2025) places hidden states and model components on unit hyperspheres and interprets transformer layers as successive displacements of normalized representations. More generally, pullback geometry characterizes how a vector-valued map locally stretches and couples directions in its input domain and has been used to analyze learned generative and autoencoding maps (Arvanitidis et al., 2018; Nazari et al., 2023). We similarly construct the proposed metric field from hidden states. For each prompt, the resulting three-component field records directional variation and is invariant to a fixed global orthogonal reparameterization. Unlike methods designed to recover encoded content or describe layer progression within one model, our objective is to estimate the systematic geometric difference over both the token and layer dimensions.

**Inferential framework.** Zhong & Raghunathan (2026) identify singular directions of the weight difference between a base model and its fine-tuned counterpart and use activation responses along those directions to detect newly acquired behaviors. Orgad et al. (2025) show that hidden states can encode information that is not expressed by the model's generated output, motivating a direct comparison between representation drift and output quantities such as negative log-likelihood, perplexity, and output-distribution divergence. Our inferential formulation draws on random-function and covariance-operator theory (Ramsay & Silverman, 2005; Hsing & Eubank, 2015) and functional mean inference (Horváth et al., 2013). Kernel and energy two-sample procedures provide alternatives for testing equality of complete distributions of complex observations (Gretton et al., 2012; Székely & Rizzo, 2013). We target a direct and interpretable quantity for inference. Given two fixed checkpoints of a model and paired prompts randomly drawn from a prespecified evaluation distribution, we test whether the expected metric-field difference is zero and report its spatial pattern.

## 3 RANDOM METRIC FIELDS

Processing a token sequence produces a two-way array of hidden states indexed by model depth and token position. Variation along the depth axis describes how the model updates its representation from layer to layer, whereas variation along the token axis describes how information is organized across the sequence. We treat this array as a discrete observation of a surface over the layer–token domain. The local behavior of this surface reveals how strongly representations change in each direction and how those directional changes interact.

Direct comparison of hidden-state surfaces is difficult because their dimensions depend on model width and architecture. Although a shared continuous layer–token domain accommodates different context lengths and observation grids, the hidden-state vectors themselves remain high-dimensional and model-specific. We therefore represent the local geometry of each surface by a metric tensor field. At every point in the domain, the tensor summarizes the magnitude of representational change along depth and token position and the coupling between those changes. The resulting three-component field is defined on a common domain regardless of hidden-state dimension, making it suitable for cross-model comparison and statistical inference.

Fix a task family $a \in \mathcal{A}$ and a nominal context length $T_0$. Let

$$Z_i \overset{\text{iid}}{\sim} \mathcal{P}_{a,T_0}, \qquad i = 1, \ldots, n,$$

where $Z_i$ denotes a random task instance. Given $Z_i$, a deterministic construction rule produces the token sequence

$$S_i = \kappa_{a,T_0}(Z_i).$$

Thus, in the present setup, stochasticity enters only through the sampled task instance. The tokenizer, model parameters, inference mode, and hidden-state extraction rule are treated as fixed.

Let $\mathcal{U} = [0,1]^2$ be the normalized layer–token domain, with $u = (\ell, t)$. For model $m$ and input sequence $i$, the hidden states are observed as an array indexed by layer depth and token position. We regard this array as finite-grid observations of a vector-valued surface

$$X_{i,m} : \mathcal{U} \to \mathbb{R}^{d_{\mathrm{hid},m}},$$

constructed through a fixed registration and interpolation procedure. The layer coordinate describes how representations evolve through the model, whereas the token coordinate describes how they vary across sequence positions. This formulation separates the underlying functional object from the model-specific observation grid and therefore accommodates different layer counts, context lengths, nonuniform token grids, and partially observed positions. We further assume that $X_{i,m} \in W^{1,2}(\mathcal{U}; \mathbb{R}^{d_{\mathrm{hid},m}})$, the vector-valued Sobolev space in which derivatives along the layer and token directions are well defined. Here, $d_{\mathrm{hid},m}$ denotes the hidden-state dimension of model $m$.

Our primary object of interest is the geometry that the random surface $X_{i,m}$ induces on $\mathcal{U}$. The proposed metric tensor field records the strength of representational change in each coordinate direction and the coupling between those changes. The Jacobian describes the local variation of the hidden-state surface:

$$J_{i,m}(u) = [\partial_\ell X_{i,m}(u) \quad \partial_t X_{i,m}(u)] \in \mathbb{R}^{d_{\mathrm{hid},m} \times 2},$$

which induces a pullback metric $g_{i,m}(u)$ as

$$g_{i,m}(u) = J_{i,m}(u)^\top J_{i,m}(u) \in \mathbb{S}^2_+, \qquad u \in \mathcal{U}, \tag{1}$$

where $\mathbb{S}^2_+$ denotes the cone of $2 \times 2$ symmetric positive-semidefinite matrices. Equivalently,

$$g_{i,m}(u) = \begin{pmatrix} g_{i,m,\ell\ell}(u) & g_{i,m,\ell t}(u) \\ g_{i,m,\ell t}(u) & g_{i,m,tt}(u) \end{pmatrix}, \tag{2}$$

with entries $g_{i,m,\ell\ell}(u) = \|\partial_\ell X_{i,m}(u)\|^2$, $g_{i,m,tt}(u) = \|\partial_t X_{i,m}(u)\|^2$, and $g_{i,m,\ell t}(u) = \langle \partial_\ell X_{i,m}(u), \partial_t X_{i,m}(u) \rangle$. The three unique entries have direct interpretations: $g_{i,m,\ell\ell}$ measures local variation along model depth, $g_{i,m,tt}$ measures local variation along token position, and $g_{i,m,\ell t}$ measures the local coupling between these two directions. Thus, $g_{i,m}(u)$ summarizes how the hidden-state surface locally changes over the layer–token domain. The Jacobian therefore encodes local directional change, and the pullback metric converts that change into an induced notion of distance and stretching on the shared layer–token domain. In the experiments, we apply this construction to the surface formed from unit-normalized hidden states, replacing $X_{i,m}$ by $\widetilde{X}_{i,m}$ in the Jacobian. The resulting metric therefore characterizes local variation in hidden-state direction rather than activation magnitude. Details of implementation and calculation are in Appendix B.

Since $g_{i,m}(u)$ is symmetric, it can be represented by its three unique entries. We define

$$\mathrm{vech} \begin{pmatrix} a & b \\ b & c \end{pmatrix} = (a, b, c)^\top, \qquad \gamma_{i,m}(u) = \mathrm{vech}\{g_{i,m}(u)\} \in \mathbb{R}^3. \tag{3}$$

Thus, $\gamma_{i,m} : \mathcal{U} \to \mathbb{R}^3$ is the metric-field representation associated with model $m$ and task instance $Z_i$. We use the unscaled vech coordinates and assign equal weight to the three interpretable components $g_{\ell\ell}$, $g_{\ell t}$, and $g_{tt}$ in the functional inner product below. The resulting representation has the same three-component form regardless of the hidden-state dimension.

## 4 Statistical inference on metric tensor fields

The metric field introduced in Section 3 provides a geometric representation of how a fixed language model maps the shared layer–token domain into hidden-state space. To compare two models, we analyze the paired difference between their metric fields evaluated on the same task instance. Fix a contrast label $c = (m_+, m_-)$, where $m_+$ and $m_-$ denote the two models being compared. For task instance $Z_i$, define the metric-difference field

$$D_i^c(u) = \gamma_{i,m_+}(u) - \gamma_{i,m_-}(u) = \mathrm{vech}\{g_{i,m_+}(u) - g_{i,m_-}(u)\}, \qquad u \in \mathcal{U}. \tag{4}$$

The ordering of $c$ determines the direction of the contrast. After vectorization, $D_i^c$ is simply a three-dimensional functional observation over $\mathcal{U}$. The population quantity of interest is the mean metric contrast

$$\mu^c(u) = \mathbb{E}_{Z \sim \mathcal{P}_{a,T_0}}\{D^c(u)\}, \tag{5}$$

which represents the systematic geometric difference between the two models over the task distribution. Individual fields $D_i^c$ may vary across task instances, but a nonzero $\mu^c$ indicates that the geometric difference persists on the population level. Thus, for each fixed contrast $c$, detecting systematic change in the metric field becomes a one-sample functional mean inference problem based on $D_1^c, \ldots, D_n^c$. The functional central limit theorem below provides the large-sample approximation for the sample mean field $\bar{D}_n^c = n^{-1} \sum_{i=1}^n D_i^c$.

### 4.1 Mean-field estimation and asymptotics

We treat each observed metric-difference field $D_i^c$ as a random element of the Hilbert space

$$\mathcal{H} = L^2(\mathcal{U}; \mathbb{R}^3), \qquad \langle f, h \rangle_{\mathcal{H}} = \frac{1}{3} \int_{\mathcal{U}} f(u)^\top h(u)\, du, \tag{6}$$

with norm $\|f\|_{\mathcal{H}}^2 = \langle f, f \rangle_{\mathcal{H}}$. We chose the equally weighted inner product for simplicity, while other choices are possible.

The population mean $\mu^c$ was defined above. The covariance kernel is

$$\Sigma^c(u, v) = \mathbb{E}_{Z \sim \mathcal{P}_{a,T_0}} \left[ \{D_i^c(u) - \mu^c(u)\} \{D_i^c(v) - \mu^c(v)\}^\top \right], \qquad u, v \in \mathcal{U}, \tag{7}$$

and the corresponding covariance operator $\mathcal{K}^c : \mathcal{H} \to \mathcal{H}$ is

$$(\mathcal{K}^c f)(u) = \frac{1}{3} \int_{\mathcal{U}} \Sigma^c(u, v) f(v)\, dv. \tag{8}$$

**Assumption 1.** *For each fixed contrast c, the fields $D_1^c, \ldots, D_n^c$ are i.i.d. random elements of $\mathcal{H}$ satisfying*

$$\mathbb{E}\|D_i^c\|_{\mathcal{H}}^2 < \infty.$$

The independence requirement concerns sampled task instances, not the layer or token coordinates within each field. With the models, tokenization, and field-construction procedures fixed, independently sampled instances from the same task induce i.i.d. random functional objects. Furthermore, the finite second-moment condition implies that the covariance operator $\mathcal{K}^c$ is self-adjoint, positive semidefinite, and trace class, with

$$\operatorname{tr}(\mathcal{K}^c) = \mathbb{E}\|D_i^c - \mu^c\|_{\mathcal{H}}^2 < \infty.$$

Consequently, $\mathcal{K}^c$ is compact and admits the spectral decomposition used in the later part of the testing procedure.

To transfer the sample-mean limit from the ideal fields to their numerical approximations, we impose the following condition.

**Assumption 2.** *When $D_i^c$ is replaced by its observed finite-difference version, we additionally assume that*

$$\sqrt{n} \left\| \frac{1}{n} \sum_{i=1}^n \left( \widehat{D}_i^c - D_i^c \right) \right\|_{\mathcal{H}} = o_p(1).$$

The approximation condition ensures that reconstruction error in the sample mean is asymptotically negligible relative to its $n^{-1/2}$ sampling errors.

**Theorem 1.** *Under Assumption 1, for each fixed contrast c, the sample mean field*

$$\bar{D}_n^c = n^{-1} \sum_{i=1}^n D_i^c \tag{9}$$

*satisfies*

$$\bar{D}_n^c \xrightarrow{p} \mu^c \qquad \text{in } \mathcal{H}. \tag{10}$$

*Furthermore,*

$$\sqrt{n}(\bar{D}_n^c - \mu^c) \xrightarrow{d} \mathbb{G}^c \qquad \text{in } \mathcal{H}, \tag{11}$$

*where $\mathbb{G}^c$ is a Gaussian process in $\mathcal{H}$ with zero mean and covariance operator $\mathcal{K}^c$.*

The theorem provides the asymptotic distribution (11) needed for inference on the mean metric-difference field (5). If Assumption 2 also holds, both conclusions remain valid with $\overline{D}_n^c$ replaced by $n^{-1} \sum_{i=1}^n \widehat{D}_i^c$. Section 4.2 uses this result to construct a global spectral test of the hypothesis that the mean contrast is zero. The proof appears in Appendix A.1.

### 4.2 GLOBAL SPECTRAL TEST FOR METRIC-FIELD CHANGE

Using the functional central limit theorem, we test whether two models differ systematically in their induced metric fields over the layer–token domain. For a fixed contrast $c$, the hypotheses are

$$H_0^c : \mu^c = 0_{\mathcal{H}} \qquad \text{versus} \qquad H_1^c : \|\mu^c\|_{\mathcal{H}} > 0. \tag{12}$$

We use the global statistic

$$T_n^c = n\|\bar{D}_n^c\|_{\mathcal{H}}^2 = \frac{n}{3}\int_{\mathcal{U}} \bar{D}_n^c(u)^\top \bar{D}_n^c(u)\, du. \tag{13}$$

Large values of $T_n^c$ indicate that the average metric-difference field is far from zero. The limiting null distribution is therefore a weighted chi-square mixture. When the global null is rejected, the estimated mean field $\bar{D}_n^c(u)$ can be inspected to characterize the spatial organization of the detected shift.

**Corollary 1.** *Under $H_0^c$ and Assumption 1,*

$$T_n^c \xrightarrow{d} \|\mathbb{G}^c\|_{\mathcal{H}}^2 = \sum_{r=1}^{\infty} \lambda_r^c \zeta_r^2, \tag{14}$$

*where $\zeta_r \overset{\text{i.i.d.}}{\sim} N(0,1)$ and $(\lambda_r^c, \phi_r^c)_{r\geq 1}$ are the eigenpairs of $\mathcal{K}^c$, with*

$$\mathcal{K}^c\phi_r^c = \lambda_r^c\phi_r^c, \qquad \langle \phi_r^c, \phi_s^c\rangle_{\mathcal{H}} = \mathbf{1}(r=s).$$

*Since $\mathcal{K}^c$ is trace class, $\sum_{r\geq 1}\lambda_r^c < \infty$.*

The limiting null distribution is therefore a weighted chi-square mixture. In practice, the eigenvalues are estimated from the empirical covariance operator. Define

$$\widehat{\Sigma}_n^c(u,v) = \frac{1}{n-1}\sum_{i=1}^{n}\{D_i^c(u) - \bar{D}_n^c(u)\}\{D_i^c(v) - \bar{D}_n^c(v)\}^\top, \tag{15}$$

and let $\widehat{\mathcal{K}}_n^c$ be the corresponding empirical covariance operator,

$$(\widehat{\mathcal{K}}_n^c f)(u) = \frac{1}{3}\int_{\mathcal{U}} \widehat{\Sigma}_n^c(u,v) f(v)\, dv.$$

Let $\widehat{\lambda}_1^c, \ldots, \widehat{\lambda}_R^c$ denote the retained empirical eigenvalues, where $R$ is chosen as the smallest integer for which the cumulative proportion of empirical variance explained is at least $95\%$. The resulting truncated spectral approximation is

$$Q_R^c = \sum_{r=1}^{R} \widehat{\lambda}_r^c \zeta_r^2, \qquad \zeta_r \overset{\text{i.i.d.}}{\sim} N(0,1). \tag{16}$$

Given $B_{\text{null}}$ Monte Carlo draws $Q_{R,1}^c, \ldots, Q_{R,B_{\text{null}}}^c$, the plug-in spectral p-value is

$$\widehat{p}_c = \frac{1 + \sum_{b=1}^{B_{\text{null}}} \mathbf{1}\{Q_{R,b}^c \geq T_{n,\text{obs}}^c\}}{B_{\text{null}} + 1}. \tag{17}$$

We reject $H_0^c$ at level $\alpha$ when $\widehat{p}_c \leq \alpha$.

**Proposition 1.** *Suppose Assumption 1 holds and the plug-in spectral approximation consistently estimates the null distribution in Corollary 1. Then the spectral test has asymptotic level $\alpha$ under $H_0^c$. Moreover, under any fixed alternative satisfying $\|\mu^c\|_{\mathcal{H}} > 0$, $\mathbb{P}_{H_1^c}(\widehat{p}_c \leq \alpha) \to 1$.*

Corollary 1 and Proposition 1 justify using the spectral statistic as a global test for model change. Proofs are given in Appendix A.2 and Appendix A.3. The next section will present the empirical studies.

# 5 EXPERIMENTS

We evaluate the testing procedure by using the proposed metric field $g$ against other baselines, including the raw hidden-state surface $X_{i,m}(u)$ and its norm as well as other similarity-based measures. Section 5.1 tests whether the procedure recovers evidence against different grades of interventions in the continuous training process. Section 5.2 monitors training processes of paired models under controlled changes. Due to the space limit, we put the comparison among established Qwen2.5-7B siblings in the appendix Appendix C. This experiment is to verify the magnitude and spatial organization of representation shifts.

## 5.1 TRAINING MONITORING UNDER CONTROLLED CHANGES

We evaluate whether the proposed inferential procedure responds systematically to a known intervention in continuation training of a model. We first train a model with the Qwen2.5-0.5B architecture from scratch on TinyStories for $1{,}500$ steps, using 32 sequences of $2{,}048$ tokens per step. We denote the resulting checkpoint by $M^\star$. Starting from $M^\star$, all variants of training scenarios continue under the same setting except for the data they see for $1000$ more steps. The fork from $M^\star$ preserve the architecture and pre-fork parameter weights. For the dose-response branches, the training setup is held fixed except for corpus substitution, so the intervention of interest is the fraction of training sequences drawn from WikiText-103.

The reference branch continues from $M^\star$ on TinyStories only ($\rho = 0$). Six treatment branches replace a fraction $\rho \in \{0.01, 0.05, 0.10, 0.25, 0.50, 1\}$ of training sequences with WikiText-103 under the same batch order. Each treatment is compared with the reference at the same continued training step. In addition, we introduce three controls to distinguish corpus-substitution effects from other variation. **C0** compares the reference with a deterministic rerun from $M^\star$ with all training settings fixed. There should be no difference between the reference and its rerun under **C0**. **C1** compares two identical $\rho = 0$ reruns with each other with nondeterministic GPU kernels to measure execution-induced variation. **C2** compares the reference with a rerun, for $\rho = 0$, differing only in batch order, providing a within-corpus baseline for substitution effects. All runs except the C1 pair use deterministic execution.

At each matched checkpoint, we evaluate all branches (or a pair of them) on the same fixed $n = 200$ sequences: 100 from TinyStories and 100 from WikiText-103, all excluded from initial and continuation training. A pair of models receives identical token inputs and yields one paired representation difference per sequence. Baselines include statistics from hidden-state norms and normalized and unnormalized raw surfaces, as well as other similarity or fixed scalar measures. Output KL compares full-vocabulary next-token distributions on identical prefixes, averaged over positions and sequences. Further implementation details appear in Appendix D. Table 1 divides each final-checkpoint diagnostic by its own C2 value. That is, ratios above one indicate greater discrepancy than that method's batch-order control.

Table 1: Final-checkpoint diagnostics normalized by each method's C2 batch-order control. Full baseline and control definitions are given in Appendix D.

| | WikiText-103 mixture dose $\rho$ (value / C2) | | | | | |
|---|---|---|---|---|---|---|
| Method | 0.01 | 0.05 | 0.10 | 0.25 | 0.50 | 1.00 |
| Metric $g$ | 184 | 201 | 261 | 393 | 454 | 557 |
| Norm $\|X\|$ | 325 | 101 | 90.8 | 106 | 125 | 337 |
| Raw | 20.0 | 15.5 | 14.4 | 14.1 | 14.9 | 28.1 |
| $1-$CKA (Kornblith et al., 2019) | 29.8 | 47.2 | 62.4 | 80.8 | 94.0 | 103 |
| $1-$SVCCA (Raghu et al., 2017) | 2.72 | 4.10 | 4.79 | 5.72 | 6.60 | 12.3 |
| $1-$PWCCA (Morcos et al., 2018) | 3.39 | 5.33 | 6.32 | 7.61 | 8.79 | 16.8 |
| Shape $\theta$ (Williams et al., 2021) | 2.89 | 3.56 | 3.92 | 4.33 | 4.62 | 6.28 |
| MMD$^2$ (Gretton et al., 2012) | 8.77 | 9.29 | 9.87 | 10.6 | 11.6 | 19.8 |
| Energy dist. (Székely & Rizzo, 2013) | 9.20 | 8.19 | 7.96 | 7.93 | 8.44 | 14.7 |
| $\|\Delta W\|$ (all matrices) (Zhong & Raghunathan, 2026) | 3.50 | 5.62 | 6.01 | 6.66 | 7.27 | 8.13 |
| Output KL (TinyStories) | 0.094 | 0.352 | 0.647 | 1.44 | 2.93 | 182 |
| Output KL (WikiText) | 24.6 | 34.9 | 39.5 | 45.6 | 50.1 | 60.2 |

At the smallest nonzero substitution dose, $\rho = 0.01$, the metric-field statistic is 184 times its C2 batch-order-control value. Its normalized response then increases monotonically across all six doses, reaching 557 at $\rho = 1$. The hidden-state norm has a larger normalized value at $\rho = 0.01$ (325 versus 184), but its response is nonmonotone across the remaining doses. The metric field has the largest C2-normalized value among the reported diagnostics at each of the other five doses. Several similarity-based diagnostics also increase monotonically with $\rho$, but with substantially smaller normalized separation from their respective C2 controls.

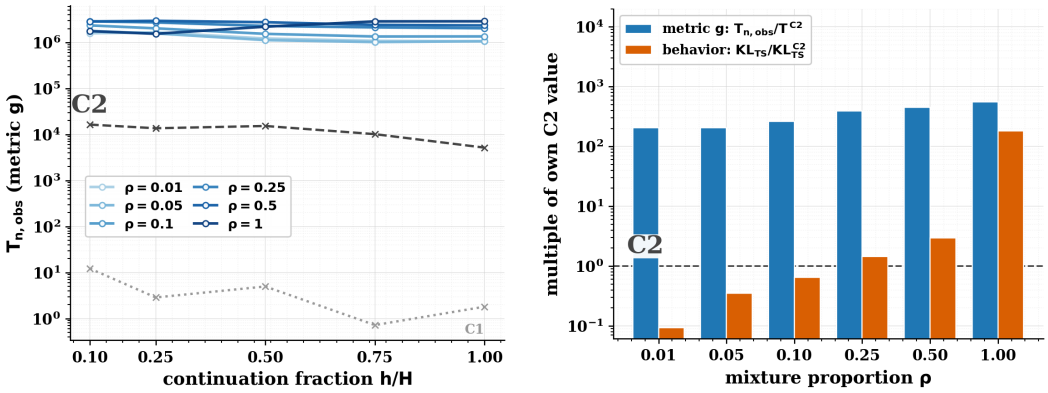


Figure 1: Continuation-training response to controlled corpus substitution. (a) Metric-field statistic $T_{n,\mathrm{obs}}$ for each substitution dose $\rho$, evaluated against the $\rho = 0$ reference at matched checkpoints; C1 and C2 are shown as controls. (b) Final-checkpoint metric-field statistic and TinyStories output KL, each divided by its own C2 batch-order-control value.

Figure 1 examines how the metric-field discrepancy evolves after the common fork checkpoint $M^\star$. Panel (a) tracks $T_{n,\mathrm{obs}}$ for each treatment branch relative to the $\rho = 0$ reference at matched checkpoints, together with the C1 and C2 controls. The temporal trajectories differ by substitution dose: lower-dose branches generally decline over the observed continuation window, whereas the largest-dose branch increases later in training. All intervention branches remain clearly separated from the controls. At the final checkpoint, larger substitution doses consistently produce larger metric-field statistics across the six treatment branches.

Panel (b) compares the metric-field statistic with TinyStories output KL at the final checkpoint after normalizing each diagnostic by its own C2 value. For the three smallest doses, $\rho \in \{0.01, 0.05, 0.10\}$, TinyStories output KL remains below its batch-order-control value, whereas the metric-field statistic exceeds its corresponding C2 value by more than two orders of magnitude.

### 5.2 Tolerance-adjusted monitoring

The zero-mean test in Section 4.2 assesses whether the mean metric-field difference is nonzero. However, statistical significance alone does not indicate whether a shift is large enough to warrant a monitoring flag. In practice, we may wish to tolerate small differences between training runs rather than flag every statistically detectable departure from zero. We therefore supplement the formal test with a tolerance-based screening rule that incorporates a chosen tolerance on the norm of the mean metric-field difference.

We evaluate this rule using eight paired-run scenarios involving Qwen2.5-0.5B models trained from scratch. The scenarios examine differences in parameter initialization, training-data order, data sampling, training corpus, and computational determinism. We use the same $n = 200$ held-out TinyStories validation sequences across scenarios and evaluate both models in each pair at matched training checkpoints. For each scenario and checkpoint, we compare the observed metric-field statistic with the corresponding tolerance-adjusted threshold. Appendix Table 7 specifies the construction of the two runs in each scenario. For scenario $c$ and checkpoint $k$, let $T_{n,\mathrm{obs}}^{c,(k)}$ denote the observed statistic (13). Given a tolerance $\Delta \geq 0$, define

$$S_n^{c,(k)}(\Delta) = T_{n,\mathrm{obs}}^{c,(k)} - n\Delta^2 = n\left\{\|\overline{D}_n^{c,(k)}\|_{\mathcal{H}}^2 - \Delta^2\right\}. \tag{18}$$

Let $\widehat{q}_{1-\alpha}^{c,(k)}$ be the $(1-\alpha)$ quantile of the plug-in spectral reference distribution under the zero-mean null at checkpoint $k$. We flag the checkpoint when

$$S_n^{c,(k)}(\Delta) > \widehat{q}_{1-\alpha}^{c,(k)}. \tag{19}$$

Equivalently, the estimated magnitude of the mean difference field must satisfy

$$\|\overline{D}_n^{c,(k)}\|_{\mathcal{H}} > \left\{ \Delta^2 + \frac{\widehat{q}_{1-\alpha}^{c,(k)}}{n} \right\}^{1/2}.$$

At $\Delta = 0$, this rule is identical to the formal spectral test at level $\alpha$.

We report screening decisions for $\Delta \in \{0, 10, 20, 50\}$ at $\alpha = 0.01$. Here, $\Delta$ is a user-specified tolerance on the $\mathcal{H}$-norm of the mean metric-field difference. Because this scale depends on the chosen representation, a numerical tolerance should not be interpreted as directly comparable across different representations. We therefore restrict the main tolerance analysis to the metric-field representation.

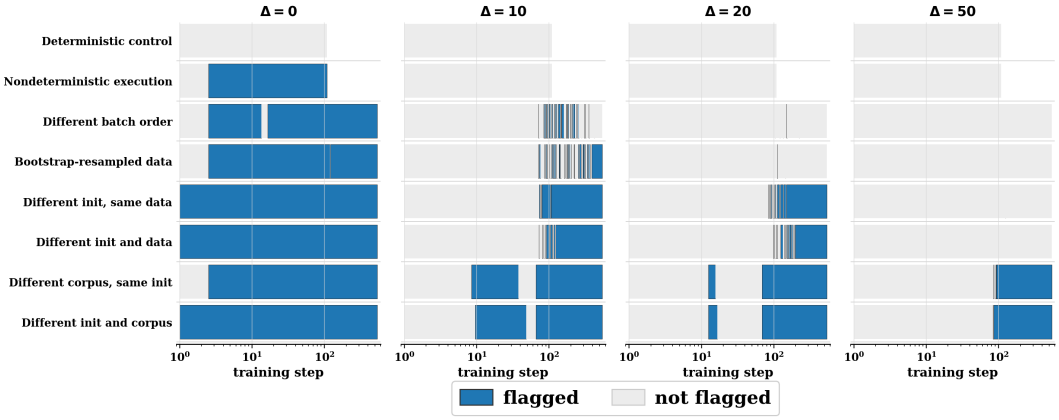


Figure 2: Tolerance-adjusted checkpoint screening for the eight paired-run scenarios in Table 7. Columns correspond to $\Delta \in \{0, 10, 20, 50\}$ and rows correspond to the eight scenarios. Blue denotes a flagged checkpoint and gray denotes an evaluated checkpoint that is not flagged. Decisions are pointwise over checkpoints at $\alpha = 0.01$. The deterministic and nondeterministic execution scenarios are evaluated for 100 training steps, whereas the remaining scenarios are evaluated for 500 steps.

At $\Delta = 0$ in Figure 2, the zero-mean spectral test flags most evaluated checkpoints in the non-control scenarios, whereas the deterministic control remains unflagged. This indicates that the procedure distinguishes the control and any departure from it. As $\Delta$ increases, flagged intervals generally begin later and become shorter or more intermittent because smaller observed departures no longer exceed the tolerance-adjusted threshold. At $\Delta = 10$ and 20, persistent flags remain in several initialization and corpus comparisons. At $\Delta = 50$, sustained flags are concentrated primarily in scenarios involving a change of training corpus. Thus, increasing the tolerance progressively filters smaller detectable differences while retaining the largest and most persistent representation shifts in these experiments.

## 6 Conclusion

We introduced a population-level inferential framework for detecting systematic shifts in transformer representations. For each sampled input, hidden states are represented as a surface over normalized layer depth and token position and summarized through its induced metric field. Paired metric-field differences then provide functional observations whose population mean defines the representation shift of interest. In controlled continuation-training experiments, the metric-field statistic responds monotonically to increasing corpus substitution at the final checkpoint and separates small substitutions from the batch-order control. The tolerance-adjusted screening analysis further shows how increasing a user-specified tolerance suppresses smaller detectable differences while retaining larger and more persistent shifts. Comparisons among established model siblings provide complementary evidence that these differences have structured organization over layer depth and token position.

### AI USE STATEMENT

Generative AI tools were used to assist with manuscript organization, language editing, and feedback on the presentation of the methodology, experiments, theoretical arguments, and proofs. They were also used to assist with checking bibliographic information and related literature. Generative AI tools were not used to generate experimental data, conduct model training, or determine the numerical results reported in the paper. The authors independently reviewed and verified all AI-assisted revisions, mathematical arguments, and empirical results and take responsibility for the final content of the submission.

## REFERENCES


Georgios Arvanitidis, Lars Kai Hansen, and Søren Hauberg. Latent space oddity: on the curvature of deep generative models. In *International Conference on Learning Representations*, 2018.

Denis Bosq. *Linear Processes in Function Spaces: Theory and Applications*. Springer, 2000.

Irene Cannistraci, Luca Moschella, Marco Fumero, Valentino Maiorca, and Emanuele Rodolà. From bricks to bridges: Product of invariances to enhance latent space communication. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/1fb591d5a6b1da9bd0d0648d122fe2a1-Abstract-Conference.html.

Nadir Durrani, Hassan Sajjad, Fahim Dalvi, and Firoj Alam. On the transformation of latent space in fine-tuned NLP models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 1495–1516. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.emnlp-main.97.

Kawin Ethayarajh. How contextual are contextualized word representations? comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 55–65. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1006.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012.

Lajos Horváth, Piotr Kokoszka, and Ron Reeder. Estimation of the mean of functional time series and a two-sample problem. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(1):103–122, 2013. doi: 10.1111/j.1467-9868.2012.01032.x.

Tailen Hsing and Randall Eubank. *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators*. Wiley Series in Probability and Statistics. Wiley, 2015.

Francesco Maria Gabriele Insulla, Shuo Huang, and Lorenzo Rosasco. Towards a learning theory of representation alignment. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/5d4f5a2de6320641566be8722d5f78dc-Abstract-Conference.html.

Fran Jelenić, Josip Jukić, Martin Tutek, Mate Puljiz, and Jan Snajder. Out-of-distribution detection by leveraging between-layer transformation smoothness. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/0a2f65c9d2313b71005e600bd23393fe-Abstract-Conference.html.

Jiachen Jiang, Jinxin Zhou, and Zhihui Zhu. Tracing representation progression: Analyzing and enhancing layer-wise similarity. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/03d113a060c0ac93a5859517a0f07271-Abstract-Conference.html.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3519–3529. PMLR, 2019.

Ilya Loshchilov, Cheng-Ping Hsieh, Simeng Sun, and Boris Ginsburg. nGPT: Normalized transformer with representation learning on the hypersphere. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/b7c12689a89e98a61bcaa65285a41b7c-Abstract-Conference.html.

Achleshwar Luthra, Priyadarsi Mishra, and Tomer Galanti. On the alignment between supervised and self-supervised contrastive learning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://iclr.cc/virtual/2026/poster/10010174.

Amil Merchant, Elahe Rahimtoroghi, Ellie Pavlick, and Ian Tenney. What happens to BERT embeddings during fine-tuning? In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pp. 33–44. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.blackboxnlp-1.4.

Ari S. Morcos, Maithra Raghu, and Samy Bengio. Insights on representational similarity in neural networks with canonical correlation. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Philipp Nazari, Sebastian Damrich, and Fred A. Hamprecht. Geometric autoencoders—what you see is what you decode. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 25834–25857. PMLR, 2023.

Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. LLMs know more than they show: On the intrinsic representation of LLM hallucinations. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/a712d461e57201efe35d429a6f1731c1-Abstract-Conference.html.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. SVCCA: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Advances in Neural Information Processing Systems*, volume 30, pp. 6076–6085, 2017.

James O. Ramsay and Bernard W. Silverman. *Functional Data Analysis*. Springer, 2 edition, 2005.

Emily Reif, Andy Coenen, Ann Yuan, Been Kim, Adam Pearce, Fernanda B. Viégas, and Martin Wattenberg. Visualizing and measuring the geometry of BERT. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Gábor J. Székely and Maria L. Rizzo. Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference*, 143(8):1249–1272, 2013. doi: 10.1016/j.jspi.2013.03.018.

Alex H. Williams, Erin Kunz, Simon Kornblith, and Scott W. Linderman. Generalized shape metrics on neural representations. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Ziqian Zhong and Aditi Raghunathan. Watch the weights: Unsupervised monitoring and control of fine-tuned LLMs. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://iclr.cc/virtual/2026/poster/10009043.

# A Proofs

## A.1 Proof of Theorem 1

*Proof.* Fix a contrast $c$, and suppress the superscript $c$ by writing

$$D_i \equiv D_i^c, \qquad \mu \equiv \mu^c, \qquad \mathcal{K} \equiv \mathcal{K}^c.$$

By Assumption 1, $D_1, \dots, D_n$ are independent and identically distributed random elements of the separable Hilbert space $\mathcal{H}$, with

$$\mathbb{E}\|D_1\|_{\mathcal{H}}^2 < \infty.$$

In particular, $\mathbb{E}\|D_1\|_{\mathcal{H}} < \infty$, so the Bochner mean $\mu = \mathbb{E}D_1$ is well defined.

The strong law of large numbers for Hilbert-space-valued random elements therefore gives

$$\left\|\bar{D}_n - \mu\right\|_{\mathcal{H}} \xrightarrow{\text{a.s.}} 0, \qquad \bar{D}_n = \frac{1}{n}\sum_{i=1}^{n} D_i.$$

Consequently,

$$\bar{D}_n \xrightarrow{p} \mu \qquad \text{in } \mathcal{H},$$

which proves (10).

For the central limit theorem, define the centered random elements

$$Y_i = D_i - \mu.$$

Then $Y_1, \ldots, Y_n$ are independent and identically distributed,

$$\mathbb{E}Y_i = 0_{\mathcal{H}}, \qquad \mathbb{E}\|Y_i\|_{\mathcal{H}}^2 < \infty.$$

Their covariance operator is $\mathcal{K}$ in (8), or equivalently, for every $f, h \in \mathcal{H}$,

$$\mathbb{E}\left[\langle Y_i, f\rangle_{\mathcal{H}} \langle Y_i, h\rangle_{\mathcal{H}}\right] = \langle \mathcal{K}f, h\rangle_{\mathcal{H}}.$$

The central limit theorem for independent and identically distributed random elements in a separable Hilbert space (Bosq, 2000; Hsing & Eubank, 2015) now yields

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} Y_i \xrightarrow{d} \mathbb{G}^c \qquad \text{in } \mathcal{H},$$

where $\mathbb{G}^c$ is a centered Gaussian random element with covariance operator $\mathcal{K}^c$. Since

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} Y_i = \sqrt{n}\left(\bar{D}_n^c - \mu^c\right),$$

this establishes (11).

If Assumption 2 also holds, write

$$\overline{\widehat{D}}_n^c = \frac{1}{n}\sum_{i=1}^{n} \widehat{D}_i^c.$$

Then

$$\sqrt{n}\left(\overline{\widehat{D}}_n^c - \mu^c\right) = \sqrt{n}\left(\bar{D}_n^c - \mu^c\right) + \sqrt{n}\left(\overline{\widehat{D}}_n^c - \bar{D}_n^c\right),$$

where the second term is $o_p(1)$ in $\mathcal{H}$. Slutsky's theorem therefore gives the same Gaussian limit for $\overline{\widehat{D}}_n^c$. Moreover, $\|\overline{\widehat{D}}_n^c - \bar{D}_n^c\|_{\mathcal{H}} = o_p(1)$, so consistency also follows.

□

### A.2 Proof of Corollary 1

*Proof.* Under $H_0^c$ in (12), $\mu^c = 0_{\mathcal{H}}$, so Theorem 1 gives

$$\sqrt{n}\,\bar{D}_n^c \xrightarrow{d} \mathbb{G}^c \qquad \text{in } \mathcal{H},$$

which is (11) under the null. The map $f \mapsto \|f\|_{\mathcal{H}}^2$ is continuous on $\mathcal{H}$, hence the continuous mapping theorem applied to the statistic (13) yields

$$T_n^c = \|\sqrt{n}\,\bar{D}_n^c\|_{\mathcal{H}}^2 \xrightarrow{d} \|\mathbb{G}^c\|_{\mathcal{H}}^2.$$

Because $\mathcal{K}^c$ in (8) is compact, self-adjoint, positive semidefinite, and trace class, the Karhunen–Loève expansion of $\mathbb{G}^c$ in the eigenbasis $(\phi_r^c)_{r\geq 1}$ gives

$$\mathbb{G}^c = \sum_{r=1}^{\infty} \sqrt{\lambda_r^c}\,\zeta_r\,\phi_r^c, \qquad \zeta_r \overset{\text{iid}}{\sim} N(0,1),$$

so Parseval's identity implies

$$\|\mathbb{G}^c\|_{\mathcal{H}}^2 = \sum_{r=1}^{\infty} \lambda_r^c\,\zeta_r^2,$$

which is the weighted chi-square mixture in (14). □

### A.3 Proof of Proposition 1

*Proof.* Let $F$ denote the continuous distribution function of the limiting null law (14), and let $\widehat{F}_n$ denote the plug-in spectral distribution function based on $\widehat{\mathcal{K}}_n^c$, constructed from (15) and (16). The consistency condition in Proposition 1 means that

$$\sup_{x\in\mathbb{R}}\left|\widehat{F}_n(x) - F(x)\right| \xrightarrow{p} 0,$$

where the truncation rule leaves an asymptotically negligible spectral tail. Equivalently, the estimated $(1-\alpha)$ quantile converges in probability to the corresponding quantile of $F$. Because the Monte Carlo $p$-value $\widehat{p}_c$ in (17) estimates a right-tail probability under $\widehat{F}_n$, standard arguments for plug-in calibration (Hsing & Eubank, 2015) yield

$$\mathbb{P}_{H_0^c}(\widehat{p}_c \le \alpha) \to \alpha.$$

For power, fix $\mu^c \neq 0_{\mathcal{H}}$. Theorem 1 implies (10), hence the statistic (13) satisfies

$$\frac{T_n^c}{n} = \|\bar{D}_n^c\|_{\mathcal{H}}^2 \xrightarrow{p} \|\mu^c\|_{\mathcal{H}}^2 > 0.$$

Hence $T_n^c \to \infty$ in probability. Therefore the observed statistic eventually lies in the far right tail of any fixed plug-in null approximation, so

$$\mathbb{P}_{H_1^c}(\widehat{p}_c \le \alpha) \to 1.$$

□

## B Finite-grid and testing implementation

The following describe the general testing procedure given a general setup. Each task instance $i$ has a context length $T_i$, which is either fixed at a common nominal value $T_0$ or allowed to vary across instances. We partition the token positions into $K_t$ bins $\{\mathcal{B}_{i,q}\}_{q=1}^{K_t}$ and average the hidden states within each bin. The resulting value $X_{i,m,pq}$ is indexed by model $m$, layer $p \in \{1,\dots,K_\ell\}$, and token bin $q$, and provides a finite-grid approximation to the continuous surface $X_{i,m}(u)$. From this surface, we estimate the metric field $\widehat{g}_{i,m}$ as in (1) and construct the paired contrast fields $\widehat{D}_1^c,\dots,\widehat{D}_n^c$ as in (4). More details are given in Appendix B.

For each contrast $c$, inference uses the sample mean field $\bar{\widehat{D}}_n^c$ and empirical covariance kernel $\hat{\Sigma}_n^c$. A spectral decomposition of the associated empirical covariance operator yields the leading eigenpairs $(\hat{\lambda}_j^c, \hat{\phi}_j^c)$, $j = 1,\dots,R$, with $R$ selected by the prespecified truncation rule. We approximate the null distribution of $T_n^c$ in (13) by Monte Carlo draws from the truncated mixture (16). For each $b = 1,\dots,B_{\text{null}}$, we generate

$$T_b^{c,\star} = \sum_{j=1}^{R} \hat{\lambda}_j^c Z_{jb}^2, \qquad Z_{jb} \overset{\text{iid}}{\sim} N(0,1).$$

We compare the observed statistic $T_{n,\text{obs}}^c$ with these simulated null values and compute the $p$-value (17) using the finite Monte Carlo correction described in Appendix B.

Before running the models, we define token bins for each instance. For positions $r = 1,\dots,T_i$,

$$\mathcal{B}_{i,q} = \left\{ r : \left\lfloor \frac{(q-1)T_i}{K_t} \right\rfloor + 1 \le r \le \left\lfloor \frac{qT_i}{K_t} \right\rfloor \right\}, \qquad q = 1,\dots,K_t.$$

Because $T_i/K_t$ need not be an integer, bin cardinalities may differ. In the established-model experiment, where $T_i \equiv T_0 = 30{,}000$ and $K_t = 64$, $|\mathcal{B}_{i,q}| \in \{468, 469\}$. The layer coordinates $\ell_p = (p-1)/(K_\ell - 1)$ are uniformly spaced and do not depend on $i$. The token coordinate for bin $q$ is its mean normalized relative position,

$$t_{i,q} = \frac{1}{|\mathcal{B}_{i,q}|} \sum_{r\in\mathcal{B}_{i,q}} \frac{r-1}{T_i - 1}, \qquad q = 1,\dots,K_t.$$

Thus, $t_{i,q} \in [0,1]$ provides a context-position coordinate that is comparable across instances of different lengths. The evaluation grid for instance $i$ is

$$\mathcal{G}_i = \{u_{i,pq} = (\ell_p, t_{i,q}) : p = 1, \ldots, K_\ell,\ q = 1, \ldots, K_t\} \subset \mathcal{U},$$

and its centered-difference interior is $\mathcal{G}_i^\circ = \{u_{i,pq} : p = 2, \ldots, K_\ell - 1,\ q = 2, \ldots, K_t - 1\}$. When $T_i \equiv T_0$, all instances share the grids $\mathcal{G}_i = \mathcal{G}$ and $\mathcal{G}_i^\circ = \mathcal{G}^\circ$. When the $T_i$ are heterogeneous, the nominal-length coordinates $t_q = t_{T_0,q}$ define a common reference grid $\mathcal{G}$. Before computing sample means or covariances, we linearly interpolate each field component from $\mathcal{G}_i^\circ$ onto $\mathcal{G}^\circ$ along the token axis. Consequently, all inferential operations use registered values on a fixed grid. We then sample $Z_1, \ldots, Z_n$ from $\mathcal{P}_{a,T_0}$, construct the sequences $S_i = \kappa_{a,T_i}(Z_i)$, and evaluate all three models $m \in \{B, I, C\}$ on every sequence.

The second step constructs bin-averaged hidden-state surfaces. Let $X_{i,m,p,r} \in \mathbb{R}^{d_{\mathrm{hid},m}}$ denote the hidden state at layer $p$ and token position $r$ after a forward pass on $S_i$. The corresponding bin average is

$$X_{i,m,pq} = \frac{1}{|\mathcal{B}_{i,q}|} \sum_{r \in \mathcal{B}_{i,q}} X_{i,m,p,r}, \qquad p = 1, \ldots, K_\ell,\ q = 1, \ldots, K_t.$$

The third step normalizes each bin-averaged hidden-state vector to unit norm:

$$\widetilde{X}_{i,m,pq} = \frac{X_{i,m,pq}}{\|X_{i,m,pq}\|}, \qquad m \in \{B, I, C\},$$

The normalized vectors are used to construct $\widehat{g}_{i,m}(u)$, $\widehat{D}_i^c(u)$, and $T_n^c$. Unit normalization is a preprocessing choice rather than a requirement of the metric-field framework. It removes the magnitude of each bin-averaged hidden-state vector while retaining its direction. Accordingly, the metric construction in (1) is applied to $\widetilde{X}_{i,m}$ rather than the unnormalized surface $X_{i,m}$. The resulting field describes local directional variation across layer depth and token position and the coupling between these variations.

The fourth step estimates derivatives and induced metric fields from the normalized surfaces. The confirmatory analysis excludes boundary points and uses only centered finite differences on $\mathcal{G}_i^\circ$, and it does not use one-sided differences. For $u_{i,pq} \in \mathcal{G}_i^\circ$, the partial derivatives are estimated using the actual coordinate spacings:

$$\widehat{\partial_\ell X}_{i,m}(u_{i,pq}) = \frac{\widetilde{X}_{i,m,p+1,q} - \widetilde{X}_{i,m,p-1,q}}{\ell_{p+1} - \ell_{p-1}},$$

and

$$\widehat{\partial_t X}_{i,m}(u_{i,pq}) = \frac{\widetilde{X}_{i,m,p,q+1} - \widetilde{X}_{i,m,p,q-1}}{t_{i,q+1} - t_{i,q-1}}.$$

The estimated pullback metric field for model $m$ is the symmetric $2 \times 2$ matrix (1),

$$\widehat{g}_{i,m}(u) = \widehat{J}_{i,m}(u)^\top \widehat{J}_{i,m}(u),$$

where

$$\widehat{J}_{i,m}(u) = \left[\widehat{\partial_\ell X}_{i,m}(u) \quad \widehat{\partial_t X}_{i,m}(u)\right] \in \mathbb{R}^{d_{\mathrm{hid},m} \times 2}.$$

Equivalently,

$$\widehat{g}_{i,m}(u) = \begin{pmatrix} \widehat{g}_{i,m,\ell\ell}(u) & \widehat{g}_{i,m,\ell t}(u) \\ \widehat{g}_{i,m,\ell t}(u) & \widehat{g}_{i,m,tt}(u) \end{pmatrix},$$

with entries

$$\widehat{g}_{i,m,\ell\ell}(u) = \left\|\widehat{\partial_\ell X}_{i,m}(u)\right\|^2,$$

$$\widehat{g}_{i,m,tt}(u) = \left\|\widehat{\partial_t X}_{i,m}(u)\right\|^2,$$

and

$$\widehat{g}_{i,m,\ell t}(u) = \left\langle \widehat{\partial_\ell X}_{i,m}(u), \widehat{\partial_t X}_{i,m}(u) \right\rangle.$$

The reference-based contrast fields are then

$$\widehat{D}_i^{I-B}(u) = \operatorname{vech}\left\{\widehat{g}_{i,I}(u) - \widehat{g}_{i,B}(u)\right\}, \qquad \widehat{D}_i^{C-B}(u) = \operatorname{vech}\left\{\widehat{g}_{i,C}(u) - \widehat{g}_{i,B}(u)\right\},$$

Both fields take values in $\mathbb{R}^3$. The coder-tuned versus instruction-tuned contrast requires no additional model evaluation:

$$\widehat{D}_i^{C-I}(u) = \widehat{D}_i^{C-B}(u) - \widehat{D}_i^{I-B}(u).$$

Thus, the labels have the orientations $\mathrm{BI} \equiv I - B$, $\mathrm{BC} \equiv C - B$, and $\mathrm{IC} \equiv C - I$. For each $c \in \{\mathrm{BI}, \mathrm{BC}, \mathrm{IC}\}$, confirmatory inference uses $\widehat{D}_i^c : \mathcal{G}^\circ \to \mathbb{R}^3$. The observations for a fixed contrast are

$$\widehat{D}_1^c, \ldots, \widehat{D}_n^c.$$

In terms of checkpoint evaluation between a pair of iterations, the same construction applies with the two members of a contrast replaced by two checkpoints at a matched training step. Let $m_+^{(k)}$ and $m_-^{(k)}$ denote the pair at iteration $k$ (a dose branch versus the $\rho = 0$ continuation, or the two from-scratch runs in a scenario). Each checkpoint is evaluated on the identical sequences $S_1, \ldots, S_n$. The metric fields $\widehat{g}_{i,m_+^{(k)}}$ and $\widehat{g}_{i,m_-^{(k)}}$ are formed as above, and the per-prompt difference is

$$\widehat{D}_i^{(k)}(u) = \operatorname{vech}\left\{\widehat{g}_{i,m_+^{(k)}}(u) - \widehat{g}_{i,m_-^{(k)}}(u)\right\}.$$

The sample $\widehat{D}_1^{(k)}, \ldots, \widehat{D}_n^{(k)}$ is the input to the spectral test at that checkpoint.

### B.1 Testing

On the common interior grid, let $J^\circ = |\mathcal{G}^\circ| = (K_\ell - 2)(K_t - 2)$ and assign equal-cell quadrature weights $\omega_{pq} = 1/J^\circ$. For three-component metric fields, the quadrature form of the inner product (6) is

$$\langle f, h\rangle_{\mathcal{H}} = \frac{1}{3} \sum_{u_{pq} \in \mathcal{G}^\circ} \omega_{pq} f(u_{pq})^\top h(u_{pq}), \qquad \|f\|_{\mathcal{H}}^2 = \langle f, f\rangle_{\mathcal{H}}.$$

The factor $1/3$ is omitted for scalar representations. These weights implement the spatial average on the normalized registered grid and make the finite-grid statistic an explicit quadrature approximation to its continuous counterpart.

For a fixed contrast $c$, we compute the sample mean

$$\bar{\widehat{D}}_n^c(u) = \frac{1}{n} \sum_{i=1}^n \widehat{D}_i^c(u),$$

and sample covariance function

$$\hat{\Sigma}_n^c(u, v) = \frac{1}{n-1} \sum_{i=1}^n \left\{\widehat{D}_i^c(u) - \bar{\widehat{D}}_n^c(u)\right\}\left\{\widehat{D}_i^c(v) - \bar{\widehat{D}}_n^c(v)\right\}^\top,$$

We then obtain the empirical eigenvalues $\hat{\lambda}_r$ and functional principal components $\hat{\phi}_r(u)$ for $r = 1, \ldots, R$. The prespecified rule selects the smallest $R$ for which the cumulative empirical percentage of variance explained is at least $95\%$.

### B.2 Power analysis

To evaluate power against structured alternatives, we separate the spatial pattern of a mean shift from its direction in metric-component space. Define the scalar spatial pattern

$$\rho(u) = \sin(2\pi\ell) + \sin(2\pi t), \qquad u = (\ell, t) \in \mathcal{G}^\circ,$$

the global all-component direction

$$v_{\mathrm{all}} = \frac{1}{\sqrt{3}} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix},$$

and the three component-specific directions $v_{\ell\ell}$, $v_{\ell t}$, $v_{tt}$. For a selected support $\mathcal{A} \subseteq \mathcal{G}^\circ$ and $v \in \{v_{\text{all}}, v_{\ell\ell}, v_{\ell t}, v_{tt}\}$, let $w_{\mathcal{A},v}(u) = \rho(u)\,\mathbb{I}(u \in \mathcal{A})\,v$ and define

$$h_{\mathcal{A},v}(u) = \frac{w_{\mathcal{A},v}(u)}{\|w_{\mathcal{A},v}\|_{\mathcal{H}}}, \qquad u \in \mathcal{G}^\circ.$$

Using the previous mentioned $\widehat{\Sigma}_y^c$, the estimated covariance, we define

$$s_0 = \sqrt{\operatorname{tr}(\Sigma_y)},$$

and set

$$\mu_{\delta,\mathcal{A},v}(u) = \delta\, s_0\, h_{\mathcal{A},v}(u), \qquad u \in \mathcal{G}^\circ. \tag{20}$$

We vary $\delta$ from 0 to 0.3. Under this parameterization, $\delta$ is the mean-field magnitude relative to the total field-level standard deviation because $\|\mu_{\delta,\mathcal{A},v}\|_{\mathcal{H}} = \delta s_0$. The three power-simulation alternatives are

$$H_{1a}:\ \mu^c(u) = \mu^c_{\delta,\mathcal{G}^\circ,v_{\text{all}}}(u) \qquad \text{(global: } \mathcal{A} = \mathcal{G}^\circ\text{)}, \tag{21}$$

$$H_{1b}:\ \mu^c(u) = \mu_{\delta,\mathcal{A},v}(u) \qquad \text{(local: } \mathcal{A} \subsetneq \mathcal{G}^\circ\text{)}, \tag{22}$$

$$H_{1c}:\ \mu^c(u) = \mu_{\delta,\mathcal{A},v}(u), \quad v \in \{v_{\ell\ell}, v_{\ell t}, v_{tt}\} \qquad \text{(component-specific)}. \tag{23}$$

The global alternative (21) shifts the full interior grid. The local alternative (22) restricts the shift to a proper subset $\mathcal{A} \subsetneq \mathcal{G}^\circ$ spanning selected depth or token positions. The component-specific alternative (23) holds the spatial support fixed and selects one $\mathrm{vech}$ direction.

Table 2 lists the factors varied in the power study.

Table 2: Power-simulation factors.

| Factor | Settings |
|---|---|
| Standardized effect size $\delta$ | swept over $\{0,\ 0.02,\ 0.04,\ 0.05,\ 0.06,\ 0.08,\ 0.10,\ 0.15,\ 0.20,\ 0.30\}$ |
| Sample size $n$ | 25, 50, 100, 200 |
| Alternative shape $\rho(u)$ | $\sin(2\pi\ell) + \sin(2\pi t)$ |
| Support size $\lvert\mathcal{A}\rvert/\lvert\mathcal{G}^\circ\rvert$ | 25% of interior grid cells (for $H_{1b}$ only) |
| Component direction $v$ | $v_{\ell\ell}$, $v_{\ell t}$, $v_{tt}$ (for $H_{1c}$ only) |
| Significance level $\alpha$ | 0.05 |
| Outer power replicates $B_{\text{pow}}$ | $B_{\text{pow}} = 1000$ (each with $B_{\text{null}} = 2000$ inner null draws) |

The local power alternatives are as follows:

Table 3: Local patches for $H_{1b}$.

| Patch | Support $\mathcal{A}$ (primary: $K_\ell = 29$, $K_t = 64$) |
|---|---|
| Early depth–early tokens | $\mathcal{A}_{\text{ee}} = \{u_{pq} : 2 \le p \le 15,\ 2 \le q \le 32\}$ |
| Late depth–late tokens | $\mathcal{A}_{\text{ll}} = \{u_{pq} : 15 \le p \le 28,\ 33 \le q \le 63\}$ |
| Early depth–late tokens | $\mathcal{A}_{\text{el}} = \{u_{pq} : 2 \le p \le 15,\ 33 \le q \le 63\}$ |

Table 4 summarizes the alternative settings used in the power analysis.

Table 4: Alternative-hypothesis settings for the power analysis.

| Alternative | Setting |
|---|---|
| $H_{1a}$ | Global shift on $\mathcal{G}^\circ$: $\mu^c(u) = \mu^c_{\delta,\mathcal{G}^\circ,v_{\mathrm{all}}}(u)$ |
| $H_{1b}$ | Local shift on a registered corner patch $\mathcal{A} \subsetneq \mathcal{G}^\circ$ (Table 3): $\mu^c(u) = \mu_{\delta,\mathcal{A},v}(u)$ with $v = v_{\ell\ell}$ |
| $H_{1c}$ | Component-specific shift: $\mu^c(u) = \mu_{\delta,\mathcal{A},v}(u)$ with $v \in \{v_{\ell\ell}, v_{\ell t}, v_{tt}\}$ |

We estimate power and empirical size using a residual bootstrap. From the difference fields $\widehat{D}^c_1, \ldots, \widehat{D}^c_n$, we form the centered residuals

$$\widehat{R}_j(u) = \widehat{D}^c_j(u) - \bar{D}^c_n(u), \qquad j = 1, \ldots, n.$$

For each outer replicate $b = 1, \ldots, B_{\mathrm{pow}}$, draw indices $i^{(b)}_1, \ldots, i^{(b)}_n$ uniformly from $\{1, \ldots, n\}$ with replacement and construct

$$\widehat{D}^{c,(b)}_i(u) = \mu^c(u) + \widehat{R}_{i^{(b)}_i}(u), \qquad i = 1, \ldots, n,$$

For size evaluation, $\mu^c$ is set to zero, and under the alternatives in Table 4, it is set to $\mu_{\delta,\mathcal{A},v}$. For each replicate, we repeat the full confirmatory procedure, including covariance estimation, FPCA truncation, plug-in spectral calibration, and $p$-value computation. We record a rejection when $\hat{p}_b < \alpha$ and estimate power by

$$\widehat{\mathrm{Power}} = \frac{1}{B_{\mathrm{pow}}} \sum_{b=1}^{B_{\mathrm{pow}}} \mathbb{I}(\hat{p}_b < \alpha),$$

with Monte Carlo standard error

$$\widehat{\mathrm{se}}(\widehat{\mathrm{Power}}) = \sqrt{\widehat{\mathrm{Power}}(1 - \widehat{\mathrm{Power}})/B_{\mathrm{pow}}}.$$

Each outer replicate uses $B_{\mathrm{null}} = 2000$ inner null draws for the plug-in $p$-value (Table 2).

Figure 3 reports residual-bootstrap power for the established Qwen2.5-7B contrasts at $n = 200$ and $T_0 = 30{,}000$, with columns indexing the three local patches in Table 3. Figure 3 shows that the empirical rejection rates remain close to the nominal level $\alpha = 0.05$ at $\delta = 0$ and increase monotonically with the standardized effect size. For the metric-field representation, the power curves are nearly identical across the $I$–$B$, $C$–$B$, and $C$–$I$ contrasts and across the global and local alternatives. The normalized raw surface shows a similarly rapid increase in power (the unnormalized raw surface is qualitatively similar), although performance at intermediate effect sizes varies more across contrasts. By comparison, the hidden-state norm is consistently less sensitive: its power increases more gradually and approaches one only for larger effects near $\delta = 0.08$–$0.10$.

## C Established-model comparison

We compare three Qwen2.5-7B siblings: the base model ($B$), the instruction-tuned model ($I$), and the coder-tuned model ($C$). For each task instance $Z_i$, $i = 1, \ldots, n$, we construct one token sequence $S_i = \kappa_{a,T_i}(Z_i) \in \{1, \ldots, |\mathcal{V}|\}^{T_i}$ using the shared vocabulary of size $|\mathcal{V}| = 152{,}064$. The identical sequence is then provided to all three models. Throughout the experiments, each compact contrast label lists the reference model first and the comparison model second: $\mathrm{BI} \equiv I - B$, $\mathrm{BC} \equiv C - B$, and $\mathrm{IC} \equiv C - I$. Hidden-state binning and the spectral test follow Section 5 and Appendix B.

Table 5 reports the spectral-test results for the metric-field representation at $n \in \{200, 500\}$ under fixed and heterogeneous context lengths. Across all three designs, the test rejects the null hypothesis for every model contrast at the minimum attainable Monte Carlo $p$-value, $1/(B_{\mathrm{null}}+1) = 1/10{,}001 < 10^{-4}$. The contrasts involving the coder-tuned model, $C - B$ and $C - I$, consistently have larger mean-field norms than the $I - B$ contrast. Their estimated population mean differences are therefore

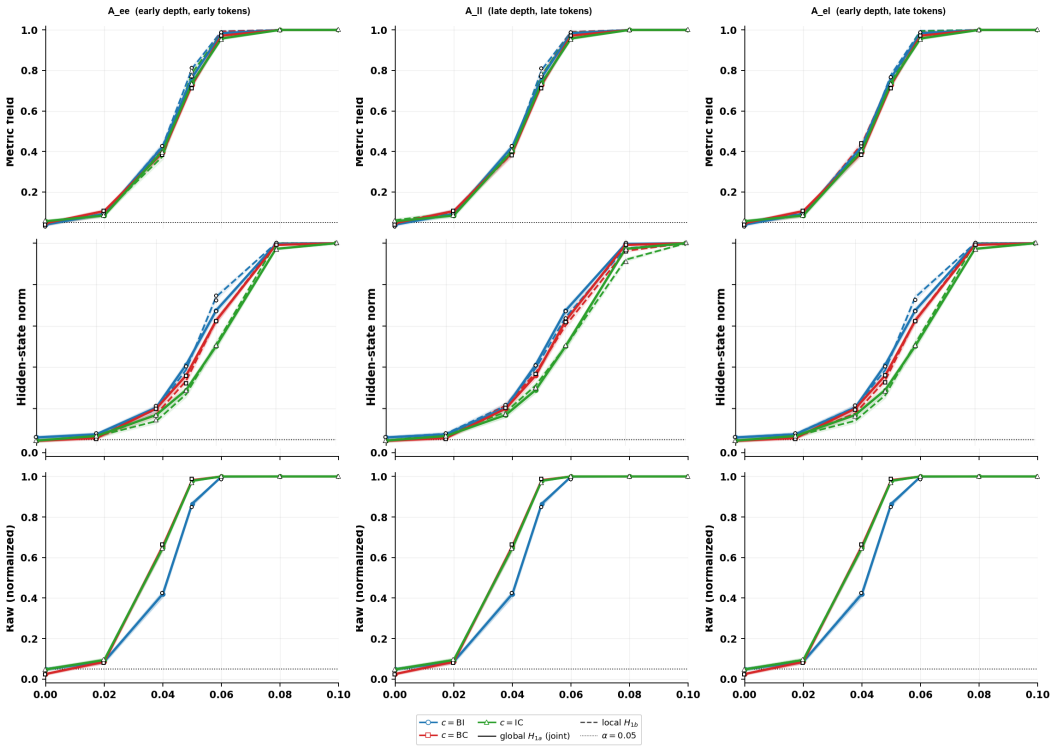


Figure 3: Residual-bootstrap power ($n = 200$, $T_0 = 30{,}000$, $\delta \leq 0.1$). Rows are representations (metric field, hidden-state norm, normalized raw surface), and columns are local patches $\mathcal{A}_{\mathrm{ee}}$, $\mathcal{A}_{\mathrm{ll}}$, and $\mathcal{A}_{\mathrm{el}}$ (Table 3) with $|\mathcal{A}| = 25\%$ and loading $v_{\ell\ell}$. Solid lines denote the global joint alternative $H_{1a}$, and dashed lines denote the column's local patch $H_{1b}$. Contrasts are distinguished in both color and marker shape: BI circles, BC squares, and IC triangles.

Table 5: Spectral tests on the metric-field representation for established Qwen2.5-7B siblings ($B$, $I$, $C$).

| | BI $\equiv I - B$ | | BC $\equiv C - B$ | | IC $\equiv C - I$ | |
|---|---|---|---|---|---|---|
| Study | $T_{\text{obs}}$ | $\|\bar{D}_n\|_{\mathcal{H}}$ | $T_{\text{obs}}$ | $\|\bar{D}_n\|_{\mathcal{H}}$ | $T_{\text{obs}}$ | $\|\bar{D}_n\|_{\mathcal{H}}$ |
| $n = 200$, $T_0 = 30{,}000$ | $5.92 \times 10^4$ | 17.211 | $8.68 \times 10^5$ | 65.881 | $7.77 \times 10^5$ | 62.326 |
| $n = 500$, $T_0 = 50{,}000$ | $7.60 \times 10^4$ | 12.331 | $1.41 \times 10^6$ | 53.047 | $1.30 \times 10^6$ | 50.934 |
| Varying length, $n = 500$, $T_i \sim \mathrm{U}[49k, 51k]$ | $7.62 \times 10^4$ | 12.342 | $1.41 \times 10^6$ | 53.072 | $1.30 \times 10^6$ | 50.951 |

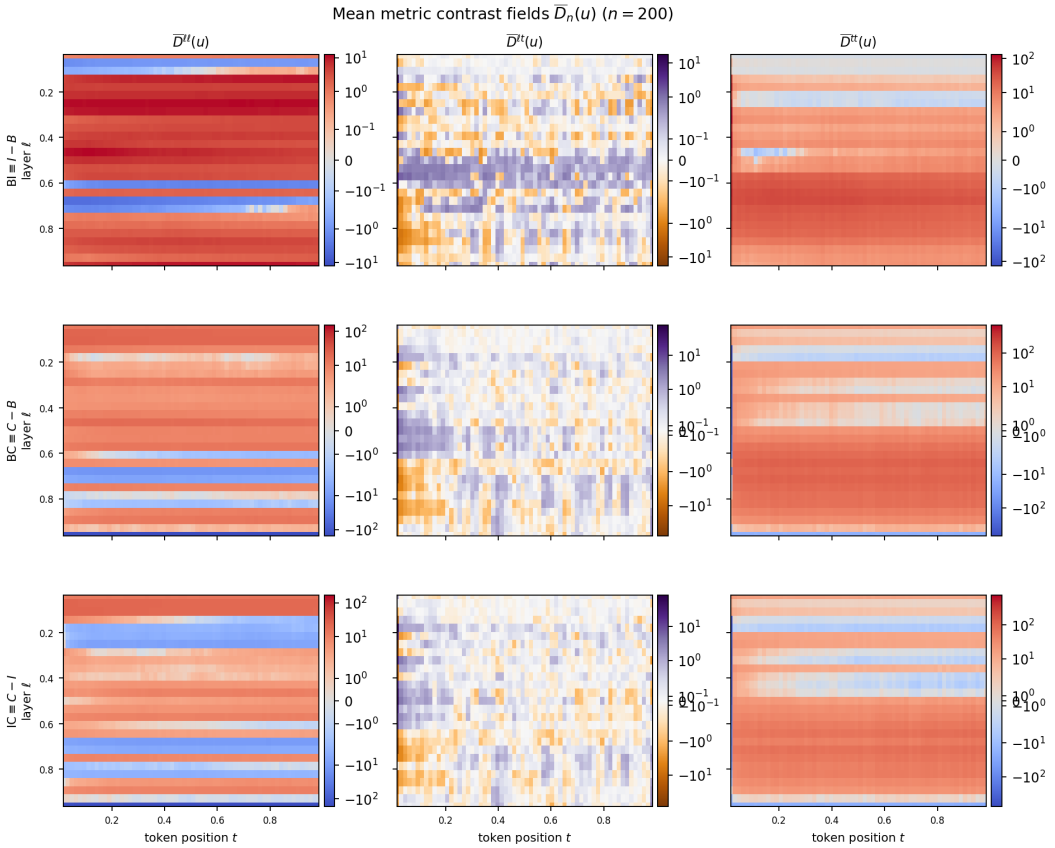


Figure 4: Mean metric contrast fields $\overline{D}_n(u)$ ($n = 200$). Rows: $c = \mathrm{BI} \equiv I - B$, $c = \mathrm{BC} \equiv C - B$, and $c = \mathrm{IC} \equiv C - I$. Columns: vech components $\overline{D}^{\ell\ell}$, $\overline{D}^{\ell t}$, and $\overline{D}^{tt}$.

larger on the metric-field scale. The two $n = 500$ experiments yield nearly identical results under fixed and heterogeneous context lengths, indicating that modest instance-level variation in context length has little effect on the estimated contrasts or testing conclusions.

Figure 4 shows the mean metric contrast fields $\overline{D}_n(u)$ for the three contrasts. Rows are contrasts (BI, BC, IC), and columns are the three vech components ($\ell\ell$, $\ell t$, $tt$).

## D Continuation-training experiment: branches and detection

Table 6 lists the different branches off the fork checkpoint $M^\star$ in Section 5.1. All branches start from the common fork checkpoint $M^\star$ and share the same training setup other than different training corpus. Except for the C1 pair, every branch runs under a bit-deterministic execution fashion.

Table 6: Continuation branches. The seven dose branches share the coupled slot schedule (exact nested mixtures, identical batch order). Contrasts: C0 = re-execution vs. $\rho = 0$, C1 = the two nondeterministic runs against each other, C2 = permuted batch order vs. $\rho = 0$, and each dose vs. $\rho = 0$.

| Branch | Dose $\rho$ | Role | Varies only |
|---|---|---|---|
| $\rho$-reference | 0 | reference continuation | — |
| dose branches ($\times 6$) | $0.01, 0.05, 0.10, 0.25, 0.50, 1$ | treatment | WikiText-103 slot substitution |
| C0 re-execution | 0 | determinism gate | nothing (byte-identical) |
| C1a, C1b | 0 | execution control | nondeterministic GPU kernels |
| C2 order | 0 | batch-order control | slot-order permutation |

**Comparison-method definitions.** For each interior layer $p$, stack the binned interior hidden states of the two checkpoints into paired matrices $A_p, B_p \in \mathbb{R}^{m \times d}$ with one row per (prompt, interior token bin), $m = n\,(K_t - 2)$ and $d = 896$, where $\tilde{A}_p$ denotes column centering. Each method below yields one value per interior layer, and the tables report the mean over the $K_\ell - 2$ interior layers. All methods are evaluated on the identical stored states used by the metric-field methodology.

- *Linear CKA* (Kornblith et al., 2019): $\mathrm{CKA}_p = \|\tilde{A}_p^\top \tilde{B}_p\|^2 / \big(\|\tilde{A}_p^\top \tilde{A}_p\|\,\|\tilde{B}_p^\top \tilde{B}_p\|\big)$, reported as $1 - \mathrm{CKA}_p$.
- *SVCCA* (Raghu et al., 2017): PCA-truncate $\tilde{A}_p, \tilde{B}_p$ to the leading components explaining $99\%$ of variance ($k_a, k_b$ directions), whiten, and take the canonical correlations $\rho_1 \geq \rho_2 \geq \cdots$ (singular values of the whitened cross-covariance). Then $\mathrm{SVCCA}_p$ is the mean of the top $\min(k_a, k_b)$ correlations, reported as $1 - \mathrm{SVCCA}_p$.
- *PWCCA* (Morcos et al., 2018): the same canonical correlations weighted by $\alpha_i \propto \sum_j |\langle h_i, \tilde{a}_j \rangle|$, where $h_i$ is the $i$-th canonical direction on the reference side (always the $\rho = 0$ branch) and $\tilde{a}_j$ ranges over the columns of $\tilde{A}_p$, reported as $1 - \mathrm{PWCCA}_p$.
- *Shape metric* (Williams et al., 2021): the angular shape distance $\theta_p = \arccos\big(\sum_k \sigma_k(\tilde{A}_p^\top \tilde{B}_p) / (\|\tilde{A}_p\|\,\|\tilde{B}_p\|)\big)$.
- *MMD$^2$ and energy distance* (Gretton et al., 2012; Székely & Rizzo, 2013): pool each prompt's interior token bins into a single $d$-vector per layer, giving $n$ paired observations per checkpoint. MMD$^2$ is the unbiased estimator with an RBF kernel and per-layer median-heuristic bandwidth on the pooled two-sample set, and the energy distance is $2\,\bar{d}(X, Y) - \bar{d}(X, X') - \bar{d}(Y, Y')$ with Euclidean distances. In both, the paired same-prompt cross terms are excluded, so byte-identical checkpoints score exactly zero.
- *Weight drift* (Zhong & Raghunathan, 2026): $\big(\sum_W \|\Delta W\|_F^2\big)^{1/2}$ over all 169 two-dimensional parameter matrices of the checkpoint pair (data-free, since the activation-response component of that method requires additional forward passes and is not computed).
- *Output KL*: the per-token full-vocabulary $\mathrm{KL}\big(p_{\mathrm{ref}}(\cdot \mid x) \,\|\, p_{\mathrm{branch}}(\cdot \mid x)\big)$, averaged over positions and prompts and reported separately per evaluation domain.

Final-checkpoint point estimates, each divided by the same method's C2 batch-order value, appear in Table 1 of Section 5.1.

## E Paired-run scenarios for tolerance-based screening

Table 7 records the eight from-scratch Qwen2.5-0.5B pairs used in Section 5.2. Both runs in every pair share the same architecture, optimizer, and $n = 200$ held-out TinyStories validation prompts. The deterministic and nondeterministic execution pairs were recorded for 100 steps, whereas the remaining pairs for 500 steps.

Table 7: Paired-run scenarios used for tolerance-based checkpoint screening.

| Scenario | Run 1 | Run 2 | Factors held fixed | Factors varied | Experimental role |
|---|---|---|---|---|---|
| Deterministic control | From-scratch TinyStories run with shared initialization, corpus, and batch order, under bit-deterministic execution | Independent re-execution of the same recipe | Architecture, initialization, corpus, batch order, and execution | None | Byte-identical negative control |
| Nondeterministic execution | Same TinyStories recipe as the control, without deterministic kernels | Independent re-execution of that recipe | Architecture, initialization, corpus, and batch order | GPU kernel nondeterminism | Execution-noise floor |
| Different batch order | TinyStories, shared initialization, first batch order | Same initialization and corpus, permuted batch order | Architecture, initialization, and corpus | Batch order | SGD-path control |
| Bootstrap-resampled data | TinyStories, shared initialization and training pool | Same initialization and pool, independent with-replacement batches | Architecture, initialization, corpus, and pool order | Per-step resampling | Sampling-stochasticity control |
| Different initialization, same data | TinyStories, first initialization, shared batch order | Second initialization, same corpus and batch order | Architecture, corpus, and batch order | Initialization | Initialization perturbation |
| Different initialization and data | TinyStories, first initialization and batch order | Second initialization and permuted batch order | Architecture and corpus | Initialization and batch order | Combined init-and-order perturbation |
| Different corpus, same initialization | TinyStories, shared initialization and batch order | WikiText-103, same initialization and batch order | Architecture, initialization, and batch order | Training corpus | Corpus shift |
| Different initialization and corpus | TinyStories, first initialization and batch order | WikiText-103, second initialization and permuted batch order | Architecture | Initialization, batch order, and corpus | Combined corpus shift |