THE INTRINSIC GEOMETRY OF A TASK

Anonymous authors
Paper under double-blind review

ABSTRACT

Understanding why deep learning models generalize on a task requires explaining how their internal representations encode information relevant to that task. Although empirical studies have revealed regularities in representation geometry, a theory connecting generalization to this geometry remains incomplete. We study this connection using the Fisher–Rao geometry induced by cross-entropy: a task already has a geometry, and that geometry is the reference a representation must match. A representation is sufficient if it recovers the task's conditional laws, and canonical if it is also minimal. Sufficiency implies isometry to this geometry; local deviations from a sufficient predictor determine excess loss. Pulling the same geometry back to parameter space yields local loss bounds and a criterion for pruning. Experiments confirm the theorems in the regime they assume and show that pullback residual length predicts pointwise excess loss. We then operationalize this geometry through pruning experiments.

1 INTRODUCTION

Understanding generalization in deep learning remains one of the central challenges in machine learning research (Goodfellow, 2016; Mohri et al., 2018). Yet the underlying reasons for recent empirical breakthroughs remain unclear, hence the characterization as a black-box (Chatterjee & Zielinski, 2022). This challenge is increasingly critical as AI models are deployed in high-stakes applications. Because a model is compositional, generalization at its output depends on its internal processes. An account of how and why a model generalizes must therefore describe the representations those processes form (Bengio et al., 2014). What those representations are, how learning forms them, and which of their properties support generalization remain open.

Recent research has focused on empirical investigations of representation learning, such as mechanistic interpretability (Bereska & Gavves, 2024; Sharkey et al., 2025). Yet a key limitation remains: post-hoc analysis uses the model itself to investigate the task. Without a ground-truth reference, it cannot provide a full explanation for why generalization is guaranteed. A full account of generalization should be predictive: it should allow us to anticipate *when* and *why* a model will succeed or fail at generalizing. Addressing this challenge requires a *theory* to explain the empirical observations.

In this work, we take a normative approach to this problem. We characterize the required geometric properties of canonical internal representations independently of any particular model architecture. This work focuses specifically on classification and autoregressive tasks.

Our approach proceeds as follows. First, we introduce the notion of a canonical, task-intrinsic representation. We define the task as a geometric object and study how models realize this object in their internal coordinates. Second, we define a metric on internal representations and relate geometric changes to changes in loss. Third, we describe and operationalize the induced parameter geometry to analyze parameter pruning. Finally, we test the resulting predictions empirically.

2 TASK GEOMETRY AND CANONICAL REPRESENTATION

We first introduce necessary background concepts. This enables us to formalize the task geometry, the intrinsic, coordinate-free manifold structure underpinning the task, and its corresponding set of coordinate charts in deep learning models.

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.

2.1 CONCEPTS

Task. A *task* is a tuple $T = (\mathcal{X}, \mathcal{Y}^n, P, \ell)$, where $\mathcal{X}$ is the input space, $\mathcal{Y}$ a finite output alphabet, P a joint distribution on $\mathcal{X} \times \mathcal{Y}^n$, and ℓ the logarithmic loss. Write P_X for the marginal of X. The task induces the conditional map $\tau_T : \mathcal{X} \to \Delta(\mathcal{Y}^n), \qquad x \mapsto P(Y_{1:n} \mid X = x)$. Classification is recovered by taking $n = 1$. For the geometric results below, we assume that the task image and model predictions under consideration lie in the simplex interior.

Model. A model specifies a hypothesis class $\mathcal{H}$ of predictors $f : \mathcal{X} \to \Delta(\mathcal{Y}^n)$ and, given a finite sample from P, returns an estimate $\hat{f} \in \mathcal{H}$. We call an intermediate map appearing in the computation of f an *internal representation*.

Generalization. Learning task T consists in finding a predictor $f \in \mathcal{H}$ with low population risk: $\mathcal{R}(f) = \mathbb{E}_X \left[D_{\mathrm{KL}}\big(\tau_T(X) \,\|\, f(X)\big) \right]$, where $\tau_T(X), f(X) \in \Delta(\mathcal{Y}^n)$. We write $\mathcal{R}^\star = \mathcal{R}(\tau_T) = 0$ for the Bayes risk.

2.2 TASK GEOMETRY

A task T maps each input to a point in the sequence simplex $\Delta(\mathcal{Y}^n)$. The simplex interior is a statistical manifold whose points are probability distributions over sequences (Amari, 2016; Tan et al., 2024). The conditional map τ_T selects the subset of distributions required by the task; we call this subset the *task image*. On smooth regions contained in the simplex interior, the task image inherits a manifold structure.

The loss ℓ equips the statistical manifold with a divergence between conditional laws. Under cross-entropy, the excess risk is the Kullback-Leibler divergence (Amari, 2016). Its local second-order expansion induces the Fisher-Rao metric. On a finite sample space, this metric is unique up to scale among Riemannian metrics monotone under Markov morphisms (Cencov, 2000).

Definition 1 (Task geometry) *The* task geometry *of T is its task image in $\Delta(\mathcal{Y}^n)$ equipped globally with the divergence induced by ℓ and, on smooth interior regions, locally with the metric induced by the divergence.*

Under cross-entropy, these are the Kullback–Leibler divergence and the Fisher–Rao metric, respectively. The task geometry quantifies how changes in the input change the required output distribution. It is uniquely determined by T and independent of any model coordinate system.

2.3 CANONICAL REPRESENTATIONS

While task geometry is coordinate-free, a model realizes it through an internal representation. A predictor $f = \pi \circ \phi$ decomposes into a representation map $\phi : \mathcal{X} \to \mathcal{Z}$ and an output readout $\pi : \mathcal{Z} \to \Delta(\mathcal{Y}^n)$. Recovering τ_T is sufficiency; that alone still allows ϕ to keep task-irrelevant variation. A canonical representation is the minimal sufficient case. It is a coordinate system for the task image without additional noise.

Definition 2 (Canonical representation) *A representation $\phi : \mathcal{X} \to \mathcal{Z}$ is* canonical *for T if there exists a measurable readout $\pi : \mathcal{Z} \to \Delta(\mathcal{Y}^n)$ such that $\pi \circ \phi = \tau_T$ (P_X-a.s.) and ϕ is constant on the fibers of τ_T: $\tau_T(x_1) = \tau_T(x_2) \implies \phi(x_1) = \phi(x_2)$ (P_X-a.s.).*

The map ϕ distinguishes two inputs if and only if the task does. In other terms, ϕ factors through τ_T and π is injective on $\phi(\mathcal{X})$. Any two such representations are therefore in bijection with the same task image, and with each other. If the restricted readouts are diffeomorphisms onto that image, the bijection is a diffeomorphism.

3 GENERALIZATION AND ISOMETRY

We now connect the canonical representation of a task and its internal geometry to generalization.

3.1 PULLBACK AND GEOMETRY

To characterize the internal geometry of task generalization, we begin by defining the pseudometric that the model inherits from the task. This sets the foundation for understanding the geometric constraints underlying generalization. The model induces a geometry on its input or representation space by pulling back the Fisher-Rao metric from the output sequence simplex $\Delta(\mathcal{Y}^n)$ via the predictor map f. That is, since f maps each input x to a distribution over sequences, we can measure distances in input space by how their images under f are separated in the output simplex, using the Fisher-Rao metric.

Definition 3 (Representation Fisher-Rao pullback) *Let $G_F(p)$ denote the matrix of the bilinear form of two tangent vectors u,v in $\Delta(\mathcal{Y}^n)$. The readout π_θ induces the pullback $g_{\mathcal{H}}(z) = J_{\pi_\theta}(z)^\top G_F(\pi_\theta(z)) J_{\pi_\theta}(z)$.*

This pullback tensor is positive semidefinite and may be degenerate because directions in $\ker J_{\pi_\theta}(z)$ leave the predictive distribution unchanged to first order and therefore have zero length. At each z, it induces an inner product on $T_z\mathcal{Z}/\ker J_{\pi_\theta}(z)$. If the readout has constant rank, the constant-rank theorem identifies these quotients locally with tangent spaces of the readout image. A global quotient additionally requires the corresponding leaf space to be regular. The pullback tensor is a metric on $\mathcal{Z}$ itself when $J_{\pi_\theta}(z)$ is injective on every tangent space.

The pullback tensor measures an infinitesimal change in z by the change it produces in the model's output distribution. Integrating these infinitesimal lengths along paths defines a pseudometric.

3.2 GENERALIZATION AND ISOMETRY

If we restrict attention to the task image, sufficiency forces the representation to preserve the task geometry. Two inputs differ for the task only insofar as their conditional laws differ, so a representation cannot collapse such a difference and still recover τ_T. Canonicity further forbids task-irrelevant variation on $\phi(\mathcal{X})$. Unused directions may still exist in the ambient space $\mathcal{Z}$ and have zero length under $g_{\mathcal{H}}$; they are simply not used by ϕ.

Proposition 1 (Canonical representations preserve task geometry) *Let ϕ_θ be canonical and let d_F be the geodesic distance induced by g_F. Define $\delta_T(x_1, x_2) = d_F(\tau_T(x_1), \tau_T(x_2))$ and $\delta_{\mathcal{H}}(z_1, z_2) = d_F(\pi_\theta(z_1), \pi_\theta(z_2))$. Then $\delta_{\mathcal{H}}(\phi_\theta(x_1), \phi_\theta(x_2)) = \delta_T(x_1, x_2)$. Hence, on the represented set $\phi_\theta(\mathcal{X})$, ϕ_θ is an isometry onto its image. If the maps are smooth, then, with $g_{\mathcal{H}} = \pi_\theta^* g_F$ and $g_T = \tau_T^* g_F$, we also have $\phi_\theta^* g_{\mathcal{H}} = g_T$.*

The global identity follows directly from $\pi_\theta \circ \phi_\theta = \tau_T$. Under smoothness, the local identity follows from composition: $\phi_\theta^* g_{\mathcal{H}} = \phi_\theta^*(\pi_\theta^* g_F) = (\pi_\theta \circ \phi_\theta)^* g_F = \tau_T^* g_F$. This is a one-way consequence of task sufficiency: distance preservation alone does not imply predictive agreement, because an isometry of the output simplex can map predictions to the wrong conditional laws. Two canonical charts are each isometric to the same task geometry, so the bijection between their images is an isometry of the pullback metrics. If the restricted readouts are smooth, that bijection is a diffeomorphism Ψ with $\Psi^* g_{\mathcal{H},2} = g_{\mathcal{H},1}$. That is the coordinate form of the *universality hypothesis* (Olah et al., 2020).

3.3 REPRESENTATIONAL DEVIATION

A learned representation need not be sufficient, let alone canonical. We define its *representational deviation* by the recovered conditional law: $\Delta_\theta = \mathbb{E}_{X \sim P_X}\left[D_{\mathrm{KL}}\big(\tau_T(X) \,\|\, \pi_\theta(\phi_\theta(X))\big)\right]$. By construction, $\Delta_\theta = \mathcal{R}(f_\theta) - \mathcal{R}^\star$. Thus $\Delta_\theta = 0$ is sufficiency, not canonicity: a predictor can match τ_T and still keep nuisance. This is an output discrepancy, not a distance in $\mathcal{Z}$ to a canonical chart.

We can thereby identify the local cost of leaving a sufficient predictor. For a small representation displacement δz about $\phi_\theta(x)$:

$$D_{\mathrm{KL}}\big(\tau_T(x) \,\|\, \pi_\theta(\phi_\theta(x) + \delta z)\big) = \frac{1}{2}\, g_{\mathcal{H}}(\phi_\theta(x))(\delta z, \delta z) + O(\|\delta z\|^3) \tag{1}$$

when $\pi_\theta(\phi_\theta(x)) = \tau_T(x)$ and the readout is sufficiently smooth. Excess risk is the expectation of this expansion. Its parameter-space form is Theorem 1.

4 Induced Parameter Geometry

We now turn to the parameters which implement this geometry and connect the intrinsic geometry to the model weights.

4.1 Intrinsic Geometry in Parameter Space

Let $\theta \in \Theta$ denote the trainable parameters of a smooth predictor f_θ. A small perturbation $\delta\theta$ changes the prediction by $J_\theta f_\theta(x)\delta\theta$ to first order. Pulling the Fisher-Rao metric back through f_θ therefore defines the parameter-space Fisher matrix $F_\theta = \mathbb{E}_X\left[J_\theta f_\theta(X)^\top G_F(f_\theta(X)) J_\theta f_\theta(X)\right]$. This matrix measures parameter directions by their effect on the predictive law rather than by their Euclidean magnitude. It induces a pseudometric: a nonzero parameter perturbation can have zero Fisher length when it leaves the model's predictions unchanged to first order. If those parameters act through the representation while the readout is held fixed, F_θ is the pullback of $g_{\mathcal{H}}$ from Section 3.1.

Theorem 1 (Local excess risk) *Suppose $f_\theta = \tau_T$ and f_θ is sufficiently smooth near θ. Then there exist a radius $\rho_0 > 0$ and a constant C such that, whenever $\|\delta\theta\| \le \rho_0$,*

$$\left|\mathcal{R}(f_{\theta+\delta\theta}) - \mathcal{R}^\star - \frac{1}{2}\,\delta\theta^\top F_\theta \delta\theta\right| \le C\|\delta\theta\|^3.$$

A sufficient predictor is a critical point of $\mathcal{R}$, so the linear term is absent and the local loss is the Fisher length of $\delta\theta$. Perturbations in low-Fisher directions therefore have little local effect, even when their Euclidean magnitude is not small. Perturbations aligned with large eigenvalues of F_θ increase the loss sharply, which is why parameter magnitude alone is not a task-sensitive pruning criterion. If $f_\theta \neq \tau_T$, the same quadratic is only the second-order term and does not replace $\nabla_\theta \mathcal{R}$.

4.2 Capacity-Constrained Compression

We now use this parameter geometry to describe compression of the parameter space. Partition the parameters into removable groups $\theta_1, \dots, \theta_m$, such as channels, neurons, or attention heads. For a retained set $S \subseteq \{1, \dots, m\}$, let M_S be the corresponding pruning mask, define $\theta_S = M_S\theta$, and write $\delta\theta_S = \theta_S - \theta$. Let $c(S)$ be the number of scalar parameters retained by the groups in S.

For a parameter budget B and local radius ρ, let $\mathcal{S}_{B,\rho} = \{S : c(S) \le B,\ \|\delta\theta_S\| \le \rho\}$ be the admissible local pruning masks. Assuming this set is nonempty, the best excess risk within this local family is

$$D_\theta(B,\rho) = \min_{S \in \mathcal{S}_{B,\rho}} \left[\mathcal{R}(f_{\theta_S}) - \mathcal{R}^\star\right].$$

The budget specifies how much parameter capacity remains, while the choice of S determines which task-sensitive directions are removed.

Theorem 2 (Capacity, compression, and loss) *Suppose the assumptions of Theorem 1 hold and $\rho \le \rho_0$. Then, with the same constant C,*

$$\left|D_\theta(B,\rho) - \frac{1}{2}\min_{S \in \mathcal{S}_{B,\rho}} \delta\theta_S^\top F_\theta \delta\theta_S\right| \le C\rho^3.$$

The theorem turns local compression into a geometry-aware selection problem. At a fixed parameter budget, the preferred admissible mask is the one with the smallest Fisher-weighted perturbation, not necessarily the one that removes the smallest weights. Large one-shot pruning steps need not satisfy this local approximation; the result applies only to groups whose removal produces a sufficiently small parameter perturbation.

For a practical approximation, suppose interactions between parameter groups are neglected, so that F_θ is block diagonal across the groups, and write $F_{\theta,gg}$ for the block on group g. Setting that group

to zero, with $\|\theta_g\| \leq \rho_0$, has local excess-risk score $s_g = \frac{1}{2}\theta_g^\top F_{\theta,gg}\theta_g$. For equal-size groups, retaining the groups with the largest scores minimizes the approximate loss at a fixed budget. For groups with different parameter costs, the same scores define a budgeted selection problem. This yields the Fisher-aware structured-pruning rule evaluated in Section 5.

5 EXPERIMENTS

We now test these predictions. First we place a predictor exactly on the task, and just off it. Then we ask where the loss of a trained model comes from. Finally we use the same geometry to decide which parameter groups to keep.

5.1 SETUP

We use three tasks. The synthetic task is a hidden semi-Markov language with 16 latent states, a 60-token vocabulary, and sequences of length 16–64. The process is known, so the forward algorithm gives the exact next-token distribution and we can score a model against the task itself. We train GRU, Transformer, and MLP predictors on $8{,}000$ and $32{,}000$ sequences, with three seeds, and save 72 checkpoints (18 completed runs).

The other two tasks take human annotations as the target distribution. ChaosNLI-M contributes all $1{,}599$ examples, with 100 annotations each. CIFAR-10H contributes all $10{,}000$ images, with roughly 50 annotations each. We smooth the human counts, so a label that nobody chose keeps a small probability. Each model is then calibrated with a single temperature, fit in five folds, so an example is never calibrated on its own labels. On CIFAR-10H we measure geometry on a fixed sample of $250{,}000$ pairs, stratified so that confident and uncertain predictions are both represented.

5.2 GEOMETRY AND GENERALIZATION

The first check is local. We start from the true conditional and pass its log-probabilities through a single gate before the softmax. When the gate equals one, the predictor is exactly the task, and the chart is canonical. Any other setting moves it off the task. We then ask whether a small move of that gate changes the loss by the amount the Fisher geometry predicts.

On the task, the loss is flat to first order: the slope is below 10^{-18} in absolute value. Curvature alone then accounts for the observed change, with relative error 0.021 and correlation 1.000.

Off the task, the slope matters as well. With the gate at 0.7, curvature alone has relative error 1.088. Adding the slope of the loss brings that error down to 0.003. Figure 1 shows both regimes. Near a canonical predictor, a small change in loss is a matter of curvature. Farther away, the same expansion also needs the slope.

5.3 REPRESENTATIONAL DEVIATION

Average excess loss is, by definition, the gap between the task distribution and the model's predictions. What we test is the local version of that gap. At one input, the size of the correction, measured in the pullback metric, should predict that input's contribution to the loss. We read the correction off the gap between the true log-probabilities and the model's logits, and we weight it by how uncertain the model is there. Euclidean distance in representation space ignores the task metric, so it is the baseline.

On a controlled linear chart whose readout stretches some directions much more than others, the pullback score correlates 0.9995 with the per-example loss (relative error 0.018). The Euclidean baseline correlates -0.082. After training a GRU on $8{,}000$ sequences, with held-out excess loss 0.092, the same score has in-distribution correlation 0.876, against 0.152 for Euclidean distance. Changing the distribution of inputs keeps most of the relationship (correlation 0.824, excess loss 0.118). Changing the task itself raises excess loss to 1.171 and drops the correlation to 0.360. The expansion describes a miss relative to a fixed task. When the task itself changes, that local account no longer applies.

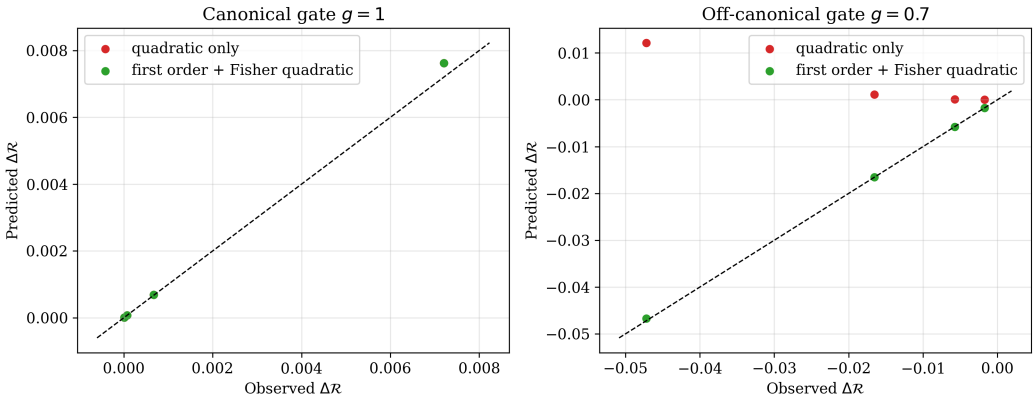


Figure 1: Local change in loss as a single gate moves the predictor. Left: the gate sits on the task, and curvature matches the change. Right: the gate is set to 0.7; curvature alone misses, and adding the slope restores the match.

5.4 Canonical-minimal masks and charts

Two local questions remain. Which groups of parameters should we keep under a budget, and do two canonical descriptions of the same task agree? We split the true log-probabilities into four blocks by frequency and put a gate on each block. With every gate open, the chart is exactly canonical: each class belongs to one block, and the readout recovers the task.

For a small move of one gate, of size 0.03, the block with the least Fisher weight is the block whose removal costs the least loss (correlation 1.0000, relative error 0.012). Ranking blocks by the size of their weights picks a different block (correlation -0.085). Deleting a block outright, by closing its gate, takes the step outside the local approximation (relative error 0.485). The Fisher ranking still names the cheapest block (correlation 0.992). Magnitude's ranking is essentially unrelated to that cost (correlation 0.003). A score that measures the actual change in the output picks the same block as the Fisher score, because on this chart the two are the same quantity. That agreement fails once the predictor is an ordinary trained model.

The same construction compares charts. Two canonical descriptions of one task, related by a linear change of coordinates that stretches some directions far more than others, agree closely in pullback distance (Spearman 1.0000) and less closely in Euclidean distance (Spearman 0.821).

5.5 Learned models and real tasks

Trained models only approximate this picture. Across the 72 synthetic checkpoints, excess loss tracks how distorted the output geometry is, with Pearson correlation 0.984. Predictors with lower loss sit closer to the task geometry, and distances measured directly in activation space lag the pullback metric. Figure 2 shows the same comparison on pairs of inputs for two GRUs. Pullback distance follows the task geodesic, Euclidean distance follows it less closely, and the two models agree with each other once both are measured in the pullback metric.

Perturbing a hidden state on held-out inputs asks a sharper question. The change predicted from that layer's pullback correlates 0.974 with the observed change along Fisher directions, and 0.991 along random directions. Agreement at the output mostly restates how close the predictor already is. The perturbation test asks whether the internal coordinates know which directions move the loss.

Table 1 reports the same comparison on the human tasks, after calibration: excess loss, and how well each model's geometry lines up with the annotators. In both domains the stronger, better-calibrated model is the closer one. Both models remain below the ceiling we get by splitting the annotators in half and comparing the halves.

Universality across models holds only in part once the models are matched by loss. The two NLI models differ in excess loss by 0.049, and the rank correlation between their charts is 0.520, well

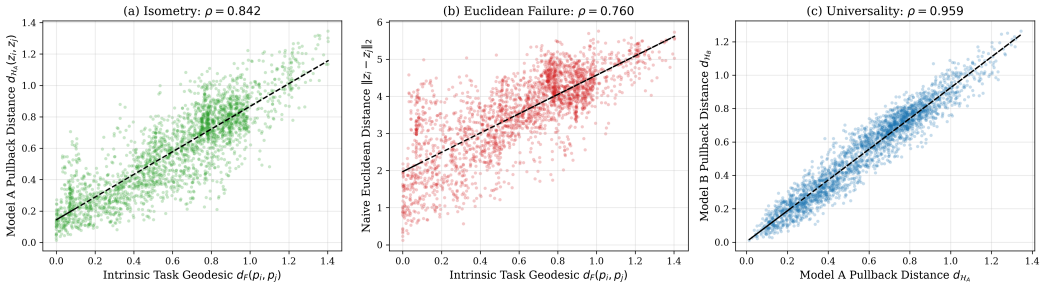


Figure 2: Pairwise distances for two GRUs on the synthetic task. (a) One model's pullback distance against the task geodesic. (b) Euclidean distance in that model's representation against the same geodesic. (c) Pullback distances of the two models against each other.

Dataset	Model	Excess KL	Spearman	Ceiling
ChaosNLI-M	DeBERTa-v3-xsmall	$0.2229\ [0.2138, 0.2321]$	$0.4311\ [0.4019, 0.4605]$	0.901
	DistilBERT	$0.2719\ [0.2622, 0.2820]$	$0.2623\ [0.2358, 0.2919]$	0.901
CIFAR-10H	MobileNetV2	$0.3242\ [0.3160, 0.3321]$	$0.8285\ [0.8239, 0.8335]$	0.890
	ResNet-20	$0.5037\ [0.4937, 0.5147]$	$0.7926\ [0.7870, 0.7985]$	0.890

Table 1: Calibrated excess loss and agreement with the human task geometry. Brackets are 95% intervals from resampling examples. The ceiling compares two halves of the annotators.

below the annotator ceiling of 0.901. The CIFAR models stay out of this comparison: their losses differ enough that agreement in geometry would be confounded with a difference in accuracy.

5.6 Structured compression

Deleting whole groups from a trained network falls outside the local argument of Section 4. Those predictors only approximate the task, and setting a group to zero is a large step. Table 2 and Figure 3 report one-shot deletion that keeps 25% of the parameters belonging to the surviving groups, after which every method receives the same budget to recover. Scoring groups by Fisher weight beats scoring them by magnitude, and beats deleting groups at random. A score that measures the actual change in the output of each deletion does slightly better, and the intervals overlap. On the canonical chart above, those two scores selected the same group. On a network that only approximates the task, measuring the finite change can still improve on the local Fisher score.

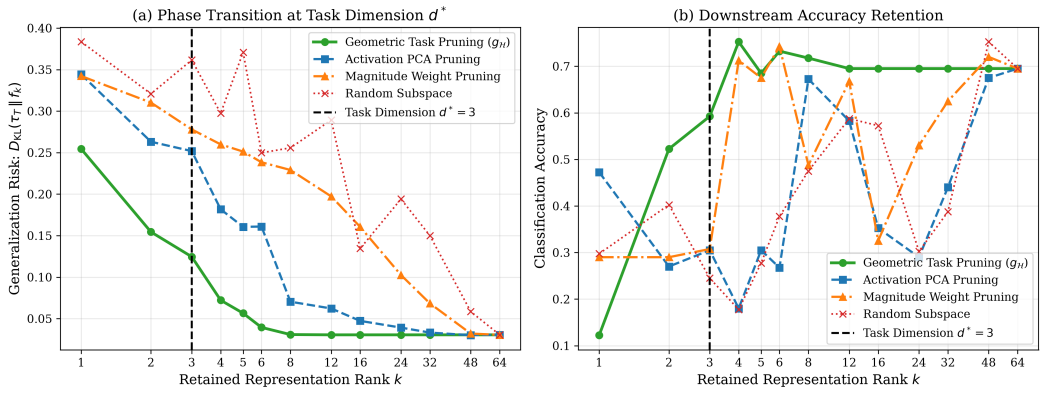


Figure 3: Excess loss after one-shot group deletion, against the fraction of parameters kept. Shaded bands are 95% intervals from resampling examples.

Method	Recovered excess KL	95% interval
Activation	0.628	[0.586, 0.680]
Fisher	0.649	[0.598, 0.705]
Taylor	0.772	[0.731, 0.810]
Uniform	0.798	[0.765, 0.838]
Random	0.820	[0.784, 0.853]
Magnitude	0.848	[0.774, 0.920]

Table 2: Excess loss after keeping 25% of the parameters in the surviving groups, with a matched recovery step.

6 Related Work

Information Geometry. Our work is inspired by information geometry, which treats families of probability distributions as Riemannian manifolds equipped with the Fisher information metric (Amari, 2016). While information geometry has traditionally studied parameter spaces (e.g., natural gradient descent; Amari 2016), we treat the *task itself* as the central object.

Deep Learning Theory. Deep learning theory extends statistical learning theory by studying how architecture, optimization, and information compression shape generalization in neural networks (Tishby & Zaslavsky, 2015; Roberts et al., 2022). Our account is complementary: rather than analyze parameter dynamics or model capacity alone, we characterize generalization through the task-induced geometry of internal representations.

Interpretability. Interpretability seeks to reverse-engineer learned computations into intelligible features and circuits (Olah et al., 2020; Nanda et al., 2023a; Geiger et al., 2024). Related work has uncovered geometric structure in learned representations (Nanda et al., 2023b; Elhage et al., 2022; Park et al., 2024b;a). We provide a model-independent reference for such analyses by specifying the task geometry against which representations can be compared.

7 Discussion

Contributions. We establish a task-intrinsic geometry that provides a model-independent reference for evaluating internal representations. Task-sufficient representations pull back the same output geometry, while representational deviation equals excess KL risk. The corresponding parameter-space pullback characterizes local loss sensitivity and motivates a geometry-aware criterion for structured compression. Experiments confirm the local expansions and the budgeted Fisher mask rule on an exactly canonical chart, show that pullback residual length predicts pointwise excess KL on-distribution, and find that Fisher-aware deletion beats magnitude but not every finite-effect baseline once the predictor is no longer canonical.

Limitations and Future Work. Our results characterize the geometry that a successful representation must preserve, but they do not establish whether a given architecture and optimization procedure can learn that geometry from finite data. Moreover, evaluating representational deviation requires the true conditional law, or a sufficiently accurate estimate of it; on realistic tasks, estimation error may therefore be confounded with distortion of the learned representation. Future work should derive finite-sample estimators and identify conditions on the task, hypothesis class, and optimization procedure under which canonical representations are identifiable and learnable. These results would connect the normative geometry developed here to an end-to-end theory of when training recovers it.

AI use statement

We used generative AI tools to help develop the framing, literature review and implement experiments. We did not use generative AI tools to generate synthetic datasets or perform analysis. We additionally used generative AI tools to draft and edit prose, write and revise experiment code. We take responsibility for the final content of this work.

Ethics statement

The experiments use only public datasets and a synthetic sequence process we generate ourselves. ChaosNLI-M and CIFAR-10H contain existing human annotations; we did not collect new human-subject data.

References

Shun-ichi Amari. *Information geometry and its applications*, volume 194. Springer, 2016.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation Learning: A Review and New Perspectives, April 2014. arXiv:1206.5538 [cs].

Leonard Bereska and Efstratios Gavves. Mechanistic Interpretability for AI Safety – A Review, August 2024. arXiv:2404.14082 [cs].

Nikolai Nikolaevich Cencov. *Statistical decision rules and optimal inference*. Number 53. American Mathematical Soc., 2000.

Satrajit Chatterjee and Piotr Zielinski. On the Generalization Mystery in Deep Learning, June 2022. arXiv:2203.10036 [cs].

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy Models of Superposition, September 2022. arXiv:2209.10652 [cs].

Atticus Geiger, Duligur Ibeling, Amir Zur, Maheep Chaudhary, Sonakshi Chauhan, Jing Huang, Aryaman Arora, Zhengxuan Wu, Noah Goodman, Christopher Potts, and Thomas Icard. Causal Abstraction: A Theoretical Foundation for Mechanistic Interpretability, August 2024. arXiv:2301.04709 [cs].

Ian Goodfellow. Deep learning, 2016.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability, October 2023a. arXiv:2301.05217 [cs].

Neel Nanda, Andrew Lee, and Martin Wattenberg. Emergent Linear Representations in World Models of Self-Supervised Sequence Models, September 2023b. arXiv:2309.00941.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 5(3):e00024–001, 2020.

Kiho Park, Yo Joong Choe, Yibo Jiang, and Victor Veitch. The Geometry of Categorical and Hierarchical Concepts in Large Language Models, June 2024a. arXiv:2406.01506 [cs, stat].

Kiho Park, Yo Joong Choe, and Victor Veitch. The Linear Representation Hypothesis and the Geometry of Large Language Models, July 2024b. arXiv:2311.03658 [cs, stat].

Daniel A Roberts, Sho Yaida, and Boris Hanin. *The principles of deep learning theory*, volume 46. Cambridge University Press Cambridge, MA, USA, 2022.

Lee Sharkey, Bilal Chughtai, Joshua Batson, Jack Lindsey, Jeff Wu, Lucius Bushnaq, Nicholas Goldowsky-Dill, Stefan Heimersheim, Alejandro Ortega, Joseph Bloom, Stella Biderman, Adria Garriga-Alonso, Arthur Conmy, Neel Nanda, Jessica Rumbelow, Martin Wattenberg, Nandi Schoots, Joseph Miller, Eric J. Michaud, Stephen Casper, Max Tegmark, William Saunders, David Bau, Eric Todd, Atticus Geiger, Mor Geva, Jesse Hoogland, Daniel Murfet, and Tom McGrath. Open Problems in Mechanistic Interpretability, January 2025. arXiv:2501.16496 [cs].

Zhiquan Tan, Chenghai Li, and Weiran Huang. The Information of Large Language Model Geometry, February 2024. arXiv:2402.03471.

Naftali Tishby and Noga Zaslavsky. Deep Learning and the Information Bottleneck Principle, March 2015. arXiv:1503.02406.