Quantifying Model Uniqueness in Heterogeneous AI Ecosystems

Lei You[1]

[1]Department of Engineering Technology, Technical University of Denmark, Lautrupvang 15, 2750 Ballerup, Capital Region, Denmark

As AI systems evolve from isolated predictors into complex, heterogeneous ecosystems of foundation models and specialized adapters, distinguishing genuine behavioral novelty from functional redundancy becomes a critical governance challenge. Here, we introduce a statistical framework for auditing model uniqueness based on In-Silico Quasi-Experimental Design (ISQED). By enforcing matched interventions across models, we isolate intrinsic model identity and quantify uniqueness as the Peer-Inexpressible Residual (PIER), i.e. the component of a target's behavior strictly irreducible to any stochastic convex combination of its peers, with vanishing PIER characterizing when such a routing-based substitution becomes possible. We establish the theoretical foundations of ecosystem auditing through three key contributions. First, we prove a fundamental limitation of observational logs: uniqueness is mathematically non-identifiable without intervention control. Second, we derive a scaling law for active auditing, showing that our adaptive query protocol achieves minimax-optimal sample efficiency ($d\sigma^2\gamma^{-2}\log(Nd/\delta)$). Third, we demonstrate that cooperative game-theoretic methods, such as Shapley values, fundamentally fail to detect redundancy. We implement this framework via the DISCO (Design-Integrated Synthetic Control) estimator and deploy it across diverse ecosystems, including computer vision models (ResNet/ConvNeXt/ViT), large language models (BERT/RoBERTa), and city-scale traffic forecasters. These results move trustworthy AI beyond explaining single models: they establish a principled, intervention-based science of auditing and governing heterogeneous model ecosystems. The code is available on https://github.com/youlei202/ISQED.

Correspondence: leiyo@dtu.dk
Date: February 2026

DTU

Introduction

Machine learning systems that affect users and institutions are rarely deployed in isolation. A content-moderation pipeline may combine pre-filters, topic classifiers, ranking models and large language models (LLMs) [1]. A recommendation platform may maintain multiple recommender models for different device classes, regions and latency budgets. As models are added, adapted and retired over time, the ecosystem can accumulate redundancy: distinct artefacts that behave identically in practice, or whose behaviour could be reproduced at negligible cost by composing existing components. Such ecosystems are increasingly shaped by foundation-model pipelines built on Transformer architectures [2] and expanded via fine-tuning, distillation [3] and parameter-efficient adapters [4].

Governance questions arise naturally in this setting. A regulator or platform operator may want to know whether a proposed target model contributes genuinely new behaviour, incurs additional risk, or is effectively a reparameterisation of existing models. A cloud provider may wish to decommission models that are functionally redundant to save energy or reduce maintenance cost [5]. A user organisation negotiating licences may need to demonstrate that a model it deploys has unique value relative to open-source peers [6–8]. These questions concern the position of a given model *inside an ecosystem* rather than the internal logic of that model.

Explainable AI (XAI) has primarily addressed single-model questions. Feature attribution methods quantify which input dimensions are important for a specific model (e.g., LIME [9], Integrated Gradients [10], and SHAP [11]). Counterfactual explanations describe which perturbations would change a prediction [12]. Representation analysis probes internal layers of neural networks (e.g., SVCCA [13] and CKA [14]). These tools are valuable for understanding an individual model in isolation. They do not, however, measure whether that model is replaceable by other models in the same ecosystem. In multi-model settings, cooperative game-theoretic methods such as Shapley value attribution [15] have been proposed to assign credit among models, but these are designed to share rewards fairly, not to detect redundancy.
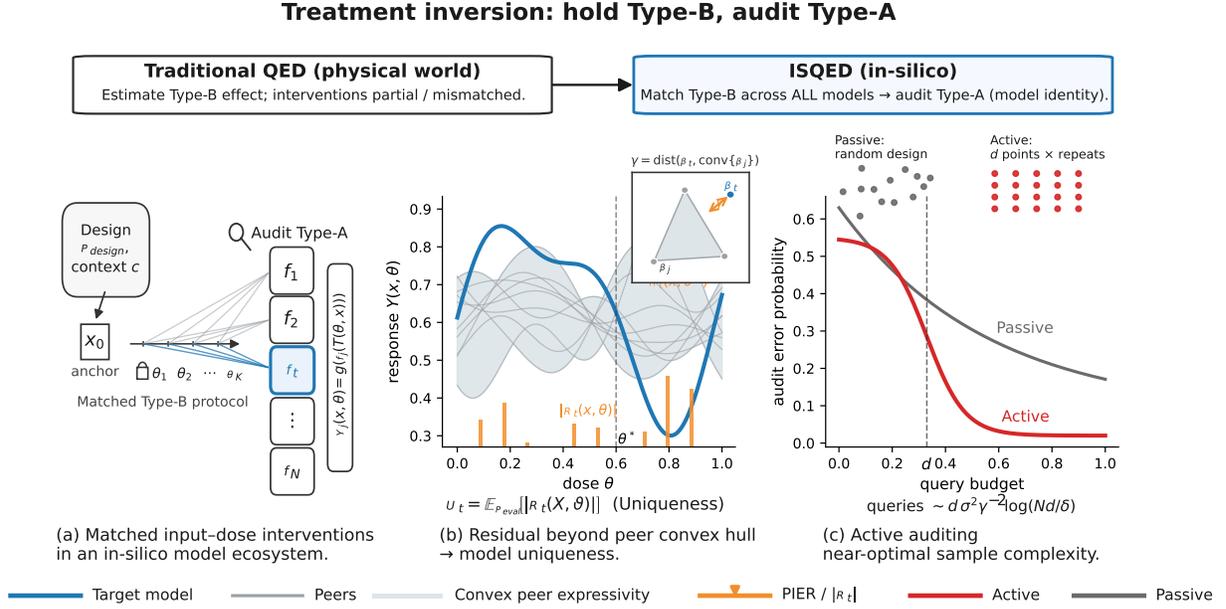
Treatment inversion: hold Type-B, audit Type-A

(a) Matched input–dose interventions in an in-silico model ecosystem.

(b) Residual beyond peer convex hull → model uniqueness.

(c) Active auditing near-optimal sample complexity.

Figure 1: **Treatment inversion: hold Type-B fixed, audit Type-A (model identity).** **(a)**, Matched input–dose interventions in an in-silico ecosystem: an auditor selects an anchor input and a Type-B intervention track (doses $\theta_1, \ldots, \theta_K$) and queries all models on the matched pairs to isolate Type-A (model-identity) differences. **(b)**, Uniqueness as a residual beyond the convex peer hull. The target response $Y_t(x, \theta)$ is projected onto the convex hull of peer responses, yielding the peer-inexpressible residual $R_t(x, \theta)$; its evaluation-average magnitude $\mathcal{U}_t$ summarizes behavioral novelty. **(c)**, Active auditing. Because the auditor controls queries in-silico, adaptive designs can detect non-zero uniqueness with far fewer queries than passive sampling; in a local linear model the minimax-optimal scaling is $d\sigma^2\gamma^{-2}\log(Nd/\delta)$.

Related work on combining multiple trained models includes weight-averaging approaches such as model soups [16].

A central difficulty is to *separate two sources of variation* in responses. The first stems from the external conditions under which models are evaluated, for example the distribution of user queries or the strength of stress-test perturbations. The second stems from the intrinsic identity of the model: its architecture, training data, objective and optimisation. In the language of causal inference [17, 18], the former corresponds to a conventional intervention applied to model inputs, which we refer to as a "Type-B" treatment, while the latter reflects a unit-level effect associated with intrinsic model identity, which we treat as a "Type-A" factor. In observational domains such as policy evaluation, quasi-experimental designs [19, 20] attempt to estimate the effect of a Type-B intervention on units that cannot all be manipulated, for example, when only one state changes its law. Here we are interested in the opposite quantity: the effect of the Type-A treatment, conditional on a controlled pattern of Type-B interventions, in an in-silico environment where we can manipulate all models.

In an in-silico ecosystem we can query every model on matched input–dose pairs and design synthetic interventions on inputs. This allows us to impose a common Type-B protocol across the ecosystem and treat differences in responses as arising from Type-A alone. The central idea of this work is to formalise this setting as an *in-silico quasi-experimental design (ISQED)* framework (Figure 1) and to turn the question *"is the target unique relative to its peers?"* into a well-posed statistical problem. Within a fixed context we define a convex peer expressivity class: the family of convex combinations of peer responses in a Hilbert space of scalarised outputs under a fitting design. The population *peer-inexpressible residual*, abbreviated *PIER*, is the orthogonal residual of the target's response after projection onto this class. A population uniqueness functional is obtained by integrating the magnitude of PIER under a possibly distinct evaluation design. These quantities capture precisely the part of the target's behaviour that cannot be reproduced by simple stochastic mixtures of peers under the chosen designs, and therefore cannot be substituted by a fixed routing over the ecosystem.

Our proposed empirical *design-integrated synthetic control (DISCO) estimator* implements this construction using finite

samples, building on the synthetic control literature [21]. A fitting sample is used to estimate projection weights subject to a simplex constraint, and an evaluation sample is used to estimate PIER values and the uniqueness functional. Under mild assumptions, DISCO has the usual statistical properties expected of a measurement instrument: the estimated projection weights converge to their population counterparts, the empirical uniqueness functional converges to the population functional, and its fluctuations admit an asymptotic normal approximation. A finite-sample design error bound separates estimation error from the contribution of finite anchor radius and dose-grid resolution. These *passive* guarantees mean that, when used with fixed designs, DISCO behaves as a calibrated measurement of model uniqueness.

The in-silico nature of the ecosystem provides more than identifiability advantages. Because we can choose where to query models, we can ask how to allocate a limited number of queries to detect uniqueness as quickly as possible [22, 23]. We analyse this question in a local linear structural model in which all models share a known feature map and differ only in their coefficients. In a noiseless setting, a small number of carefully chosen design points suffice to recover all coefficients exactly, and uniqueness reduces to a convex hull membership test in coefficient space. In a noisy setting with sub-Gaussian perturbations, repeated queries at these design points yield an *active auditing* procedure whose sample complexity scales as $d\sigma^2\gamma^{-2}\log(Nd/\delta)$, where d is feature dimension, σ^2 is noise level, γ is a uniqueness margin, N the number of models, and δ is the target error probability. A minimax lower bound and a detection-versus-estimation result show that this dependence on σ^2, γ and δ is optimal up to constants: in the worst case, detecting non-zero uniqueness cannot be easier than estimating the corresponding local-linear parameter. This reveals a fundamental scaling law for in-silico uniqueness audits.

Beyond rate statements, structural properties of the uniqueness functional matter for ecosystem governance. Because the projection is taken onto a convex hull, the resulting PIER is monotone in the peer set: adding peers can only decrease uniqueness. The convex hull is also conservative relative to richer peer expressivity classes such as linear spans or kernel-induced function spaces: uniqueness defined via convex combinations provides an upper bound on uniqueness defined via any nested richer class. The projection weights admit an interpretation as a stochastic routing policy over peers. If convex PIER is close to zero, the target can be replaced by routing queries to peers according to these weights without changing the expected response under the chosen designs. Conversely, if convex PIER remains large, even such simple mixtures fail to explain the target. These properties align convex PIER with the notion of actionable redundancy.

We further show that popular model-level attribution methods are not suited to this task. In particular, a model can be fully redundant in the sense of lying exactly in the convex hull of its peers, hence having zero PIER and zero uniqueness, and nevertheless receive strictly positive Shapley value. This illustrates that cooperative-game attribution is designed for sharing rewards fairly among contributors, not for identifying units that can be removed without loss. We also prove that, when models are only observed under unmatched observational designs, the population uniqueness functional with respect to a common reference design is not identifiable, even given arbitrarily large logs. In other words, the ability to impose matched interventions among models is not merely convenient but mathematically necessary for uniqueness auditing in the sense studied here.

The remainder of the main text is organised as follows. The Results section introduces the ISQED framework, defines PIER and model uniqueness, establishes passive and active theoretical guarantees, examines key structural and conceptual properties, and empirically audits heterogeneous ecosystems across language, vision, and spatiotemporal prediction tasks. The Methods section formalises the constructions and describes the DISCO estimator and active auditing protocols. Detailed proofs and additional technical material are provided in the Supplementary Information.

Results

ISQED Framework for AI Ecosystems

Notation. We use uppercase symbols for random variables and lowercase for their realisations; in particular, $(X, \vartheta) \sim \mathbb{P}_{\text{fit}}$ or $\mathbb{P}_{\text{eval}}$, while (x, θ) denotes a fixed input–dose pair. We reserve θ for a fixed dose value and use ϑ for a random dose drawn from a design distribution. We do not use boldface: vectors and matrices are identified by their declared domains (e.g., $w \in \Delta^{N-1}$, $\phi(x, \theta) \in \mathbb{R}^d$, $\Phi \in \mathbb{R}^{d \times d}$, $\mathsf{X}_{-t}^{\text{fit}} \in \mathbb{R}^{m \times |\mathcal{J}_{-t}|}$). Throughout, $R_t(x, \theta) = R_t(x, \theta)$ denotes the (signed) peer-inexpressible residual, and the scalar uniqueness score is $\mathcal{U}_t = \mathbb{E}_{\mathbb{P}_{\text{eval}}}[|R_t(X, \vartheta)|]$; when we plot a single "PIER" value, we report an aggregate residual magnitude (typically a mean absolute residual).

We consider an input space $\mathcal{X}$ and a dose space Θ. A deterministic intervention map $T : \Theta \times \mathcal{X} \to \mathcal{X}$ perturbs an input x according to a dose parameter θ, for example by scaling, masking or otherwise modifying selected features. A model ecosystem consists of functions $f_j : \mathcal{X} \to \mathcal{Y}$ indexed by $j \in \mathcal{J} = \{1, \ldots, N\}$. A scalarisation map $g : \mathcal{Y} \to \mathbb{R}$

converts model outputs into real-valued scores. The scalarised response of model j at an input–dose pair (x, θ) is

$$Y_j(x, \theta) = g\big(f_j(T(\theta, x))\big) \in \mathbb{R}.$$

The ecosystem is evaluated within a context C taking values in a space $\mathcal{C}_{\text{ctx}}$. Intuitively, the context encodes a local region of the input–dose space, such as a neighbourhood around an anchor input combined with an intervention track. A global design distribution $\mathbb{P}_{\text{design}}$ on $\mathcal{X} \times \Theta \times \mathcal{C}_{\text{ctx}}$ describes how inputs, doses and contexts are generated. Conditioning on a fixed context value $c \in \mathcal{C}_{\text{ctx}}$ yields a design distribution on $\mathcal{X} \times \Theta$. Throughout the results section we fix such a context and suppress it in the notation.

Within the fixed context, we distinguish three roles for design distributions. A fitting design $\mathbb{P}_{\text{fit}}$ determines where responses are observed for learning a convex peer surrogate that best approximates the target's behaviour. An evaluation design $\mathbb{P}_{\text{eval}}$ determines where PIER and uniqueness are assessed. Both $\mathbb{P}_{\text{fit}}$ and $\mathbb{P}_{\text{eval}}$ are typically derived from $\mathbb{P}_{\text{design}}$, which represents the population of interest and provides a conceptual reference, for example the distribution of real user queries. The in-silico setting allows all three to be chosen by the auditor, subject to computational constraints. An honesty condition requires that fitting and evaluation samples are drawn independently from $\mathbb{P}_{\text{fit}}$ and $\mathbb{P}_{\text{eval}}$ respectively, and that the fitting sample is not reused for evaluation. This separates the process of learning a convex peer surrogate from the process of auditing uniqueness.

A target index $t \in \mathcal{J}$ is fixed and its peers are $\mathcal{J}_{-t} = \mathcal{J} \setminus \{t\}$. The vector of peer responses at (x, θ) is

$$\Phi_{-t}(x, \theta) = (Y_j(x, \theta))_{j \in \mathcal{J}_{-t}} \in \mathbb{R}^{|\mathcal{J}_{-t}|}.$$

Convex combinations of these responses form the basic building blocks of the peer surrogate.

A Convex Notion of PIER

The core population quantities are defined in a Hilbert space. For the fixed context and fitting design $\mathbb{P}_{\text{fit}}$, consider the space $L^2(\mathbb{P}_{\text{fit}})$ of measurable functions $h : \mathcal{X} \times \Theta \to \mathbb{R}$ with finite second moment under $\mathbb{P}_{\text{fit}}$. The inner product is

$$\langle h_1, h_2 \rangle = \mathbb{E}_{\mathbb{P}_{\text{fit}}}\big[h_1(X, \vartheta)h_2(X, \vartheta)\big],$$

with induced norm $\|h\|_2 = \sqrt{\langle h, h \rangle}$.

For a weight vector $w \in \mathbb{R}^{|\mathcal{J}_{-t}|}$ and an input–dose pair (x, θ), the convex combination of peer responses is $h_w(x, \theta) = w^\top \Phi_{-t}(x, \theta)$. The convex peer expressivity class is the set of such functions with weights in the simplex,

$$\mathcal{C} = \{h_w : w \in \Delta^{|\mathcal{J}_{-t}|-1}\} \subset L^2(\mathbb{P}_{\text{fit}}),$$

where $\Delta^{|\mathcal{J}_{-t}|-1}$ is the set of non-negative vectors in $\mathbb{R}^{|\mathcal{J}_{-t}|}$ that sum to one. This class captures *all behaviours that can be realised by stochastic routing* over peers with a fixed context-independent routing distribution w.

The squared $L^2(\mathbb{P}_{\text{fit}})$ distance between the target response and a peer mixture is

$$L(w) = \mathbb{E}_{\mathbb{P}_{\text{fit}}}\big[(Y_t(X, \vartheta) - w^\top \Phi_{-t}(X, \vartheta))^2\big].$$

Under a mild identifiability condition, this residual has a unique minimiser w^* in the relative interior of the simplex, and the residual is locally strongly convex around w^*. The function h_{w^*} is the population projection of Y_t onto the closure of $\mathcal{C}$ in $L^2(\mathbb{P}_{\text{fit}})$. The population PIER is defined as the corresponding residual,

$$R_t(x, \theta) = Y_t(x, \theta) - w^{*\top} \Phi_{-t}(x, \theta).$$

If R_t is almost surely zero under $\mathbb{P}_{\text{fit}}$, then the target behaves as a convex mixture of peers under the fitting design. If R_t is not almost surely zero, then the target exhibits behaviour that cannot be expressed by such mixtures.

To aggregate PIER into a scalar measure, an evaluation design $\mathbb{P}_{\text{eval}}$ on $\mathcal{X} \times \Theta$ is introduced. Under mild integrability conditions, the *population uniqueness functional* is defined as

$$\mathcal{U}_t = \mathbb{E}_{\mathbb{P}_{\text{eval}}}\big[|R_t(X, \vartheta)|\big].$$

This functional depends on the fitting design through w^*, on the evaluation design and on the peer set. It may be interpreted as an average magnitude of peer-inexpressible behaviour along the dose–response trajectory specified by $\mathbb{P}_{\text{eval}}$.

The convex formulation has a useful probabilistic interpretation. The weights w^* define a stochastic routing policy that sends a query to peer j with probability w_j^* and returns its scalarised response. The conditional expectation of this routed response is exactly $h_{w^*}(x, \theta)$. When R_t is small in $L^2(\mathbb{P}_{\text{fit}})$ and $\mathcal{U}_t$ is small under $\mathbb{P}_{\text{eval}}$, the target can be substituted, in expectation, by randomised routing over peers according to w^* under both designs. When R_t is exactly zero, such routing is indistinguishable from the target along the PIER metric.

Passive Statistical Guarantees for DISCO

We now propose the DISCO estimator that implements the population construction using finite samples. For a fixed context, a fitting sample $(X_i^{\text{fit}}, \vartheta_i^{\text{fit}})_{i=1}^m$ is drawn independently from $\mathbb{P}_{\text{fit}}$ and used to compute responses $Y_{j,i}^{\text{fit}} = Y_j(X_i^{\text{fit}}, \vartheta_i^{\text{fit}})$ for all models. The peer response vectors are stacked into a design matrix $\mathsf{X}_{-t}^{\text{fit}} \in \mathbb{R}^{m \times |\mathcal{J}_{-t}|}$ and the target responses into a vector $y_t^{\text{fit}} \in \mathbb{R}^m$. The empirical projection weights are obtained by solving a regularised least-squares problem over the simplex,

$$\widehat{w} \in \underset{w \in \Delta^{|\mathcal{J}_{-t}|-1}}{\arg\min} \left\{ \frac{1}{m} \sum_{i=1}^m \left(Y_{t,i}^{\text{fit}} - w^\top \Phi_{-t,i}^{\text{fit}} \right)^2 + \lambda_m \|w\|_2^2 \right\},$$

with a regularisation parameter λ_m that tends to zero as m increases.

An independent evaluation sample $(X_i^{\text{eval}}, \vartheta_i^{\text{eval}})_{i=1}^n$ is drawn from $\mathbb{P}_{\text{eval}}$. The empirical PIER at the i-th evaluation point is

$$\widehat{R}_{ti} = Y_t(X_i^{\text{eval}}, \vartheta_i^{\text{eval}}) - \widehat{w}^\top \Phi_{-t}(X_i^{\text{eval}}, \vartheta_i^{\text{eval}}),$$

and the empirical uniqueness functional is the average absolute PIER,

$$\widehat{\mathcal{U}}_t = \frac{1}{n} \sum_{i=1}^n |\widehat{R}_{ti}|.$$

Under mild regularity, honesty and identifiability conditions, the DISCO estimator behaves as expected of a sound statistical instrument. Appendix A.1 shows that the empirical residual converges uniformly to the population residual on the simplex, the regularisation term vanishes, and the argmin theorem for M-estimators implies that $\widehat{w}$ converges in probability to the population projection weights w^*. A continuous mapping argument then yields convergence in probability of $\widehat{\mathcal{U}}_t$ to $\mathcal{U}_t$. Appendix A.2 shows that under additional smoothness assumptions on the dependence of $|Y_t - w^\top \Phi_{-t}|$ on w, a central limit theorem holds for $\widehat{\mathcal{U}}_t$: when both m and n grow and their ratio tends to a finite positive constant, the scaled difference $\sqrt{n}(\widehat{\mathcal{U}}_t - \mathcal{U}_t)$ converges in distribution to a centred Gaussian law with finite variance. This justifies normal approximations and bootstrap-based confidence intervals for $\mathcal{U}_t$.

A finite-sample design error bound clarifies the role of anchors and dose grids. In many practical setups, evaluation points are associated with nearby anchor points in the input–dose space, and responses at anchors are used as proxies for responses at evaluation points. Under Lipschitz continuity assumptions on the scalarised responses, Appendix A.3 shows that the error between the anchor-based empirical PIER and the population PIER at an evaluation point is bounded by a term proportional to $\|\widehat{w} - w^*\|_2$ (estimation error) plus a term proportional to the input distance between the evaluation point and its anchor and the dose-grid spacing (design error). This decomposition shows that improving either estimation accuracy or anchor/grid resolution decreases PIER error, and that the contributions of statistical and geometric design choices can be analysed separately.

Taken together, these results place PIER and DISCO on a standard statistical footing. The quantities that appear throughout the Results section can be interpreted as population parameters, estimators of these parameters and formal error controls, rather than as heuristic scores.

Active Auditing Achieves Optimal Sample Efficiency

The passive theory treats $\mathbb{P}_{\text{fit}}$ and $\mathbb{P}_{\text{eval}}$ as fixed reference designs. The in-silico setting allows a different regime in which the auditor actively chooses where to query models. To analyse the benefits of this control, we consider a local

linear structural model within a fixed context. A feature map $\phi : \mathcal{X} \times \Theta \to \mathbb{R}^d$ is assumed known, and each model j is parametrised by a coefficient vector $\beta_j \in \mathbb{R}^d$ such that

$$Y_j(x, \theta) = \phi(x, \theta)^\top \beta_j$$

for all (x, θ) in the region of interest. The feature map may arise from hand-crafted basis functions, a representation learnt in a pretraining phase, or from linearisation of differentiable models via Jacobians or neural tangent kernels [24]. The target and peer behaviours are thus encoded in coefficient space. Remark that we do not assume that real-world ecosystems are globally linear. The local linear structural model is a stylised approximation, motivated by first-order Taylor expansions and shared-feature architectures, and is used to expose the best-case sample complexity of active in-silico auditing.

The peer coefficients $\{\beta_j : j \in \mathcal{J}_{-t}\}$ form a finite set in $\mathbb{R}^d$, and their convex hull, denoted $\mathcal{H}_{\text{peer}}$, represents all behaviours attainable by convex mixtures of peers in coefficient space. The target coefficient is β_t. The distance between β_t and $\mathcal{H}_{\text{peer}}$ defines a uniqueness margin γ, which is zero if the target lies in the convex hull and strictly positive otherwise. The question of uniqueness reduces to deciding whether γ is zero or positive using as few queries as possible.

In a noiseless setting, the auditor can exploit the linear structure fully. By choosing d design points (x_ℓ, θ_ℓ) such that the resulting feature matrix

$$\Phi = \begin{bmatrix} \phi(x_1, \theta_1)^\top \\ \vdots \\ \phi(x_d, \theta_d)^\top \end{bmatrix}$$

is invertible, the auditor can recover each coefficient vector exactly. Evaluating model j at these d points yields a response vector $y_j = (Y_j(x_1, \theta_1), \ldots, Y_j(x_d, \theta_d))^\top$. Solving the linear system $\Phi \beta_j = y_j$ gives $\beta_j = \Phi^{-1} y_j$. This holds for all models simultaneously, and the uniqueness question becomes a convex feasibility problem in coefficient space: the target is non-unique if and only if its coefficient vector lies in the convex hull of peer coefficients. Appendix B.1 formalises this argument and shows that exactly d noiseless queries per model suffice for an exact uniqueness decision in the linear structural model.

In a noisy setting, model evaluations are perturbed by randomness. For example, stochastic components in the system or approximate inference may introduce zero-mean noise with variance proxy σ^2. The auditor can still query each model at a fixed set of d design points, but now multiple repetitions are needed to average out noise. Under sub-Gaussian noise assumptions, repeated queries at each design point and a simple averaging-and-inversion scheme yield coefficient estimates $\widehat{\beta}_j$ for all models. The empirical distance from the target estimate to the convex hull of peer estimates is then compared to a threshold. Appendix B.2 shows that, if the uniqueness margin γ is strictly positive, then there exists a constant C depending only on the feature matrix such that

$$r \geq C \frac{\sigma^2}{\gamma^2} \log\left(\frac{Nd}{\delta}\right)$$

repetitions per design point suffice to distinguish $\gamma = 0$ from $\gamma > 0$ with error probability at most δ. The total number of noisy queries per model is dr, which yields a sample-complexity bound of order $d\sigma^2\gamma^{-2} \log(Nd/\delta)$. Larger uniqueness margins permit fewer queries, and the dependence on the number of models and the target error probability is only logarithmic. To validate this scaling law, we simulated audits in a controlled linear structural environment. As shown in Figure 2A, the active auditing strategy (red) achieves a sharp phase transition, driving the error rate to near-zero once the query budget exceeds the feature dimension d. In contrast, the passive strategy (blue) converges significantly slower, requiring an order of magnitude more queries to reach comparable reliability. The empirical convergence rate of the active strategy perfectly matches our theoretical prediction (dashed arrow), confirming the efficiency gain of the in-silico design.

Active auditing might appear to enjoy unique advantages over passive estimation, and one may wonder whether detection of uniqueness could be substantially easier than estimation of the corresponding structural parameter. Appendix B.3 establishes that this is not the case in the worst-case local linear model. In one dimension with constant features and a trivial peer set, the problem reduces to distinguishing whether a Gaussian mean parameter is zero or equal to γ. Classical information-theoretic inequalities yield a minimax lower bound of order $\sigma^2\gamma^{-2} \log(1/\delta)$ noisy queries for any test that decides between these hypotheses with error probability at most δ. At the same time, any estimator of the mean that aims to achieve accuracy of order γ with confidence level $1 - \delta$ requires the same order of queries. Based on this result,
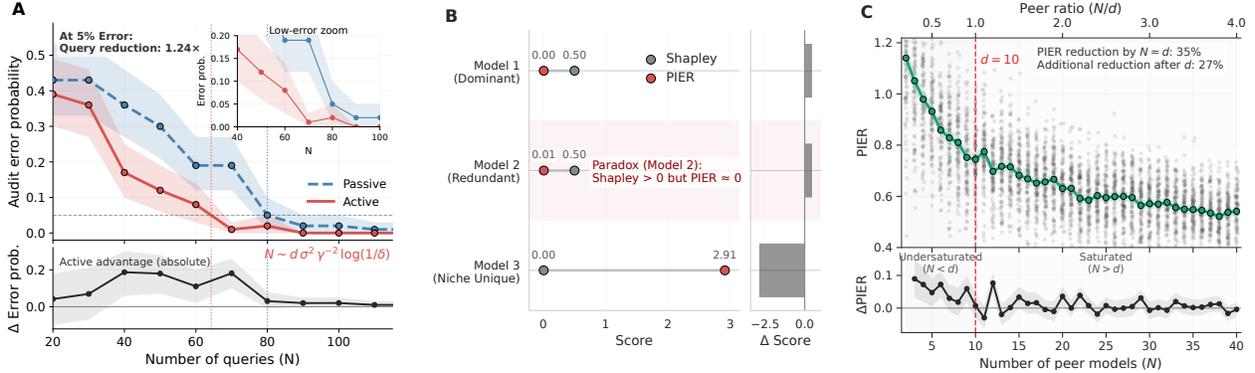
Figure 2: **DISCO and PIER: auditing, attribution, and ecosystem effects.** (A) Active auditing reduces query complexity compared to passive sampling, achieving a $1.24\times$ query reduction at 5% audit error. (B) Attribution versus auditing in a toy ecosystem: Shapley assigns positive value to a redundant model, while PIER assigns near-zero uniqueness and correctly identifies a niche, non-substitutable model. (C) Ecosystem saturation: as the number of peer models N grows beyond the intrinsic task dimension d (here $d = 10$), average uniqueness (mean absolute PIER) collapses, revealing diminishing returns from adding peers.

Appendix B.4 shows that, in this regime, there is no rate separation between detection and estimation: any auditing procedure that reliably detects non-zero uniqueness must operate at essentially the same sample-complexity scale as one that estimates the underlying parameter.

These results identify a fundamental scaling law for in-silico uniqueness audits in a local linear regime. While the constants and precise dependence on d and N depend on the chosen feature map and design, the key message is that active control over design can dramatically reduce the dependence on the ambient input dimension and can achieve optimal dependence on noise level, uniqueness margin and desired error probability.

Conservatism, Monotonicity and Probabilistic Routing

The convex formulation of the peer expressivity class has several structural properties that are useful for ecosystem governance. First, the projection weights admit an operational interpretation. Because w^* lies in the simplex, it can be interpreted as a routing distribution over peers. If PIER is small, a system designer can replace the target by a stochastic router that, given any input–dose pair in the context, sends the query to peer j with probability w_j^* and returns its response. Under both $\mathbb{P}_{\text{fit}}$ and $\mathbb{P}_{\text{eval}}$, the expected scalarised response of this router matches the convex combination h_{w^*}, and small PIER means that this expected response is close to that of the target. The uniqueness functional thus measures how well such a simple routing policy can substitute the target.

Second, uniqueness is monotone in the peer set. Consider two peer sets $\mathcal{J}_1 \subseteq \mathcal{J}_2$ and the corresponding convex expressivity classes $\mathcal{C}(\mathcal{J}_1)$ and $\mathcal{C}(\mathcal{J}_2)$. Enlarging the peer set enlarges the convex hull, and orthogonal projection onto a larger closed convex set in a Hilbert space cannot increase residual norm. The PIER defined with respect to $\mathcal{J}_2$ therefore has no larger $L^2(\mathbb{P}_{\text{fit}})$ norm than the PIER defined with respect to $\mathcal{J}_1$. Boundedness of responses implies the same ordering for the uniqueness functional under $\mathbb{P}_{\text{eval}}$. Appendix A.4 formalises this argument. Empirically, we observe that this monotonic decay slows down significantly near the task's intrinsic dimension, marking a clear saturation threshold for the ecosystem (Figure 2C). This characteristic provides a structural justification for greedy ecosystem pruning strategies in which peers are removed one by one based on their impact on the uniqueness of remaining models.

Third, convex PIER is conservative with respect to richer peer expressivity classes. In addition to $\mathcal{C}_{\text{conv}}$, one may consider the linear span of peer responses or the closure of a reproducing-kernel Hilbert space that contains them. Projections of Y_t onto these larger classes yield alternative residuals and uniqueness functionals $\mathcal{U}_t^{\text{lin}}$ and $\mathcal{U}_t^{\text{ker}}$. Because projection onto larger closed convex sets can only reduce residual norms, the corresponding uniqueness functionals satisfy

$$\mathcal{U}_t^{\text{ker}} \leq \mathcal{U}_t^{\text{lin}} \leq \mathcal{U}_t^{\text{conv}}.$$

Appendix A.4 proves this relationship. It follows that convex PIER provides an upper bound on uniqueness defined with respect to any nested richer class. This is desirable for auditing. If $\mathcal{U}_t^{\text{conv}}$ is small, then the target can be replaced by a

very simple mechanism (a convex mixture or a stochastic router), which is an actionable notion of redundancy. If only a very rich non-linear peer class can explain the target, the cost of replacement may be comparable to training a new large model, and convex PIER will not declare the target redundant.

Uniqueness Versus Attribution and the Limits of Observational Logs

The uniqueness functional studied here measures peer-inexpressible behaviour under a controlled in-silico design. It differs in aim and construction from model-level attribution methods based on cooperative game theory. In Shapley-based attribution [11, 15], models are treated as players in a cooperative game and a characteristic function assigns a performance score to each subset of models. The Shapley value of a model is its average marginal contribution to performance across all coalitions, subject to axioms of efficiency, symmetry and additivity. This notion of contribution is well suited to sharing rewards but does not encode redundancy.

A simple example illustrates the gap. Consider two models M_1 and M_2 with identical scalarised responses on all inputs and doses. Suppose that any system using either one alone achieves score 1 and that using both together yields no further gain. Cooperative game theory assigns a Shapley value of $1/2$ to each model. From an auditing perspective, if M_1 is taken as target and M_2 as peer, the convex peer expressivity class contains M_2, the population PIER of M_1 is identically zero, and its uniqueness functional is zero. In other words, M_1 is fully redundant relative to M_2, yet its Shapley value is strictly positive. Appendix C.1 makes this argument precise. This demonstrates that a positive Shapley value does not imply non-zero PIER and that Shapley-based attribution cannot serve as a sufficient statistic for redundancy detection.

We empirically validated this theoretical divergence in a controlled linear structural ecosystem (Figure 2B). The experiment extends the theoretical setup to a realistic scenario involving a dominant model (M_1), its redundant clone (M_2), and a niche specialist (M_3). Consistent with the theoretical prediction, Shapley value assigns high credit to the redundant clone M_2 due to its performance contribution, masking its redundancy. In contrast, PIER correctly identifies M_2 as having vanishing uniqueness. Furthermore, the experiment reveals a converse failure mode: Shapley neglects the specialist M_3 due to low marginal utility, while PIER identifies it as the most geometrically unique component. This confirms that utility-based attribution and geometry-based auditing are orthogonal objectives.

The ability to impose matched interventions across models is also crucial. In many practical deployments, an auditor may only have access to historical logs in which each model has been queried under its own observational design, for example as part of different A/B tests or routing policies. In such a regime the joint distribution of $(X, \vartheta, Y_j(X, \vartheta))$ for each model j is determined by an unknown design distribution Q_j on $\mathcal{X} \times \Theta$ and the deterministic response function Y_j. One might hope to reconstruct the population uniqueness functional with respect to a reference design $\mathbb{P}^{\star}_{\text{design}}$ from these logs. Appendix C.2 shows that this is impossible in general. There exist two ecosystems that induce identical joint distributions over observational logs for all models and all sample sizes, yet whose uniqueness functionals with respect to $\mathbb{P}^{\star}_{\text{design}}$ differ, with one equal to zero and the other strictly positive. Any procedure based purely on observational logs must therefore fail on at least one of these ecosystems. This non-identifiability result highlights that in-silico control over interventions is not merely a convenience but a structural prerequisite for uniqueness auditing in the sense studied here.

Uniqueness, Robustness and the Shape of the PIER Trajectory

The uniqueness functional $\mathcal{U}_t$ compresses the entire residual trajectory $(x, \theta) \mapsto R_t(x, \theta)$ into a single scalar via an expectation under $\mathbb{P}_{\text{eval}}$. This is a natural summary for many purposes but it inevitably loses information. In particular, a model may exhibit a large uniqueness score either because it consistently behaves differently from peers in a smooth and robust way, or because it reacts erratically to small perturbations, generating high-magnitude residuals only in narrow regions of the dose space.

Appendix D makes this ambiguity explicit. For any prescribed pair of constants $0 < L_{\text{low}} < L_{\text{high}}$, it constructs two targets that share the same peer set and the same uniqueness functional $\mathcal{U}_t$ under a given $\mathbb{P}_{\text{eval}}$, yet whose PIER functions have dramatically different regularity along the dose axis. In one case, the residual is constant in θ and hence perfectly stable, with Lipschitz constant bounded by L_{low}. In the other, the residual oscillates rapidly with θ and has Lipschitz constant at least L_{high}. The two models are indistinguishable at the level of the scalar uniqueness score but have very different robustness profiles. This shows that, without additional inductive bias, the magnitude of $\mathcal{U}_t$ alone cannot separate "capability-driven" uniqueness from uniqueness arising from sensitivity to perturbations.

In practice, this suggests two complementary uses of the PIER framework. The scalar functional $\mathcal{U}_t$ is appropriate for ranking models by their average degree of peer-inexpressible behaviour in a given context and for certifying strong
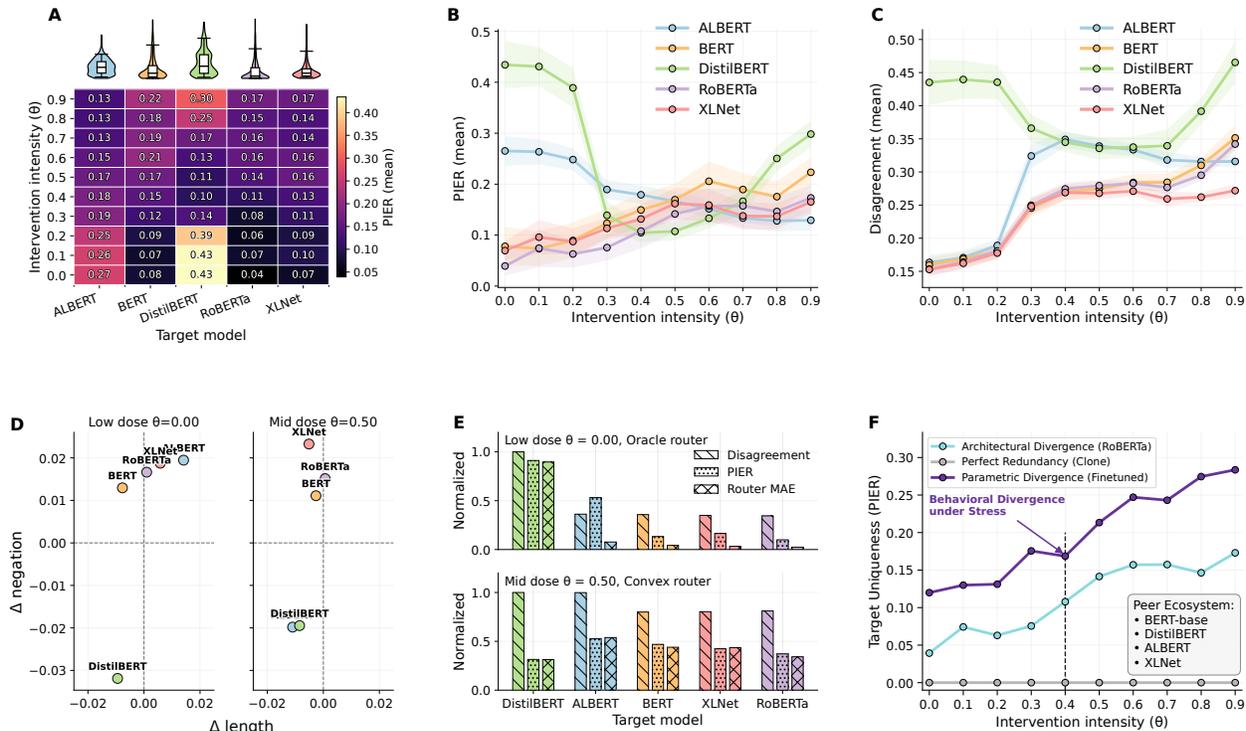
Figure 3: **Auditing an SST-2 ecosystem with DISCO under matched token masking.** PIER reveals dose-dependent uniqueness (A,B) that can differ from raw disagreement (C). Context fingerprints at $\theta = 0.0$ and 0.5 localize which linguistic regimes drive residuals (D). Routing checks validate the interpretation: oracle routing still fails for ALBERT at $\theta = 0.0$, while DISCO-weight convex routing matches all targets at $\theta = 0.5$ (E). Controlled targets separate redundancy, architectural divergence, and parametric divergence (F).

forms of redundancy when it is close to zero. For more fine-grained auditing, especially when robustness to interventions matters, it is natural to inspect the full dose–response profile $\theta \mapsto R_t(x, \theta)$ at representative inputs or to regularise the residual trajectory, for example by penalising its variation along θ. The in-silico design, which treats θ as an explicit controlled coordinate, makes such analyses straightforward to implement.

Use Case Study: Auditing Large Language Models

To demonstrate the practical utility of DISCO, we audited a heterogeneous ecosystem of language models on SST-2 [25, 26] under matched token-masking interventions. We treated the masking intensity θ as a controlled Type-B stress coordinate and used each model's positive-class probability as the scalarized response. We considered a five-model ecosystem comprising BERT-base [27], DistilBERT [28], ALBERT [29], RoBERTa [30], and XLNet [31]. For Fig. 3A–E, we performed a cross-audit on this ecosystem by treating each model in turn as the target and using the remaining four models as its peers, namely an "establishment" peer set that changes with the target. For Fig. 3F, we fixed a single establishment peer set, namely BERT-base, DistilBERT, ALBERT, and XLNet, and we audited three controlled targets to separate sources of uniqueness: (1) a DistilBERT *Clone*, which is a bit-wise copy of the peer DistilBERT and represents perfect redundancy; (2) a RoBERTa model, which represents architectural divergence from the BERT lineage; and (3) a DistilBERT *Finetuned*, which shares the same architecture as DistilBERT but differs by fine-tuning, representing parametric divergence.

Across the cross-audit, Fig. 3A–B shows that uniqueness is strongly dose-dependent. We observe qualitatively different PIER trajectories across models, and these trajectory shapes carry audit-relevant information that a single scalar score cannot capture. For example, DistilBERT exhibits a pronounced mid-dose redundancy window where its PIER drops sharply relative to its low-dose behavior, while BERT, RoBERTa, and XLNet show increasing peer-inexpressible

9

residuals as masking becomes more severe. In contrast, ALBERT exhibits its largest PIER at low dose and then decays as θ increases. This heterogeneity supports a key practical message of our framework: we should treat "uniqueness" as a property *conditional on a controlled intervention protocol*, not as a static label attached to a model.

A second message in Fig. 3B–C is that "being different" is not the same as "being irreplaceable." Disagreement (Fig. 3C) measures the average deviation from individual peers, so it answers a pairwise question. PIER (Fig. 3B) instead measures the residual after projecting onto the convex hull of peer behaviors, so it answers a substitution question. These two views can disagree at specific stress levels, and Fig. 3E turns those disagreements into testable, operational claims about routing.

At low dose $\theta = 0.0$, Fig. 3B and Fig. 3C give different conclusions for ALBERT. PIER ranks ALBERT as highly peer-inexpressible, while disagreement does not single it out. To test what this gap means, we ask the strongest possible substitution question: if we could route each input to the best peer using hindsight, could we match ALBERT's responses? We implement this as an oracle router that selects, for each example, the peer with the smallest absolute error. Fig. 3E (top) shows that ALBERT still incurs a large routing error under this oracle policy. This observation sharpens the interpretation of the low-dose PIER signal. ALBERT is not merely "far" from a particular peer; rather, the ecosystem lacks the behavioral pieces needed to cover ALBERT even with per-example routing. In other words, the residual we observe is not an artifact of a poor mixture choice, and it reflects a genuine uniqueness margin relative to the available peers.

At mid dose $\theta = 0.5$, we observe the converse mismatch. Disagreement suggests that DistilBERT and ALBERT remain especially different from the rest of the ecosystem, but PIER does not treat them as uniquely peer-inexpressible in this regime. Here we test the constructive implication of small PIER, namely *actionable redundancy*. We build a convex router that samples peers using the DISCO projection weights learned from an independent fitting split, and we evaluate its mean absolute error on held-out inputs at $\theta = 0.5$. Fig. 3E (bottom) shows that this simple context-independent routing policy achieves uniformly small error across targets, and DistilBERT is among the easiest models to substitute. This result explains why disagreement can overstate uniqueness: a model can be moderately far from every single peer while still lying close to their convex hull. More importantly, it illustrates a practical advantage of PIER over heuristic difference scores. PIER does not only rank models, but it also produces a concrete replacement policy and it predicts when that policy should work.

Finally, Fig. 3D and Fig. 3F show how our audit goes beyond global ranking. In Fig. 3D, we summarize how each model's residual shifts across two interpretable linguistic contrasts, namely sentence length and the presence of explicit negation. At $\theta = 0.0$, ALBERT lies in the upper-right quadrant, which indicates that its peer-inexpressible behavior concentrates on longer sentences and negation, while DistilBERT lies in the lower-left quadrant, indicating the opposite trend. At $\theta = 0.5$, the models rearrange, and ALBERT moves toward the lower-left cluster, which shows that stress can change which linguistic regimes drive residual behavior. This context sensitivity suggests that ecosystem governance does not need to reduce to a binary keep-or-remove decision. When uniqueness concentrates in specific regimes, we can route those regimes to the unique model and route the remaining traffic to cheaper peers.

In Fig. 3F, we isolate three sources of uniqueness with controlled targets. The DistilBERT clone stays at near-zero PIER across all θ, which validates that DISCO does not hallucinate uniqueness when perfect redundancy exists. RoBERTa shows a stable, non-zero trajectory, consistent with architectural divergence. The fine-tuned DistilBERT variant displays a sharp increase in PIER once θ enters a higher-stress regime, indicating stress-induced behavioral divergence even within the same architecture. Taken together, Fig. 3 shows how DISCO converts a geometric definition of uniqueness into a practical audit workflow: we can characterize dose-dependent uniqueness, localize it to interpretable contexts, and validate replaceability claims through explicit routing policies.

Use Case Study: Auditing Computer Vision Models

We evaluate whether PIER provides actionable auditing signals for modern computer vision ecosystems. We build a peer set spanning standard ImageNet [32] classifiers (ResNet [33], EfficientNet [34], ConvNeXt [35], ViT [36]) and robust [37] or shape-biased [38] variants, and we probe these models under matched interventions. We use a scalarized behavioral response (a confidence score for the ground-truth class) and compute PIER within each target model's ecosystem.

Figure 4A treats adversarial strength ϵ [37, 39] as a continuous stress coordinate and reports relative uniqueness $\text{PIER}(\epsilon)/\text{PIER}(0)$. We see two qualitatively different regimes. For a standard ResNet-50, uniqueness collapses under attack, with PIER falling to roughly $0.55\times$ its clean baseline. This pattern suggests that, once inputs cross a robustness
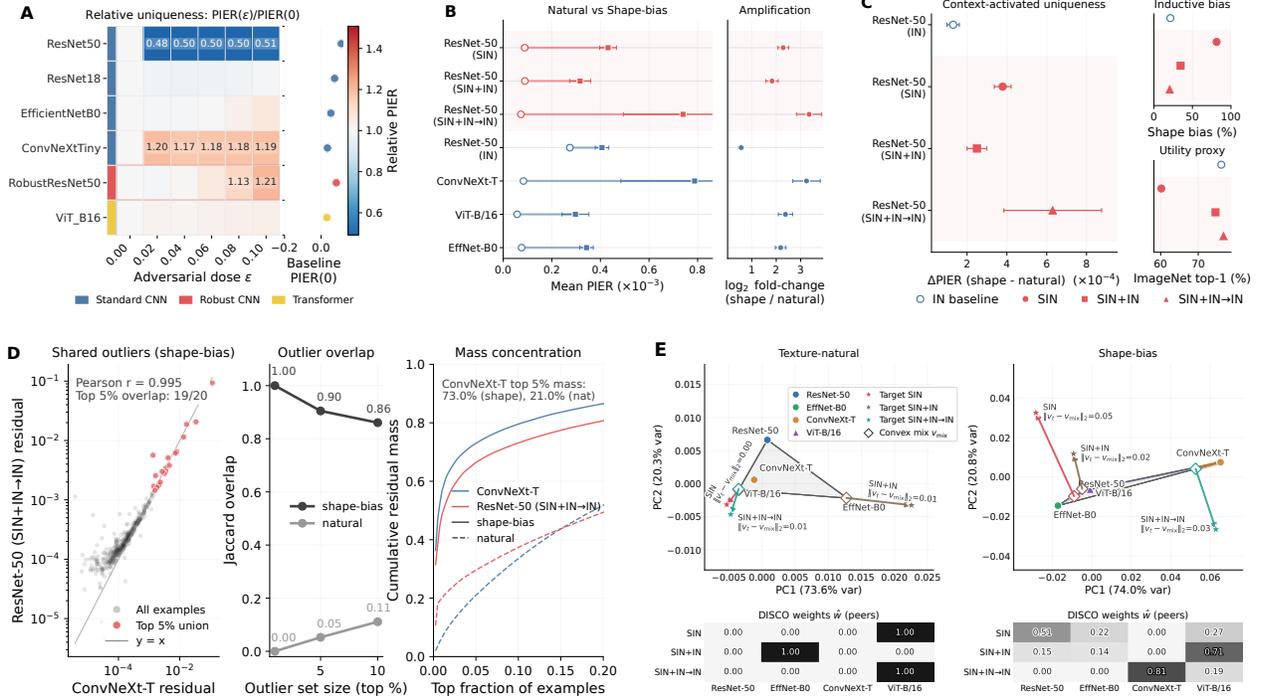
Figure 4: **Auditing vision ecosystems under controlled stressors.** (A) Relative uniqueness across a matched adversarial dose ϵ, reported as PIER(ϵ)/PIER(0), with baseline PIER(0) shown at right. (B) Context audit comparing texture-natural versus shape-biased evaluation, plus the within-model amplification $\log_2(\text{PIER}_{\text{shape}}/\text{PIER}_{\text{nat}})$. (C) Context-activated uniqueness ΔPIER decouples from both inductive bias (shape bias score) and a utility proxy (ImageNet top-1). (D) Heavy-tail diagnostics show that a small set of shared outliers dominates residual mass under shape-bias. (E) Convex-hull geometry: targets move outside the peer hull in the shape-bias context, and DISCO weights explain the closest convex surrogate.

threshold, ResNet-50 fails in ways that look increasingly similar to its peers, so the ecosystem offers little incremental behavioral diversity. In contrast, ConvNeXt shows amplified uniqueness across the same range, with ratios above $1.2\times$, and the robust model shows late-stage amplification. This contrast matters for auditing. If we only measure uniqueness on clean data, we can miss models whose distinctive behavior only appears under stress, which is exactly when deployment risk increases.

In Figure 4B, we switch from adversarial stress to a dataset-context stressor, comparing a texture-natural context to a shape-biased context [38]. We plot a paired dumbbell per target model, so the comparison controls for that target's peer set. Across models, the shape-biased context increases PIER, but the amplification is highly non-uniform. Shape-biased ResNet variants exhibit the strongest lift, and some standard architectures also show large fold-changes. This result supports a core claim of PIER: uniqueness is not an intrinsic model attribute, but a *context-relative* property of the model within a given ecosystem.

Figure 4C summarizes context-activated uniqueness as ΔPIER $= \text{PIER}_{\text{shape}} - \text{PIER}_{\text{nat}}$ and compares it against two common summary axes. First, it does not track an inductive-bias score (shape bias), even though the experimental manipulation directly targets shape versus texture cues. Second, it does not track ImageNet top-1 accuracy, which we use as a simple utility proxy. This decoupling matters for governance. A team that optimizes for accuracy or for a single bias metric can still assemble an ecosystem with large redundancy, or overlook models whose behavior becomes uniquely risky or uniquely valuable under shift. PIER instead measures substitutability against a peer hull, so it captures complementarity that these one-dimensional summaries miss.

To explain why certain models show large context amplification, Figure 4D inspects per-example residuals between each target and its best convex surrogate. Under the shape-bias context, residuals become heavy-tailed: the top few percent

of examples account for most of the residual mass, and ConvNeXt and the shape-biased ResNet share these outliers almost perfectly. Under the texture-natural context, the same overlap collapses toward zero. This diagnostic provides a concrete interpretation for "context-activated" uniqueness. The context shift does not uniformly change behavior across the dataset. Instead, it activates a rare but decisive set of test cases where the ecosystem's failure modes align or diverge, and these cases dominate the uniqueness signal.

Finally, Fig. 4E visualizes the geometric object that PIER measures. We embed the peer models and each target model into a 2D PCA space for visualization, shade the peers' convex hull, and plot the DISCO-induced convex-combination point that best approximates the target in this 2D projection. In the texture-natural context, target points typically lie near (or inside) the peers' hull, and the convex-combination point nearly overlaps with the target, indicating near-complete substitutability. In the shape-bias context, targets more often fall outside the peers' hull; the gap becomes larger and the weights redistribute, revealing which peers explain the closest surrogate and what residual behavior remains irreducible. This panel closes the loop between theory and practice: PIER behaves as a geometric distance-to-hull signal, and DISCO produces an explicit, auditable substitution certificate.

Together, Figure 4 shows that PIER supports high-level governance questions that accuracy and bias metrics do not answer. We can stress-test ecosystems, identify contexts that activate uniqueness, localize the small set of examples that drive irreducible behavior, and extract interpretable substitution structures that explain which models the ecosystem can safely consolidate.

Use Case Study: Multi-City Traffic Predictions

To stress-test DISCO in a high-stakes *model-governance* setting, we audited an ecosystem of $N = 31$ city-level traffic forecasters. Each city model served as a target in turn, and we treated the remaining city models together with a pooled GLOBAL model as its peers. For every target city, DISCO returned (i) a uniqueness score, PIER, and (ii) an interpretable convex decomposition over peers, $\hat{w}$, which we deployed as a *convex router* to substitute the target. This setting let us connect geometric uniqueness to concrete operational questions, namely: which city models are redundant, which ones are irreplaceable, and how much consolidation we can achieve under a fixed performance budget.

In Fig. 5A we plot *relative uniqueness* on the x-axis (PIER normalized by the typical flow scale) against the *replacement impact* on the y-axis, which we define as the signed MAE change when we replace a target city model with its convex router. A positive impact means substitution degrades accuracy, while a negative impact means the router improves over the local model. We observe essentially no linear association (Pearson $r = 0.05$), which highlights a central message of our paper: PIER measures *behavioral* uniqueness relative to peers, not task utility. This decoupling exposes three practically distinct regimes that a single utility score cannot separate. First, many cities sit near the origin in the inset, which indicates low uniqueness and small replacement impact; these models are natural consolidation candidates. Second, a small set of cities occupy the upper-right region (for example, paris), where high uniqueness co-occurs with large positive replacement impact; these models are irreplaceable because no peer mixture reproduces their behavior without paying a clear error penalty. Third, we find *unique-but-unhealthy* cities in the lower-right region (for example, bern), where PIER is high but replacement impact is negative and the GLOBAL model already outperforms the local model (triangle markers). In this regime, uniqueness reflects idiosyncratic failure modes rather than valuable specialization, and DISCO flags these cases precisely because it separates ecosystem geometry from utility.

Fig. 5B demonstrates the mechanism behind many apparent "unique" cities under naive metrics. The left panel compares an ecosystem-level uniqueness score (PIER) to a pairwise proxy that ignores mixtures, and we color points by the *convexity gain* $\log_2(\text{proxy}/\text{PIER})$. Large gains indicate cities that look far from every single peer, yet fall close to the *convex hull* of the peer set. These are exactly the cases where pairwise comparisons systematically overestimate uniqueness, because the true substitute is not one peer but a sparse mixture of several. The right panel makes this geometric fact operational: DISCO returns $\hat{w}$ and shows that high-gap cities admit a compact top-3 decomposition that explains most of the mixture mass. This weight-level explanation matters for governance, because it converts "this city is redundant" into a concrete substitution plan and reveals which peer models form the effective backbone of the ecosystem.

Fig. 5C ties PIER back to convex geometry using two complementary notions of mixture complexity. We summarize how many peers the router effectively uses via $n_{\text{eff}} := 1/\sum_j \hat{w}_j^2$, and we summarize weight concentration via the top-3 mass $m_{\text{top3}} := \sum_{k=1}^3 \hat{w}_{(k)}$ (sorted in descending order). We observe a clear monotone relationship: as relative uniqueness grows, the convex mix becomes less diffuse (Spearman $\rho_s(\text{PIER}, n_{\text{eff}}) = -0.60$), and m_{top3} almost deterministically predicts n_{eff} ($\rho_s(m_{\text{top3}}, n_{\text{eff}}) = -0.94$). This pattern matches the convex-hull view of DISCO. When a target lies
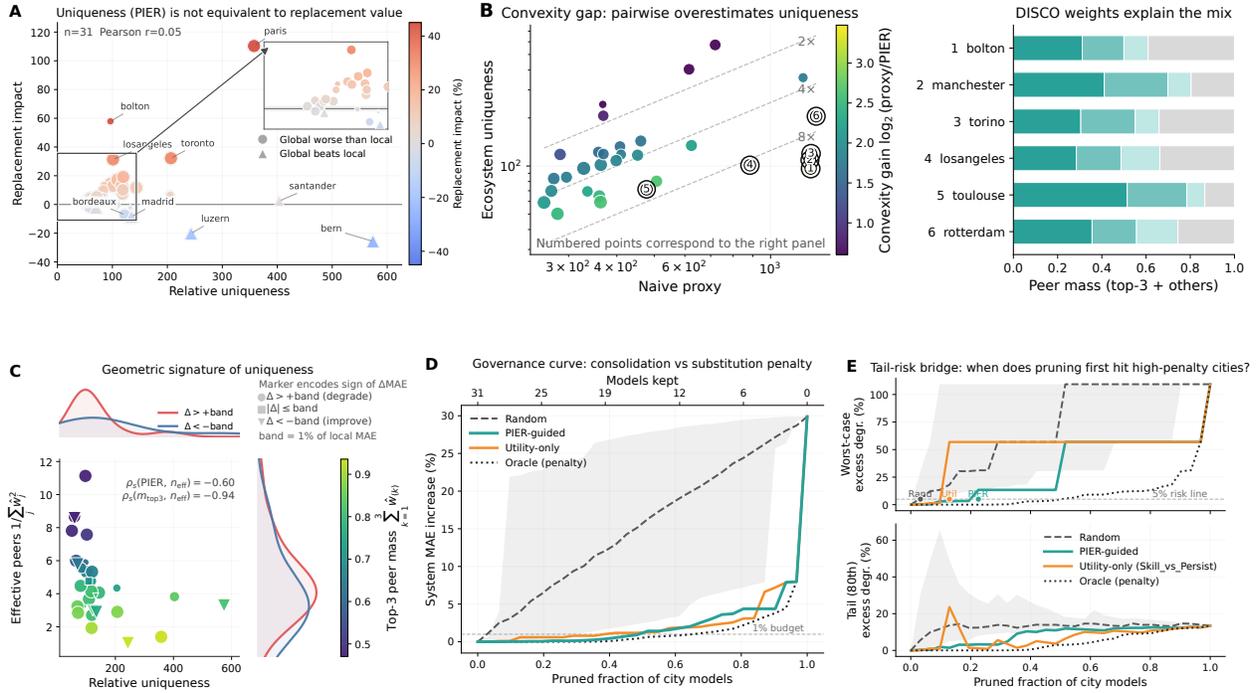
Figure 5: **Auditing traffic models trained on multiple cities.** (A) Relative uniqueness versus signed replacement impact when substituting a local city model with the DISCO convex router (inset shows the dense region; marker shape indicates whether GLOBAL beats the local model). (B) Convexity gap between a pairwise proxy and ecosystem uniqueness (left), with DISCO weights that explain representative high-gap cases (right). (C) Geometric signature linking uniqueness to convex-mix complexity; we report Spearman's rank correlation ρ_s. (D) Consolidation curve: system MAE increase as we prune city models under different pruning orders. *Utility-only* ranks cities by their standalone forecasting skill relative to a persistence baseline, ignoring substitution/replacement penalties. *Oracle (penalty)* ranks cities using access to the true replacement penalty (available only for evaluation, not in practice). (E) Tail-risk bridge: worst-case and tail excess degradation along the pruning path.

outside the peer hull, its projection typically lands on a low-dimensional face supported by a small subset of peers, which concentrates $\hat{w}$ and reduces n_{eff}. Importantly, the marker encoding reveals that high uniqueness can correspond to either degradation or improvement under substitution, which reinforces that PIER is not a utility proxy. Instead, PIER and the weight geometry tell us *how* a model differs from the ecosystem and *how* substitution will depend on a small set of supporting peers.

Fig. 5D turns the audit into a decision procedure. We progressively prune city models and substitute each removed model with a convex router built from the remaining ecosystem. We then measure the system-level MAE increase as a function of the pruned fraction. A random pruning order degrades steadily, while the PIER-guided order stays near the oracle lower bound for most of the path and exhibits a sharp knee only when pruning starts to remove high-PIER cities. A utility-only heuristic based on forecasting skill also looks competitive on mean degradation, but it provides a fundamentally different guarantee. Utility is city-intrinsic, while substitution depends on ecosystem coverage. PIER directly targets that coverage by ordering models according to convex-hull novelty, which is why it produces a clean governance curve that supports statements of the form: "we can remove a large fraction of models while staying within a fixed MAE budget."

Finally, Fig. 5E shows why we need a geometric notion of uniqueness even when a utility heuristic seems close on average. We report the *excess degradation* beyond a 1% local-MAE band and track both the worst-case city and the 80th-percentile tail as pruning proceeds. The worst-case curve exhibits step-like jumps that reveal a heavy-tailed substitution-risk landscape. Utility-only pruning hits a high-penalty city early, producing an immediate large worst-case

jump, while PIER-guided pruning delays this first catastrophic event to a much later pruning fraction. This behavior is consistent with the convex-hull interpretation: high-PIER cities sit on extreme faces that the ecosystem cannot reconstruct, so pruning them early exposes the system to abrupt SLA violations. By contrast, PIER-guided consolidation preferentially removes interior, well-covered cities first, which makes the tail-risk profile substantially more stable.

Taken together, Fig. 5 demonstrates how DISCO turns a heterogeneous set of deployed predictors into a coherent geometric object. We use that geometry to (i) decouple uniqueness from utility, (ii) avoid false uniqueness from pairwise proxies, (iii) produce interpretable substitution backbones via $\hat{w}$, and (iv) deliver governance curves that quantify consolidation budgets and tail-risk tradeoffs in a directly actionable way.

Discussion

We introduced a statistical framework for quantifying model uniqueness in heterogeneous AI ecosystems via in-silico quasi-experimental design. By separating intrinsic model identity (Type-A treatment) from extrinsic input interventions (Type-B treatment), and by applying matched Type-B patterns across all models, we obtain a convex peer expressivity class and a residual notion of peer-inexpressible behavior. This yields a population uniqueness functional (PIER) with a direct operational meaning: whether a target can be substituted by a probabilistic routing policy over its peers. Under mild regularity conditions, PIER can be estimated reliably by the DISCO estimator.

On the statistical side, the passive theory establishes that, under fixed designs, DISCO behaves like a standard projection-based estimator: the projection weights and uniqueness converge to their population counterparts and admit asymptotic normal approximations. Finite-sample bounds further clarify how design choices (e.g., anchor radius and dose-grid spacing) control estimation error, supporting the use of PIER as a calibrated audit statistic rather than an arbitrary score.

A second advantage of in-silico design is algorithmic: when the auditor controls interventions, auditing can be sample-efficient. In a local linear structural model, a small set of carefully chosen design points suffices for exact uniqueness detection in the noiseless regime. With noise, repeated queries at these points yield explicit non-asymptotic sample-complexity guarantees whose dependence on the noise level, uniqueness margin, and confidence is minimax optimal. These results characterize when uniqueness auditing is both principled and efficient.

Several structural properties make the convex formulation conservative yet actionable for ecosystem governance. Monotonicity in the peer set connects PIER to pruning and consolidation decisions, while the routing interpretation of projection weights provides an explicit, auditable substitution certificate. At the same time, our non-identifiability results and the separation between PIER and Shapley-based attribution emphasize that uniqueness auditing is distinct from credit assignment and that matched intervention control is essential.

Empirically, simulations further suggest an "ecosystem saturation" effect, where average uniqueness decreases rapidly as the peer set grows and then exhibits diminishing returns. We conjecture that analogous phase-transition behavior persists under suitable sampling assumptions on low-dimensional task manifolds; making this precise and deriving dimension-dependent rates is an interesting direction for future work.

From a practical standpoint, the framework suggests a two-stage audit pipeline in large model ecosystems: a coarse screening stage to identify low-uniqueness candidates, followed by targeted local audits and robustness analyses of the PIER trajectory for models of interest. Future work includes incorporating richer structural priors in the expressivity class (while retaining conservative aggregation), developing adaptive context selection strategies that focus audit budget on consequential regions of the intervention space, and connecting uniqueness capacity to downstream trustworthy-AI notions such as robustness, privacy, and policy constraints. Applying the framework to industrial-scale ecosystems of large language models [1], vision models, and multimodal systems will require careful engineering of scalarization maps, intervention families, and approximate inference under resource constraints.

Methods

This section provides the formal definitions, assumptions, and estimation procedures underlying the Results. To limit repetition, we keep the introductory setup in the Results section and focus here on the technical construction of PIER/uniqueness, the DISCO estimator, and the active-auditing protocol.

Model Ecosystems and Scalarised Responses

We work with the model-ecosystem setup introduced in the Results section (subsection "ISQED Framework for AI Ecosystems") and only recall the notation that is used repeatedly below. For each model $j \in \mathcal{J}$ and input–dose pair

$(x, \theta) \in \mathcal{X} \times \Theta$, let

$$Y_j(x, \theta) \in \mathbb{R}$$

denote the *scalarised response* of model j under the intervention indexed by θ. Fix a target index $t \in \mathcal{J}$ and let $\mathcal{J}_{-t} = \mathcal{J} \setminus \{t\}$ be its peer set. We write

$$\Phi_{-t}(x, \theta) = (Y_j(x, \theta))_{j \in \mathcal{J}_{-t}} \in \mathbb{R}^{|\mathcal{J}_{-t}|}$$

for the vector of peer responses at (x, θ).

In-Silico Quasi-Experimental Design

We use the in-silico quasi-experimental design (ISQED) framework described in the Results section. A *context* variable C takes values in a context space $\mathcal{C}_{\mathrm{ctx}}$; throughout we condition on a fixed realised context value and suppress it in the notation. Conditional on this context, the auditor specifies two designs on $\mathcal{X} \times \Theta$: the *fitting design* $\mathbb{P}_{\mathrm{fit}}$ (used to learn a convex peer surrogate) and the *evaluation design* $\mathbb{P}_{\mathrm{eval}}$ (used to assess PIER and uniqueness).

The honesty condition is enforced by an explicit sample split. We draw a fitting sample $(X_i^{\mathrm{fit}}, \vartheta_i^{\mathrm{fit}})_{i=1}^m$ i.i.d. from $\mathbb{P}_{\mathrm{fit}}$ and an evaluation sample $(X_i^{\mathrm{eval}}, \vartheta_i^{\mathrm{eval}})_{i=1}^n$ i.i.d. from $\mathbb{P}_{\mathrm{eval}}$, with the two samples independent. This ensures that the data used to estimate projection weights are not reused when reporting PIER or uniqueness.

Peer Expressivity, Population PIER and Uniqueness

For the fixed context and fitting design $\mathbb{P}_{\mathrm{fit}}$, consider the Hilbert space $L^2(\mathbb{P}_{\mathrm{fit}})$ of measurable functions $h : \mathcal{X} \times \Theta \to \mathbb{R}$ with finite second moment under $\mathbb{P}_{\mathrm{fit}}$. The inner product is

$$\langle h_1, h_2 \rangle = \mathbb{E}_{\mathbb{P}_{\mathrm{fit}}} \big[h_1(X, \vartheta) h_2(X, \vartheta) \big],$$

with induced norm $\|h\|_2 = \sqrt{\langle h, h \rangle}$.

For a weight vector $w \in \mathbb{R}^{|\mathcal{J}_{-t}|}$, define the peer combination function

$$h_w(x, \theta) = w^\top \Phi_{-t}(x, \theta).$$

The convex peer expressivity class is

$$\mathcal{C} = \{ h_w : w \in \Delta^{|\mathcal{J}_{-t}|-1} \} \subset L^2(\mathbb{P}_{\mathrm{fit}}),$$

where $\Delta^{|\mathcal{J}_{-t}|-1}$ is the simplex of non-negative vectors in $\mathbb{R}^{|\mathcal{J}_{-t}|}$ that sum to one. This class collects all behaviours reachable by stochastic routing over peers with a context-independent routing distribution.

The population residual of a weight vector w is the mean-squared deviation between the target and the convex peer combination,

$$L(w) = \mathbb{E}_{\mathbb{P}_{\mathrm{fit}}} \big[(Y_t(X, \vartheta) - w^\top \Phi_{-t}(X, \vartheta))^2 \big].$$

Under an identifiability condition, the residual function L has a unique minimiser w^* in the relative interior of $\Delta^{|\mathcal{J}_{-t}|-1}$, and the residual is locally strongly convex around w^*. The function h_{w^*} is the $L^2(\mathbb{P}_{\mathrm{fit}})$-projection of Y_t onto the closure of $\mathcal{C}$.

The population peer-inexpressible residual (PIER) is defined pointwise by

$$R_t(x, \theta) = Y_t(x, \theta) - w^{*\top} \Phi_{-t}(x, \theta).$$

By construction, $R_t \in L^2(\mathbb{P}_{\mathrm{fit}})$ and is orthogonal to the closure of $\mathcal{C}$ in $L^2(\mathbb{P}_{\mathrm{fit}})$.

To aggregate PIER into a scalar measure, an evaluation design $\mathbb{P}_{\mathrm{eval}}$ on $\mathcal{X} \times \Theta$ is introduced. Under an integrability condition ensuring that $\mathbb{E}_{\mathbb{P}_{\mathrm{eval}}}[|R_t(X, \vartheta)|] < \infty$, the population uniqueness functional is defined as

$$\mathcal{U}_t = \mathbb{E}_{\mathbb{P}_{\mathrm{eval}}} \big[|R_t(X, \vartheta)| \big],$$

where $(X, \vartheta) \sim \mathbb{P}_{\mathrm{eval}}$. The functional $\mathcal{U}_t$ depends on the target, the peer set, the fitting design via w^*, and the evaluation design.

The DISCO Estimator

The DISCO estimator implements empirical counterparts of w^*, R_t and $\mathcal{U}_t$ using finite fitting and evaluation samples. For the fixed context, a fitting sample $(X_i^{\text{fit}}, \vartheta_i^{\text{fit}})_{i=1}^m$ is drawn independently from $\mathbb{P}_{\text{fit}}$. For each model j, the scalarised responses

$$Y_{j,i}^{\text{fit}} = Y_j(X_i^{\text{fit}}, \vartheta_i^{\text{fit}})$$

are computed. The peer responses are assembled into vectors

$$\Phi_{-t,i}^{\text{fit}} = \Phi_{-t}(X_i^{\text{fit}}, \vartheta_i^{\text{fit}}),$$

which are stacked row-wise into a design matrix $\mathsf{X}_{-t}^{\text{fit}} \in \mathbb{R}^{m \times |\mathcal{J}_{-t}|}$. The corresponding target responses form a vector $y_t^{\text{fit}} \in \mathbb{R}^m$ with entries $(y_t^{\text{fit}})_i = Y_{t,i}^{\text{fit}}$.

For a regularisation parameter $\lambda_m \geq 0$, the empirical projection weights are defined as any measurable solution to the convex optimisation problem

$$\widehat{w} \in \operatorname*{arg\,min}_{w \in \Delta^{|\mathcal{J}_{-t}|-1}} \left\{ \frac{1}{m} \left\| y_t^{\text{fit}} - \mathsf{X}_{-t}^{\text{fit}} w \right\|_2^2 + \lambda_m \|w\|_2^2 \right\}.$$

The regularisation term stabilises the estimator in finite samples and is chosen so that $\lambda_m \to 0$ and $m\lambda_m \to 0$ as m increases for the asymptotic theory.

An independent evaluation sample $(X_i^{\text{eval}}, \vartheta_i^{\text{eval}})_{i=1}^n$ is drawn from $\mathbb{P}_{\text{eval}}$. For each model j,

$$Y_{j,i}^{\text{eval}} = Y_j(X_i^{\text{eval}}, \vartheta_i^{\text{eval}}), \qquad \Phi_{-t,i}^{\text{eval}} = \Phi_{-t}(X_i^{\text{eval}}, \vartheta_i^{\text{eval}})$$

are computed. The empirical PIER at the i-th evaluation point is

$$\widehat{R}_{ti} = Y_{t,i}^{\text{eval}} - \widehat{w}^\top \Phi_{-t,i}^{\text{eval}}.$$

The empirical uniqueness functional is the average absolute PIER,

$$\widehat{\mathcal{U}}_t = \frac{1}{n} \sum_{i=1}^n |\widehat{R}_{ti}|.$$

In practice, evaluation points are often associated with nearby anchor points in the input–dose space, and responses at anchors are used as proxies for responses at evaluation points. Under Lipschitz regularity of the scalarised responses, Appendix A.3 bounds the resulting additional error in terms of the anchor radius and the dose-grid spacing.

Local Linear Structural Model and Active Auditing Protocol

For active auditing analyses, a local linear structural model is considered within a fixed context. A feature map $\phi : \mathcal{X} \times \Theta \to \mathbb{R}^d$ of dimension $d \geq 1$ is assumed known, and each scalarised response admits a linear parametrisation,

$$Y_j(x, \theta) = \phi(x, \theta)^\top \beta_j,$$

for some coefficient vector $\beta_j \in \mathbb{R}^d$. The set of peer coefficients is

$$\mathcal{B}_{\text{peer}} = \{\beta_j : j \in \mathcal{J}_{-t}\} \subset \mathbb{R}^d,$$

and their convex hull is denoted $\mathcal{H}_{\text{peer}}$. The target coefficient vector is $\beta_t \in \mathbb{R}^d$.

The uniqueness margin is defined as the Euclidean distance from the target coefficient to the convex hull of peer coefficients,

$$\gamma = \inf_{w \in \Delta^{|\mathcal{J}_{-t}|-1}} \left\| \beta_t - \sum_{j \in \mathcal{J}_{-t}} w_j \beta_j \right\|_2.$$

The target is non-unique in this model if and only if $\gamma = 0$.

To perform active auditing, the auditor chooses a set of d design points (x_ℓ, θ_ℓ) such that the resulting feature matrix

$$\Phi = \begin{bmatrix} \phi(x_1, \theta_1)^\top \\ \vdots \\ \phi(x_d, \theta_d)^\top \end{bmatrix} \in \mathbb{R}^{d \times d}$$

is invertible. In a noiseless regime, querying each model once at these design points yields response vectors y_j satisfying $y_j = \Phi \beta_j$, so that $\beta_j = \Phi^{-1} y_j$ for all j. Uniqueness reduces to a convex feasibility problem in coefficient space.

In a noisy regime, each model evaluation at a design point (x_ℓ, θ_ℓ) yields a noisy observation

$$\widetilde{Y}_j(x_\ell, \theta_\ell) = Y_j(x_\ell, \theta_\ell) + \varepsilon_j(x_\ell, \theta_\ell),$$

where the noise variables are independent across models, design points and repetitions, have zero mean and are sub-Gaussian with variance proxy σ^2. The auditor queries each model r times at each design point and averages the responses to form

$$\bar{y}_{j,\ell} = \frac{1}{r} \sum_{s=1}^{r} \widetilde{Y}_j^{(s)}(x_\ell, \theta_\ell).$$

The averaged response vector $\bar{y}_j$ is then used to estimate the coefficients by $\widehat{\beta}_j = \Phi^{-1} \bar{y}_j$. The empirical convex hull of peer coefficients is formed as $\mathrm{conv}\{\widehat{\beta}_j : j \in \mathcal{J}_{-t}\}$, and the empirical distance from $\widehat{\beta}_t$ to this hull is compared to a threshold based on a presumed margin γ.

Experimental Protocols

Synthetic experiments can be designed to validate the theoretical predictions. In one class of experiments, an input space is sampled from a low-dimensional Euclidean distribution and synthetic scalarised responses are generated according to prescribed coefficient vectors and feature maps, with additive noise. DISCO is applied with various fitting and evaluation sample sizes, anchor radii and dose-grid resolutions, and empirical behaviour of $\widehat{w}$, $\widehat{\mathcal{U}}_t$ and anchor-based PIER estimates is compared to Monte Carlo approximations of population quantities. In another class of experiments, the local linear structural model is enforced exactly by choosing ϕ and β_j directly, and the active auditing protocol is used to recover coefficients and test uniqueness under varying noise levels and repetition counts. These setups allow empirical examination of convergence, normal approximation, design-error scaling and the sample-complexity bounds derived in the Supplementary Information.

References

[1] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2022.

[2] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. volume 30, 2017.

[3] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *stat*, 1050:9, 2015.

[4] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR, 2019.

[5] Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for deep learning in nlp. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 3645–3650, 2019.

[6] Florian Tramèr, Fan Zhang, Ari Juels, Michael K Reiter, and Thomas Ristenpart. Stealing machine learning models via prediction API. In *25th USENIX security symposium (USENIX Security 16)*, pages 601–618, 2016.

[7] Yusuke Uchida, Yuki Nagai, Shigeyuki Sakazawa, and Shin'ichi Satoh. Embedding watermarks into deep neural networks. In *Proceedings of the 2017 ACM on international conference on multimedia retrieval*, pages 269–277, 2017.

[8] Bita Darvish Rouhani, Huili Chen, and Farinaz Koushanfar. Deepsigns: An end-to-end watermarking framework for ownership protection of deep neural networks. In *Proceedings of the twenty-fourth international conference on architectural support for programming languages and operating systems*, pages 485–497, 2019.

[9] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[10] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017.

[11] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30, pages 4765–4774, 2017.

[12] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.*, 31:841, 2017.

[13] Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. *Advances in neural information processing systems*, 30, 2017.

[14] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International conference on machine learning*, pages 3519–3529. PMlR, 2019.

[15] LS Shapley. A value for n-person games'. *Contributions to the Theory of Games*, (28):307, 1953.

[16] Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, et al. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *ICML*, pages 23965–23998. PMLR, 2022.

[17] Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Cambridge, UK, 2 edition, 2009.

[18] Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge university press, Cambridge, UK, 2015.

[19] Donald T Campbell and Julian C Stanley. *Experimental and quasi-experimental designs for research*. Ravenio books, County Dublin, Ireland, 2015.

[20] Joshua D Angrist and Jörn-Steffen Pischke. *Mostly harmless econometrics: An empiricist's companion*. Princeton university press, Princeton, USA, 2009.

[21] Alberto Abadie, Alexis Diamond, and Jens Hainmueller. Synthetic control methods for comparative case studies: Estimating the effect of california's tobacco control program. *Journal of the American statistical Association*, 105(490):493–505, 2010.

[22] Burr Settles. Active learning literature survey. Technical Report 1648, University of Wisconsin–Madison, 2009.

[23] Friedrich Pukelsheim. *Optimal Design of Experiments*. SIAM, PA 19104 USA, 2006.

[24] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. volume 31, 2018.

[25] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642, 2013.

[26] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP workshop BlackboxNLP: Analyzing and interpreting neural networks for NLP*, pages 353–355, 2018.

[27] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

[28] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

[29] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. ALBERT: A lite BERT for self-supervised learning of language representations. In *International Conference on Learning Representations*, 2020.

[30] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[31] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. XLNet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[32] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[33] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[34] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

[35] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022.

[36] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[37] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

[38] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *International conference on learning representations*, 2018.

[39] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *stat*, 1050:20, 2015.

[40] Aad W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, Cambridge, UK, 2000.

Appendix: Quantifying Model Uniqueness in Heterogeneous AI Ecosystems

Appendix A   Passive Theory for DISCO

This note provides detailed proofs of the passive statistical guarantees and structural properties of the DISCO estimator described in the main text. Throughout, the context value is fixed and suppressed from the notation. Expectations and probabilities are conditional on this context.

A.1   Consistency of DISCO Under Fixed Designs

We first establish that, under mild assumptions, the empirical projection weights $\widehat{w}$ converge in probability to the population projection weights w^* and the empirical uniqueness functional $\widehat{\mathcal{U}}_t$ converges in probability to the population uniqueness $\mathcal{U}_t$ as the sizes of the fitting and evaluation samples increase.

Proposition 1 (Consistency under fixed designs). Suppose that the residual function $L(w) = \mathbb{E}_{\mathbb{P}_{\text{fit}}}[(Y_t(X,\vartheta) - w^\top \Phi_{-t}(X,\vartheta))^2]$ has a unique minimiser w^* in the relative interior of the simplex $\Delta^{|\mathcal{J}_{-t}|-1}$ and that L is locally strongly convex around w^*. Assume that the response functions are bounded and continuous, that $\mathbb{P}_{\text{fit}}$ and $\mathbb{P}_{\text{eval}}$ have full support on a compact subset of $\mathcal{X} \times \Theta$ and are absolutely continuous with respect to the corresponding marginal of $\mathbb{P}_{\text{design}}$, that the honesty condition holds, and that the regularisation parameter in the DISCO optimisation satisfies $\lambda_m \to 0$ and $m\lambda_m \to 0$ as $m \to \infty$. Then $\widehat{w} \xrightarrow{p} w^*$ as $m \to \infty$, and, as $m, n \to \infty$,

$$\widehat{\mathcal{U}}_t \xrightarrow{p} \mathcal{U}_t.$$

Proof. The proof proceeds by first showing uniform convergence of the empirical residual to the population residual on the simplex, then invoking an argmin theorem for M-estimators, and finally transferring this convergence to $\widehat{\mathcal{U}}_t$ via continuity and a law of large numbers.

For each $w \in \Delta^{|\mathcal{J}_{-t}|-1}$, define the unregularised empirical residual

$$\widetilde{L}_m(w) = \frac{1}{m} \sum_{i=1}^{m} \left(Y_{t,i}^{\text{fit}} - w^\top \Phi_{-t,i}^{\text{fit}} \right)^2.$$

By boundedness of the scalarised responses, there exists a constant B such that $|Y_j(x,\theta)| \leq B$ for all (x,θ) and all j. Hence, for each fixed $w \in \Delta^{|\mathcal{J}_{-t}|-1}$, the random variables $(Y_{t,i}^{\text{fit}} - w^\top \Phi_{-t,i}^{\text{fit}})^2$ are independent and identically distributed with finite expectation under $\mathbb{P}_{\text{fit}}$, and the strong law of large numbers implies that $\widetilde{L}_m(w)$ converges almost surely to $L(w)$ as $m \to \infty$.

To control the convergence uniformly over $w \in \Delta^{|\mathcal{J}_{-t}|-1}$, note that for each i the map $w \mapsto (Y_{t,i}^{\text{fit}} - w^\top \Phi_{-t,i}^{\text{fit}})^2$ is a continuous function on the compact set $\Delta^{|\mathcal{J}_{-t}|-1}$. The family of functions $\{(Y_{t,i}^{\text{fit}} - w^\top \Phi_{-t,i}^{\text{fit}})^2 : w \in \Delta^{|\mathcal{J}_{-t}|-1}\}$ is dominated by the integrable envelope $(2B + 2B\sqrt{|\mathcal{J}_{-t}|})^2$, which does not depend on w or i. A uniform law of large numbers for such parametric families yields that the supremum over $w \in \Delta^{|\mathcal{J}_{-t}|-1}$ of $|\widetilde{L}_m(w) - L(w)|$ converges to zero almost surely as $m \to \infty$.

The regularised empirical residual is defined by

$$\widehat{L}_m(w) = \widetilde{L}_m(w) + \lambda_m \|w\|_2^2.$$

Since the simplex is compact, there is a constant K such that $\|w\|_2^2 \leq K$ for all $w \in \Delta^{|\mathcal{J}_{-t}|-1}$. The condition $\lambda_m \to 0$ therefore implies that

$$\sup_{w \in \Delta^{|\mathcal{J}_{-t}|-1}} |\widehat{L}_m(w) - \widetilde{L}_m(w)| \leq \lambda_m K \to 0$$

as $m \to \infty$. Combining this with the uniform convergence of $\widetilde{L}_m$ to L shows that $\widehat{L}_m$ converges to L uniformly on $\Delta^{|\mathcal{J}_{-t}|-1}$ in probability. In particular, for any $\varepsilon > 0$ and any neighbourhood U of w^*, there exists m_0 such that for $m \geq m_0$ the set of approximate minimisers of $\widehat{L}_m$ lies inside U with high probability.

The residual function L is assumed to have a unique minimiser w^* in the relative interior of $\Delta^{|\mathcal{J}_{-t}|-1}$ and to be locally strongly convex around w^*, so that $|L(w) - L(w^*)| \geq \kappa \|w - w^*\|_2^2$ for some $\kappa > 0$ and all w in a neighbourhood of w^*. Standard argmin theorems for M-estimators (see, for example, [40]) then yield that any measurable selection of minimisers of $\widehat{L}_m$ converges in probability to w^*. The empirical projection weights $\widehat{w}$ are such a selection, hence $\widehat{w} \xrightarrow{p} w^*$ as $m \to \infty$.

To pass from $\widehat{w}$ to $\widehat{\mathcal{U}}_t$, introduce the function

$$\psi(x,\theta,w) = |Y_t(x,\theta) - w^\top \Phi_{-t}(x,\theta)|.$$

For each evaluation pair $(X_i^{\text{eval}}, \vartheta_i^{\text{eval}})$, the empirical uniqueness estimator can be written as

$$\widehat{\mathcal{U}}_t = \frac{1}{n} \sum_{i=1}^n \psi(X_i^{\text{eval}}, \vartheta_i^{\text{eval}}, \widehat{w}).$$

The map $w \mapsto \psi(X_i^{\text{eval}}, \vartheta_i^{\text{eval}}, w)$ is continuous because it is the absolute value of an affine function in w. Since $\widehat{w} \xrightarrow{p} w^*$, the continuous mapping theorem implies that for each fixed i the random quantity $\psi(X_i^{\text{eval}}, \vartheta_i^{\text{eval}}, \widehat{w})$ converges in probability to $\psi(X_i^{\text{eval}}, \vartheta_i^{\text{eval}}, w^*)$ as $m \to \infty$.

Boundedness of responses ensures that there exists a constant M such that $|\psi(X_i^{\text{eval}}, \vartheta_i^{\text{eval}}, w)| \leq M$ almost surely for all $w \in \Delta^{|\mathcal{J}_{-t}|-1}$. The difference $|\psi(X_i^{\text{eval}}, \vartheta_i^{\text{eval}}, \widehat{w}) - \psi(X_i^{\text{eval}}, \vartheta_i^{\text{eval}}, w^*)|$ is therefore bounded by $2M$. Dominated convergence then shows that the expectation of this difference converges to zero as $(m,n) \to \infty$.

Decomposing $\widehat{\mathcal{U}}_t - \mathcal{U}_t$ into the sum of two terms,

$$\widehat{\mathcal{U}}_t - \mathcal{U}_t = A_n + B_n,$$

where A_n is the average of $\psi(X_i^{\text{eval}}, \vartheta_i^{\text{eval}}, \widehat{w}) - \psi(X_i^{\text{eval}}, \vartheta_i^{\text{eval}}, w^*)$ and B_n is the average of $\psi(X_i^{\text{eval}}, \vartheta_i^{\text{eval}}, w^*) - \mathcal{U}_t$, one observes that B_n converges almost surely to zero by the law of large numbers applied under $\mathbb{P}_{\text{eval}}$, and that the expectation of $|A_n|$ converges to zero by Jensen's inequality and dominated convergence. Markov's inequality then implies that $A_n \xrightarrow{p} 0$. Slutsky's theorem applied to $A_n + B_n$ yields $\widehat{\mathcal{U}}_t \xrightarrow{p} \mathcal{U}_t$. This completes the proof of Proposition 1. $\qquad \square$

A.2 Asymptotic Normality of the Uniqueness Estimator

We next establish a central limit theorem for $\widehat{\mathcal{U}}_t$ under additional differentiability assumptions.

Proposition 2 (Asymptotic normality under fixed designs). Under the conditions of Proposition 1, suppose in addition that there exists a neighbourhood U of w^* in $\Delta^{|\mathcal{J}_{-t}|-1}$ such that, for each (x, θ), the map $w \mapsto |Y_t(x, \theta) - w^\top \Phi_{-t}(x, \theta)|$ is differentiable on U with gradient that is continuous in w, and that the squared norm of this gradient has finite expectation under $\mathbb{P}_{\text{eval}}$ for all $w \in U$. If $m, n \to \infty$ with $m/n \to \rho \in (0, \infty)$, then

$$\sqrt{n}(\widehat{\mathcal{U}}_t - \mathcal{U}_t) \rightsquigarrow \mathcal{N}(0, \sigma^2)$$

for some finite variance $\sigma^2 > 0$.

Proof. The proof combines the asymptotic normality of the M-estimator $\widehat{w}$ with a delta-method argument for the plug-in estimator $\widehat{\mathcal{U}}_t$. Under the regularity and identifiability assumptions, and the rate condition on λ_m, standard theory for M-estimators implies that

$$\sqrt{m}(\widehat{w} - w^*) \rightsquigarrow \mathcal{N}(0, \Sigma_w)$$

for some positive-definite covariance matrix Σ_w. This results from a second-order expansion of the empirical residual around w^* and an application of a multivariate central limit theorem to the empirical gradient.

Define the functional

$$\Gamma(w) = \mathbb{E}_{\mathbb{P}_{\text{eval}}}\big[\psi(X, \vartheta, w)\big] = \mathbb{E}_{\mathbb{P}_{\text{eval}}}\big[|Y_t(X, \vartheta) - w^\top \Phi_{-t}(X, \vartheta)|\big].$$

By the assumed differentiability and boundedness of the gradient, Γ is differentiable in a neighbourhood of w^* with gradient

$$\nabla\Gamma(w^*) = \mathbb{E}_{\mathbb{P}_{\text{eval}}}\big[\nabla_w \psi(X, \vartheta, w^*)\big].$$

For any fixed w, an application of the central limit theorem to the average of $\psi(X_i^{\text{eval}}, \vartheta_i^{\text{eval}}, w)$ yields that

$$\sqrt{n}\big(\widehat{\mathcal{U}}_t(w) - \Gamma(w)\big) \rightsquigarrow \mathcal{N}(0, \sigma^2(w)),$$

where $\widehat{\mathcal{U}}_t(w)$ denotes the empirical average of ψ evaluated at w and $\sigma^2(w)$ is the variance of $\psi(X, \vartheta, w)$ under $\mathbb{P}_{\text{eval}}$.

Returning to $\widehat{\mathcal{U}}_t = \widehat{\mathcal{U}}_t(\widehat{w})$, decompose

$$\widehat{\mathcal{U}}_t - \mathcal{U}_t = \big(\widehat{\mathcal{U}}_t(\widehat{w}) - \Gamma(\widehat{w})\big) + \big(\Gamma(\widehat{w}) - \Gamma(w^*)\big).$$

Conditionally on the fitting sample, the evaluation sample is independent and identically distributed from $\mathbb{P}_{\text{eval}}$, and the conditional central limit theorem implies that the first term, scaled by $\sqrt{n}$, converges in distribution to a centred Gaussian with variance $\sigma^2(w^*)$. This is because $\widehat{w} \xrightarrow{p} w^*$ and $\sigma^2(\widehat{w})$ converges to $\sigma^2(w^*)$ by continuity.

To control the second term, expand Γ in a first-order Taylor expansion around w^*,

$$\Gamma(\widehat{w}) - \Gamma(w^*) = \nabla\Gamma(\tilde{w}_m)^\top(\widehat{w} - w^*),$$

where $\tilde{w}_m$ lies on the line segment between $\widehat{w}$ and w^*. As $\widehat{w} \xrightarrow{p} w^*$, the continuity of $\nabla\Gamma$ implies that $\nabla\Gamma(\tilde{w}_m) \xrightarrow{p} \nabla\Gamma(w^*)$. Combining this with the asymptotic normality of $\sqrt{m}(\widehat{w} - w^*)$ shows that

$$\sqrt{n}\big(\Gamma(\widehat{w}) - \Gamma(w^*)\big) = \sqrt{n}\,\nabla\Gamma(w^*)^\top(\widehat{w} - w^*) + o_p(1).$$

Since $m/n \to \rho$, the term $\sqrt{n}(\widehat{w} - w^*)$ behaves like $\sqrt{n/m}$ times a Gaussian vector, leading to a Gaussian limit for this component as well with variance matrix given by $\rho^{-1}\,\nabla\Gamma(w^*)^\top\Sigma_w\nabla\Gamma(w^*)$.

The two leading terms in the decomposition of $\sqrt{n}(\widehat{\mathcal{U}}_t - \mathcal{U}_t)$ come from independent sources: the evaluation sample and the fitting sample. The honesty assumption ensures their independence asymptotically. Each converges in distribution to a centred Gaussian variable, and their sum is therefore Gaussian with variance equal to the sum of the variances. The variance of the limiting Gaussian is finite by the boundedness and integrability assumptions. This establishes the claimed central limit theorem. □

A.3 Finite-Sample Design Error Bound

We now provide a bound on the error incurred when anchor points and finite dose grids are used to approximate PIER at evaluation points.

Proposition 3 (Finite-sample design error bound). Assume that the spaces $\mathcal{X}$ and Θ are compact metric spaces, that the intervention map and response functions are continuous, that each scalarised response Y_j is Lipschitz in (x, θ) with Lipschitz constants L_x and L_θ in the input and dose coordinates respectively, and that the peer response vectors are uniformly bounded in norm by a constant C_Φ. Let (x^*, θ^*) be an evaluation point and $(\tilde{x}, \tilde{\theta})$ be an anchor point satisfying $\|x^* - \tilde{x}\| \leq r_x$ and $|\theta^* - \tilde{\theta}| \leq r_\theta$. Define the anchor-based empirical PIER

$$\widehat{R}_t^{\mathrm{anchor}}(x^*, \theta^*) = Y_t(\tilde{x}, \tilde{\theta}) - \widehat{w}^\top \Phi_{-t}(\tilde{x}, \tilde{\theta}).$$

Then

$$\left| \widehat{R}_t^{\mathrm{anchor}}(x^*, \theta^*) - R_t(x^*, \theta^*) \right| \leq C_\Phi \|\widehat{w} - w^*\|_2 + 2L_x r_x + 2L_\theta r_\theta.$$

Proof. By definition of $\widehat{R}_t^{\mathrm{anchor}}$ and R_t, the difference can be expressed as

$$\widehat{R}_t^{\mathrm{anchor}}(x^*, \theta^*) - R_t(x^*, \theta^*) = A_1 + A_2 + A_3,$$

where

$$A_1 = (w^* - \widehat{w})^\top \Phi_{-t}(\tilde{x}, \tilde{\theta}),$$
$$A_2 = Y_t(\tilde{x}, \tilde{\theta}) - Y_t(x^*, \theta^*),$$

and

$$A_3 = w^{*\top} \Phi_{-t}(x^*, \theta^*) - w^{*\top} \Phi_{-t}(\tilde{x}, \tilde{\theta}).$$

The term A_1 reflects the contribution of weight estimation error. By the Cauchy–Schwarz inequality and the uniform bound on peer response norms, one has

$$|A_1| \leq \|w^* - \widehat{w}\|_2 \, \|\Phi_{-t}(\tilde{x}, \tilde{\theta})\|_2 \leq C_\Phi \|w^* - \widehat{w}\|_2.$$

The term A_2 measures the change in the target response between the anchor and the evaluation point. The Lipschitz assumption on Y_t implies that

$$|A_2| = |Y_t(\tilde{x}, \tilde{\theta}) - Y_t(x^*, \theta^*)| \leq L_x \|\tilde{x} - x^*\| + L_\theta |\tilde{\theta} - \theta^*| \leq L_x r_x + L_\theta r_\theta.$$

The term A_3 accounts for the change in peer responses between the anchor and the evaluation point, weighted by w^*. It can be written as

$$A_3 = \sum_{j \in \mathcal{J}_{-t}} w_j^* \big(Y_j(x^*, \theta^*) - Y_j(\tilde{x}, \tilde{\theta}) \big).$$

Each summand has absolute value bounded by $L_x r_x + L_\theta r_\theta$ by the Lipschitz property, and the weights are non-negative and sum to one. Therefore,

$$|A_3| \leq \sum_{j \in \mathcal{J}_{-t}} w_j^* (L_x r_x + L_\theta r_\theta) = L_x r_x + L_\theta r_\theta.$$

Applying the triangle inequality to $A_1 + A_2 + A_3$ and combining the bounds gives

$$\left| \widehat{R}_t^{\mathrm{anchor}}(x^*, \theta^*) - R_t(x^*, \theta^*) \right| \leq C_\Phi \|\widehat{w} - w^*\|_2 + 2L_x r_x + 2L_\theta r_\theta,$$

as claimed. $\qquad\square$

A.4 Monotonicity and Conservatism of the Uniqueness Functional

We now formalise two structural properties of the uniqueness functional described in the main text.

Proposition 4 (Monotonicity in the peer set). *Let $\mathcal{J}_1$ and $\mathcal{J}_2$ be peer sets such that $\mathcal{J}_1 \subseteq \mathcal{J}_2$. Let $\mathcal{U}_t(\mathcal{J}_1)$ and $\mathcal{U}_t(\mathcal{J}_2)$ denote the corresponding population uniqueness functionals defined with respect to the convex expressivity classes generated by $\mathcal{J}_1$ and $\mathcal{J}_2$ and the same fitting and evaluation designs. Then*

$$\mathcal{U}_t(\mathcal{J}_2) \leq \mathcal{U}_t(\mathcal{J}_1).$$

Proof. Let $\mathcal{C}(\mathcal{J}_1)$ and $\mathcal{C}(\mathcal{J}_2)$ denote the convex peer expressivity classes associated with the two peer sets, defined as sets of convex combinations of corresponding peer responses in $L^2(\mathbb{P}_{\text{fit}})$. The inclusion $\mathcal{J}_1 \subseteq \mathcal{J}_2$ implies that $\mathcal{C}(\mathcal{J}_1) \subseteq \mathcal{C}(\mathcal{J}_2)$. Let h_1^* and h_2^* be the orthogonal projections of Y_t onto the closures of $\mathcal{C}(\mathcal{J}_1)$ and $\mathcal{C}(\mathcal{J}_2)$ respectively, and let $R_t^{(1)} = Y_t - h_1^*$ and $R_t^{(2)} = Y_t - h_2^*$ be the corresponding residuals.

In a Hilbert space, projection onto a closed convex set minimises the norm of the residual over that set. Since $\mathcal{C}(\mathcal{J}_1)$ is contained in $\mathcal{C}(\mathcal{J}_2)$, the residual norm associated with $\mathcal{J}_2$ cannot exceed that associated with $\mathcal{J}_1$, that is

$$\|R_t^{(2)}\|_{L^2(\mathbb{P}_{\text{fit}})} \leq \|R_t^{(1)}\|_{L^2(\mathbb{P}_{\text{fit}})}.$$

The uniqueness functional is defined as the $L^1(\mathbb{P}_{\text{eval}})$ norm of the residual. Under the boundedness and absolute continuity assumptions on responses and designs, there exists a constant depending only on the support of $\mathbb{P}_{\text{eval}}$ such that the $L^1(\mathbb{P}_{\text{eval}})$ norm of any residual is controlled by a constant multiple of its $L^2(\mathbb{P}_{\text{fit}})$ norm. Applying this control to $R_t^{(1)}$ and $R_t^{(2)}$ yields

$$\mathcal{U}_t(\mathcal{J}_2) = \mathbb{E}_{\mathbb{P}_{\text{eval}}}[|R_t^{(2)}(X, \vartheta)|] \leq \mathbb{E}_{\mathbb{P}_{\text{eval}}}[|R_t^{(1)}(X, \vartheta)|] = \mathcal{U}_t(\mathcal{J}_1).$$

This proves monotonicity. □

Proposition 5 (Conservatism of convex peer aggregation). *Let $\mathcal{C}_{\text{conv}}$, $\mathcal{C}_{\text{lin}}$ and $\mathcal{C}_{\text{ker}}$ be three closed convex subsets of $L^2(\mathbb{P}_{\text{fit}})$ such that $\mathcal{C}_{\text{conv}} \subseteq \mathcal{C}_{\text{lin}} \subseteq \mathcal{C}_{\text{ker}}$ and all peer responses belong to these sets. Let R_t^{conv}, R_t^{lin} and R_t^{ker} denote the residuals of the orthogonal projections of Y_t onto these sets, and let $\mathcal{U}_t^{\text{conv}}$, $\mathcal{U}_t^{\text{lin}}$ and $\mathcal{U}_t^{\text{ker}}$ be the corresponding uniqueness functionals defined under the same evaluation design $\mathbb{P}_{\text{eval}}$. Then*

$$\mathcal{U}_t^{\text{ker}} \leq \mathcal{U}_t^{\text{lin}} \leq \mathcal{U}_t^{\text{conv}}.$$

Proof. Projection onto a larger closed convex set in a Hilbert space cannot increase the norm of the residual. The inclusions $\mathcal{C}_{\text{conv}} \subseteq \mathcal{C}_{\text{lin}} \subseteq \mathcal{C}_{\text{ker}}$ therefore imply

$$\|R_t^{\text{lin}}\|_{L^2(\mathbb{P}_{\text{fit}})} \leq \|R_t^{\text{conv}}\|_{L^2(\mathbb{P}_{\text{fit}})}, \qquad \|R_t^{\text{ker}}\|_{L^2(\mathbb{P}_{\text{fit}})} \leq \|R_t^{\text{lin}}\|_{L^2(\mathbb{P}_{\text{fit}})}.$$

As in the proof of Proposition 4, boundedness and absolute continuity allow us to transfer these $L^2(\mathbb{P}_{\text{fit}})$ inequalities to the corresponding $L^1(\mathbb{P}_{\text{eval}})$ norms. This yields

$$\mathcal{U}_t^{\text{ker}} = \mathbb{E}_{\mathbb{P}_{\text{eval}}}[|R_t^{\text{ker}}(X, \vartheta)|] \leq \mathbb{E}_{\mathbb{P}_{\text{eval}}}[|R_t^{\text{lin}}(X, \vartheta)|] = \mathcal{U}_t^{\text{lin}} \leq \mathbb{E}_{\mathbb{P}_{\text{eval}}}[|R_t^{\text{conv}}(X, \vartheta)|] = \mathcal{U}_t^{\text{conv}}.$$

This establishes the stated inequalities. □

Appendix B   Active Auditing Theory

This note develops the active auditing results summarised in the main text. We work within the local linear structural model described in the Methods.

B.1 Exact Uniqueness Detection in a Noiseless Regime

We first show that, in the noiseless local linear model, a finite number of carefully chosen design points suffices to recover all model coefficients exactly and to decide uniqueness.

Proposition 6 (Exact uniqueness detection with finitely many noiseless queries). Fix a context and suppose that each scalarised response satisfies $Y_j(x, \theta) = \phi(x, \theta)^\top \beta_j$ for some known feature map $\phi : \mathcal{X} \times \Theta \to \mathbb{R}^d$ and coefficient vectors $\beta_j \in \mathbb{R}^d$. Let $(x_\ell, \theta_\ell)_{\ell=1}^d$ be a set of design points such that the feature matrix

$$\Phi = \begin{bmatrix} \phi(x_1, \theta_1)^\top \\ \vdots \\ \phi(x_d, \theta_d)^\top \end{bmatrix}$$

is invertible. For each model $j \in \mathcal{J}$, evaluate Y_j at these points to obtain $y_{j,\ell} = Y_j(x_\ell, \theta_\ell)$, assemble the response vector $y_j = (y_{j,1}, \ldots, y_{j,d})^\top$, and compute $\widehat{\beta}_j = \Phi^{-1} y_j$. Then $\widehat{\beta}_j = \beta_j$ for all j. In particular, the target lies in the convex hull of peer coefficients if and only if there exists $w \in \Delta^{|\mathcal{J}_{-t}|-1}$ such that $\beta_t = \sum_{j \in \mathcal{J}_{-t}} w_j \beta_j$, and this condition can be checked exactly using finitely many noiseless queries per model.

Proof. For a fixed model j and design point index ℓ, the local linear model implies that

$$Y_j(x_\ell, \theta_\ell) = \phi(x_\ell, \theta_\ell)^\top \beta_j.$$

Collecting these equalities for $\ell = 1, \ldots, d$ in vector form yields $y_j = \Phi \beta_j$. The matrix Φ is invertible by assumption, so left-multiplying by Φ^{-1} gives $\widehat{\beta}_j = \Phi^{-1} y_j = \beta_j$. This holds for each $j \in \mathcal{J}$.

The uniqueness margin in coefficient space is defined as the distance from β_t to the convex hull of $\{\beta_j : j \in \mathcal{J}_{-t}\}$. In the noiseless setting, we have access to all coefficients β_j exactly via the reconstruction $\widehat{\beta}_j$. The condition $\gamma = 0$ is equivalent to the existence of a weight vector $w \in \Delta^{|\mathcal{J}_{-t}|-1}$ such that $\beta_t = \sum_{j \in \mathcal{J}_{-t}} w_j \beta_j$. This is a linear equality constraint together with simplex constraints on w and can be checked via a convex feasibility problem. Therefore, exactly d noiseless queries per model suffice to determine whether the uniqueness margin is zero or strictly positive. $\square$

B.2 Sample Complexity of Active Auditing with Noise

We now add noise to model evaluations and derive a non-asymptotic sample-complexity bound for active uniqueness detection.

Proposition 7 (Sample complexity of active auditing in the noisy linear model). Fix a context and suppose that $Y_j(x, \theta) = \phi(x, \theta)^\top \beta_j$ for a known feature map $\phi : \mathcal{X} \times \Theta \to \mathbb{R}^d$ and coefficients $\beta_j \in \mathbb{R}^d$. Let $(x_\ell, \theta_\ell)_{\ell=1}^d$ be a set of design points such that the feature matrix Φ is invertible. Assume that, when model j is queried at (x_ℓ, θ_ℓ), the auditor observes

$$\widetilde{Y}_j(x_\ell, \theta_\ell) = Y_j(x_\ell, \theta_\ell) + \varepsilon_j(x_\ell, \theta_\ell),$$

where the noise variables are independent across models, design points and repetitions, have zero mean and are sub-Gaussian with variance proxy σ^2. Suppose that the uniqueness margin

$$\gamma = \inf_{w \in \Delta^{|\mathcal{J}_{-t}|-1}} \left\| \beta_t - \sum_{j \in \mathcal{J}_{-t}} w_j \beta_j \right\|_2$$

is strictly positive. For each model and each design point, perform r independent noisy evaluations, compute the average response at each design point, and reconstruct coefficient estimates $\widehat{\beta}_j = \Phi^{-1} \bar{y}_j$ as described in the Methods. Form the empirical convex hull of peer coefficients and the empirical distance

$$\widehat{\text{dist}} = \inf_{w \in \Delta^{|\mathcal{J}_{-t}|-1}} \left\| \widehat{\beta}_t - \sum_{j \in \mathcal{J}_{-t}} w_j \widehat{\beta}_j \right\|_2.$$

There exists a constant $C > 0$, depending only on Φ, such that if

$$r \geq C \frac{\sigma^2}{\gamma^2} \log\left(\frac{Nd}{\delta}\right),$$

then the decision rule that declares the target non-unique when $\widehat{\text{dist}} \leq \gamma/2$ and unique when $\widehat{\text{dist}} > \gamma/2$ has misclassification probability at most δ.

Proof. For each model j and design point index ℓ, denote by $\widetilde{Y}_j^{(s)}(x_\ell, \theta_\ell)$ the s-th noisy evaluation at (x_ℓ, θ_ℓ) and define the sample average

$$\bar{y}_{j,\ell} = \frac{1}{r} \sum_{s=1}^{r} \widetilde{Y}_j^{(s)}(x_\ell, \theta_\ell).$$

The local linear model implies that $Y_j(x_\ell, \theta_\ell) = \phi(x_\ell, \theta_\ell)^\top \beta_j$, so the averaged response can be written as

$$\bar{y}_{j,\ell} = \phi(x_\ell, \theta_\ell)^\top \beta_j + \bar{\varepsilon}_{j,\ell},$$

where $\bar{\varepsilon}_{j,\ell} = r^{-1} \sum_{s=1}^{r} \varepsilon_j^{(s)}(x_\ell, \theta_\ell)$ is the average of r independent sub-Gaussian noise variables with variance proxy σ^2. Standard properties of sub-Gaussian variables state that $\bar{\varepsilon}_{j,\ell}$ is sub-Gaussian with variance proxy σ^2/r. Collecting the averages into a vector $\bar{\varepsilon}_j = (\bar{\varepsilon}_{j,1}, \ldots, \bar{\varepsilon}_{j,d})^\top$ and the responses into $\bar{y}_j$, one has

$$\bar{y}_j = \Phi \beta_j + \bar{\varepsilon}_j.$$

The coefficient estimate is

$$\widehat{\beta}_j = \Phi^{-1} \bar{y}_j = \beta_j + \Phi^{-1} \bar{\varepsilon}_j.$$

To control the estimation error, observe that the linear transformation Φ^{-1} maps the sub-Gaussian vector $\bar{\varepsilon}_j$ to a sub-Gaussian vector with a covariance proxy scaled by $\|\Phi^{-1}\|_{\text{op}}^2$. In particular, there exists a constant $K > 0$ depending only on Φ and d such that the Euclidean norm of $\Phi^{-1} \bar{\varepsilon}_j$ satisfies a tail bound of the form

$$\mathbb{P}\big(\|\widehat{\beta}_j - \beta_j\|_2 > u\big) \leq 2 \exp\left(-c_1 \frac{ru^2}{\sigma^2}\right)$$

for all $u > 0$, where c_1 is a positive constant independent of j, u and r. This follows from concentration inequalities for the Euclidean norm of sub-Gaussian vectors and from the bounded operator norm of Φ^{-1}.

Set $u = \gamma/8$. Then for each model j,

$$\mathbb{P}\big(\|\widehat{\beta}_j - \beta_j\|_2 > \gamma/8\big) \leq 2 \exp\left(-c_1 \frac{r\gamma^2}{64\sigma^2}\right).$$

Choose r such that the right-hand side is at most $\delta/(2Nd)$. This is achieved by requiring

$$r \geq C \frac{\sigma^2}{\gamma^2} \log\left(\frac{2Nd}{\delta}\right),$$

for a sufficiently large constant C depending on c_1. Applying a union bound over all models $j \in \mathcal{J}$ shows that, with probability at least $1 - \delta/(2d)$, the inequality $\|\widehat{\beta}_j - \beta_j\|_2 \leq \gamma/8$ holds simultaneously for all j.

On this high-probability event, consider any weight vector $w \in \Delta^{|\mathcal{J}_{-t}|-1}$. The convexity of the norm and the simplex properties give

$$\left\| \sum_{j \in \mathcal{J}_{-t}} w_j \widehat{\beta}_j - \sum_{j \in \mathcal{J}_{-t}} w_j \beta_j \right\|_2 \leq \sum_{j \in \mathcal{J}_{-t}} w_j \|\widehat{\beta}_j - \beta_j\|_2 \leq \frac{\gamma}{8} \sum_{j \in \mathcal{J}_{-t}} w_j = \frac{\gamma}{8}.$$

Moreover, $\|\widehat{\beta}_t - \beta_t\|_2 \leq \gamma/8$ holds as well.

Let d^* denote the true distance from β_t to the convex hull of $\{\beta_j : j \in \mathcal{J}_{-t}\}$, so that $d^* \geq \gamma$. For any $w \in \Delta^{|\mathcal{J}_{-t}|-1}$, the triangle inequality implies

$$\left\| \widehat{\beta}_t - \sum_{j \in \mathcal{J}_{-t}} w_j \widehat{\beta}_j \right\|_2 \geq \left\| \beta_t - \sum_{j \in \mathcal{J}_{-t}} w_j \beta_j \right\|_2 - \|\widehat{\beta}_t - \beta_t\|_2 - \left\| \sum_{j \in \mathcal{J}_{-t}} w_j \widehat{\beta}_j - \sum_{j \in \mathcal{J}_{-t}} w_j \beta_j \right\|_2$$

$$\geq d^* - \frac{\gamma}{8} - \frac{\gamma}{8} \geq \gamma - \frac{\gamma}{4} = \frac{3\gamma}{4}.$$

Taking the infimum over $w \in \Delta^{|\mathcal{J}_{-t}|-1}$ shows that, on the high-probability event, $\widehat{\text{dist}} \geq 3\gamma/4$ whenever $\gamma > 0$.

Under the null hypothesis of zero uniqueness margin, there exists $w^\dagger \in \Delta^{|\mathcal{J}_{-t}|-1}$ such that $\beta_t = \sum_{j \in \mathcal{J}_{-t}} w_j^\dagger \beta_j$. On the same high-probability event,

$$\left\| \widehat{\beta}_t - \sum_{j \in \mathcal{J}_{-t}} w_j^\dagger \widehat{\beta}_j \right\|_2 \leq \|\widehat{\beta}_t - \beta_t\|_2 + \left\| \sum_{j \in \mathcal{J}_{-t}} w_j^\dagger \widehat{\beta}_j - \sum_{j \in \mathcal{J}_{-t}} w_j^\dagger \beta_j \right\|_2$$

$$\leq \frac{\gamma}{8} + \frac{\gamma}{8} = \frac{\gamma}{4}.$$

Thus, under the null, $\widehat{\text{dist}} \leq \gamma/4$ on the high-probability event.

The decision rule compares $\widehat{\text{dist}}$ to the threshold $\gamma/2$. On the event where all coefficient estimation errors are bounded by $\gamma/8$, the rule outputs the correct decision both under the null and under the alternative. The only way the rule can err is if the event of simultaneous coefficient estimation accuracy fails. The probability of this event is at most δ by the choice of r and the union bound. Since the total number of queries per model is dr, the stated sample-complexity bound follows. $\qquad\square$

B.3 Minimax Lower Bound for Noisy Active Auditing

We next show that the dependence on σ^2, γ and δ in Proposition 7 cannot be improved in general, even in a one-dimensional special case.

Proposition 8 (Minimax lower bound for noisy active auditing). Consider the one-dimensional local linear model with feature map $\phi(x, \theta) \equiv 1$, so that $Y_j(x, \theta) = \beta_j$ for all (x, θ). Suppose that all peer coefficients are zero, $\beta_j = 0$ for $j \in \mathcal{J}_{-t}$. When the target is queried at any design point, the auditor observes $\widetilde{Y}_t = \beta_t + \varepsilon$, where $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ is Gaussian noise. Consider testing
$$H_0 : \beta_t = 0 \quad \textit{versus} \quad H_1 : \beta_t = \gamma,$$
for a fixed $\gamma > 0$. An active auditing procedure is allowed to adaptively select up to r queries and observe independent noisy evaluations at each query. For any such procedure, let ψ denote the resulting test and let

$$\text{err}(\psi) = \mathbb{P}_{H_0}(\psi = 1) + \mathbb{P}_{H_1}(\psi = 0)$$

be the sum of type I and type II error probabilities. If $\text{err}(\psi) \leq 2\delta$ for some $\delta \in (0, 1/4)$, then there exists a universal constant $c > 0$ such that

$$r \geq c \frac{\sigma^2}{\gamma^2} \log\left(\frac{1}{\delta}\right).$$

Proof. In this one-dimensional setting, the feature map is constant and equal to one, so the scalarised target response does not depend on the design point. Regardless of which design points are chosen, each noisy observation of the target has the form
$$Z_i = \beta_t + \varepsilon_i, \qquad i = 1, \ldots, r,$$
where the ε_i are independent $\mathcal{N}(0, \sigma^2)$ random variables. The joint distribution under H_0 is the product measure $P_0 = \bigotimes_{i=1}^r \mathcal{N}(0, \sigma^2)$, and under H_1 it is $P_1 = \bigotimes_{i=1}^r \mathcal{N}(\gamma, \sigma^2)$. Active choice of design points does not change these distributions.

The Kullback–Leibler divergence between $\mathcal{N}(\gamma, \sigma^2)$ and $\mathcal{N}(0, \sigma^2)$ is $\gamma^2/(2\sigma^2)$. Independence implies that the divergence between P_1 and P_0 is

$$D(P_1 \| P_0) = r \frac{\gamma^2}{2\sigma^2}.$$

The Bretagnolle–Huber inequality states that, for any two probability measures P and Q and any measurable test ψ,

$$P(\psi = 1) + Q(\psi = 0) \geq \frac{1}{2} \exp(-D(Q \| P)).$$

Applying this inequality with $P = P_0$ and $Q = P_1$ yields

$$\mathrm{err}(\psi) = P_0(\psi = 1) + P_1(\psi = 0) \geq \frac{1}{2} \exp\left(-D(P_1 \| P_0)\right) = \frac{1}{2} \exp\left(-\frac{r\gamma^2}{2\sigma^2}\right).$$

If $\mathrm{err}(\psi) \leq 2\delta$, then

$$2\delta \geq \frac{1}{2} \exp\left(-\frac{r\gamma^2}{2\sigma^2}\right).$$

Rearranging gives

$$\exp\left(-\frac{r\gamma^2}{2\sigma^2}\right) \leq 4\delta,$$

hence

$$-\frac{r\gamma^2}{2\sigma^2} \leq \log(4\delta).$$

For $\delta \in (0, 1/4)$, the term $\log(4\delta)$ is negative and bounded above by a negative multiple of $\log(1/\delta)$. More precisely, there exists a universal constant $c' > 0$ such that $-\log(4\delta) \geq c' \log(1/\delta)$ for all δ in this range. Substituting yields

$$r \geq c \frac{\sigma^2}{\gamma^2} \log\left(\frac{1}{\delta}\right)$$

for a constant c depending on c'. This shows that any active auditing procedure with total error at most 2δ requires at least this order of queries, establishing the lower bound. □

B.4 Detection Versus Estimation in One Dimension

We finally show that, in the same one-dimensional setting, estimating the parameter β_t up to accuracy of order γ with high probability requires the same order of queries as reliable detection.

Proposition 9 (No separation between detection and estimation). Consider the one-dimensional model and observation scheme of Proposition 8. Let $\widehat{\beta}$ be any estimator of β_t based on at most r noisy observations. Fix $\gamma > 0$ and $\delta \in (0, 1/4)$. Suppose that

$$\max\left\{\mathbb{P}_{\beta_t=0}\left(|\widehat{\beta} - 0| \geq \tfrac{\gamma}{2}\right), \ \mathbb{P}_{\beta_t=\gamma}\left(|\widehat{\beta} - \gamma| \geq \tfrac{\gamma}{2}\right)\right\} \leq \delta.$$

Then there exists a universal constant $c > 0$ such that

$$r \geq c \frac{\sigma^2}{\gamma^2} \log\left(\frac{1}{\delta}\right).$$

Proof. Assume that the estimator $\widehat{\beta}$ satisfies the stated accuracy property. Construct a hypothesis test for $H_0 : \beta_t = 0$ versus $H_1 : \beta_t = \gamma$ by thresholding the magnitude of the estimator. Specifically, define a test ψ by

$$\psi = \begin{cases} 1, & \text{if } |\widehat{\beta}| \geq \gamma/2, \\ 0, & \text{if } |\widehat{\beta}| < \gamma/2. \end{cases}$$

Under H_0, the type I error probability of this test is

$$P_0(\psi = 1) = \mathbb{P}_{\beta_t=0}\left(|\widehat{\beta}| \geq \gamma/2\right) = \mathbb{P}_{\beta_t=0}\left(|\widehat{\beta} - 0| \geq \gamma/2\right),$$

which is at most δ by assumption. Under H_1, the type II error probability is

$$P_1(\psi = 0) = \mathbb{P}_{\beta_t = \gamma}(|\widehat{\beta}| < \gamma/2).$$

On the event $\{|\widehat{\beta} - \gamma| < \gamma/2\}$, the reverse triangle inequality gives

$$|\widehat{\beta}| \geq |\gamma| - |\widehat{\beta} - \gamma| > \gamma - \gamma/2 = \gamma/2,$$

so on this event the test outputs $\psi = 1$. Therefore the event $\{\psi = 0\}$ is contained in $\{|\widehat{\beta} - \gamma| \geq \gamma/2\}$, and

$$P_1(\psi = 0) \leq \mathbb{P}_{\beta_t = \gamma}(|\widehat{\beta} - \gamma| \geq \gamma/2) \leq \delta.$$

The sum of type I and type II error probabilities of ψ is thus at most 2δ.

The test ψ is computed from $\widehat{\beta}$ and therefore uses at most r noisy observations of the target. Proposition 8 shows that any test based on at most r observations and satisfying $P_0(\psi = 1) + P_1(\psi = 0) \leq 2\delta$ must obey the lower bound

$$r \geq c \frac{\sigma^2}{\gamma^2} \log\left(\frac{1}{\delta}\right)$$

for some universal constant $c > 0$. Since ψ has been constructed from $\widehat{\beta}$ and meets the error condition, the same lower bound must hold for r. This proves that estimating β_t up to accuracy of order γ with high probability cannot be achieved with fewer queries than required for reliable detection of non-zero uniqueness margin. $\qquad\square$

Appendix C Observational Non-Identifiability and Shapley Redundancy

This note collects two results: a simple example showing that Shapley-value attribution cannot detect redundancy as defined by PIER, and a non-identifiability statement for uniqueness under purely observational designs.

C.1 Shapley Value Does Not Detect Redundancy

We now formalise an example showing that Shapley-value attribution does not capture redundancy in the sense of PIER.

Proposition 10 (Shapley value cannot detect redundancy)**.** Consider an ecosystem consisting of two models M_1 and M_2 with identical scalarised responses $Y_1(x, \theta) = Y_2(x, \theta)$ for all (x, θ). Define a characteristic function $v : 2^{\{M_1, M_2\}} \to \mathbb{R}$ by

$$v(\emptyset) = 0, \qquad v(\{M_1\}) = 1, \qquad v(\{M_2\}) = 1, \qquad v(\{M_1, M_2\}) = 1.$$

Let $\phi_{M_1}(v)$ and $\phi_{M_2}(v)$ denote the Shapley values of M_1 and M_2 under v. Then $\phi_{M_1}(v) = \phi_{M_2}(v) = 1/2 > 0$. If M_1 is treated as target and M_2 as the sole peer, the population PIER of M_1 is identically zero and its uniqueness functional vanishes.

Proof. In a two-player cooperative game with players $\{M_1, M_2\}$, the Shapley value for M_1 can be written as

$$\phi_{M_1}(v) = \frac{1}{2}\big(v(\{M_1\}) - v(\emptyset)\big) + \frac{1}{2}\big(v(\{M_1, M_2\}) - v(\{M_2\})\big),$$

and similarly for M_2. Substituting the values of v yields

$$\phi_{M_1}(v) = \frac{1}{2}(1 - 0) + \frac{1}{2}(1 - 1) = \frac{1}{2},$$

and symmetry gives $\phi_{M_2}(v) = 1/2$ as well. Each model therefore receives strictly positive credit under Shapley-value attribution.

To evaluate PIER, take M_1 as the target and $\mathcal{J}_{-t} = \{M_2\}$ as the peer set. The convex peer expressivity class associated with $\mathcal{J}_{-t}$ consists of functions of the form $h_w(x, \theta) = wY_2(x, \theta)$ with $w \in [0, 1]$. Because Y_1 and Y_2 coincide everywhere, the population residual associated with a weight w is

$$L(w) = \mathbb{E}\big[(Y_1(X, \vartheta) - wY_2(X, \vartheta))^2\big] = \mathbb{E}\big[(Y_1(X, \vartheta) - wY_1(X, \vartheta))^2\big],$$

which is uniquely minimised at $w^* = 1$. The corresponding convex combination is $h_{w^*}(x, \theta) = Y_1(x, \theta)$ for all (x, θ), so the PIER is

$$R_{M_1}(x, \theta) = Y_1(x, \theta) - h_{w^*}(x, \theta) = 0$$

everywhere. The uniqueness functional $\mathcal{U}_{M_1} = \mathbb{E}_{\mathbb{P}_{\text{eval}}}[|R_{M_1}(X, \vartheta)|]$ is therefore zero under any evaluation design. The target is fully redundant relative to its peer in the sense of PIER, yet Shapley-value attribution assigns it non-zero credit. $\qquad\square$

C.2 Non-Identifiability of Uniqueness Without Matched Interventions

We work with a reference design distribution $\mathbb{P}^\star_{\text{design}}$ on $\mathcal{X} \times \Theta$ with support S and assume that there exists a measurable subset $\mathsf{S}_0 \subset \mathsf{S}$ such that $\mathbb{P}^\star_{\text{design}}(\mathsf{S}_0) > 0$ and $\mathbb{P}^\star_{\text{design}}(\mathsf{S} \setminus \mathsf{S}_0) > 0$.

Proposition 11 (Non-identifiability under unmatched observational designs). Consider an auditor who, for each model $j \in \mathcal{J}$, only has access to observational logs of the form

$$\{(X_{j,i}, \vartheta_{j,i}, Y_j(X_{j,i}, \vartheta_{j,i}))\}_{i=1}^{n_j},$$

where $(X_{j,i}, \vartheta_{j,i})$ are independent and identically distributed draws from an unknown design distribution Q_j on $\mathcal{X} \times \Theta$ and $Y_j : \mathcal{X} \times \Theta \to \mathbb{R}$ are deterministic scalarised responses. Suppose that Q_j is supported on S_0 for all j. Then there exist two ecosystems, indexed by $k \in \{0, 1\}$, consisting of response functions $\{Y_j^{(k)}\}$ and observational designs $\{Q_j^{(k)}\}$, such that

- for each j and all sample sizes $\{n_j\}$, the joint distribution of observational logs is the same in the two ecosystems,

- for a fixed target index t and peer set $\mathcal{J}_{-t}$, the population uniqueness functional $U_t^{(0)}$ of the target under $\mathbb{P}^\star_{\text{design}}$ is zero in ecosystem $k = 0$, while $U_t^{(1)}$ is strictly positive in ecosystem $k = 1$.

Consequently, the population uniqueness functional with respect to $\mathbb{P}^\star_{\text{design}}$ is not identifiable from purely observational logs generated under unmatched designs Q_j.

Proof. We construct two ecosystems explicitly. For each j and $k \in \{0, 1\}$, take $Q_j^{(k)}$ to be any fixed distribution supported on S_0, for example the normalised restriction of $\mathbb{P}^\star_{\text{design}}$ to S_0. Set $Q_j^{(0)} = Q_j^{(1)}$ for all j so that the observational designs are identical across ecosystems and concentrated on S_0.

Fix a target index t and a non-empty peer set $\mathcal{J}_{-t} \subseteq \mathcal{J} \setminus \{t\}$. In ecosystem $k = 0$, define all peer responses by $Y_j^{(0)}(x, \theta) = 0$ for all (x, θ) and all $j \in \mathcal{J}_{-t}$, and the target response by $Y_t^{(0)}(x, \theta) = 0$ for all (x, θ). Under $\mathbb{P}^\star_{\text{design}}$, the convex peer expressivity class contains only the zero function, so the projection of $Y_t^{(0)}$ is zero and the PIER is identically zero. The uniqueness functional $U_t^{(0)} = \mathbb{E}_{\mathbb{P}^\star_{\text{design}}}[|R_t^{(0)}(X, \vartheta)|]$ is therefore zero.

In ecosystem $k = 1$, keep the peer responses unchanged by setting $Y_j^{(1)}(x, \theta) = 0$ for all (x, θ) and $j \in \mathcal{J}_{-t}$. Define the target response by $Y_t^{(1)}(x, \theta) = 0$ for $(x, \theta) \in \mathsf{S}_0$, $Y_t^{(1)}(x, \theta) = 1$ for $(x, \theta) \in \mathsf{S} \setminus \mathsf{S}_0$, and an arbitrary value outside S. Under $\mathbb{P}^\star_{\text{design}}$, all peer responses are zero, whereas the target response coincides with the indicator of $\mathsf{S} \setminus \mathsf{S}_0$. The convex peer class again consists only of the zero function, so the projection of $Y_t^{(1)}$ is zero and the PIER is equal to $Y_t^{(1)}$. The uniqueness functional becomes

$$U_t^{(1)} = \mathbb{E}_{\mathbb{P}^\star_{\text{design}}}[|Y_t^{(1)}(X, \vartheta)|] = \mathbb{P}^\star_{\text{design}}(\mathsf{S} \setminus \mathsf{S}_0),$$

which is strictly positive by assumption on S_0.

To compare observational logs, note that in both ecosystems $(X_{j,i}^{(k)}, \vartheta_{j,i}^{(k)}) \sim Q_j$ are supported on S_0. On this set, the peer responses are identically zero in both ecosystems, and the target responses are also zero because $Y_t^{(1)}$ equals zero on S_0. Therefore, for each j and each i, the triples $(X_{j,i}^{(0)}, \vartheta_{j,i}^{(0)}, Y_j^{(0)}(X_{j,i}^{(0)}, \vartheta_{j,i}^{(0)}))$ and $(X_{j,i}^{(1)}, \vartheta_{j,i}^{(1)}, Y_j^{(1)}(X_{j,i}^{(1)}, \vartheta_{j,i}^{(1)}))$ have the same distribution. Independence across i and j implies that the joint distribution of all observational logs is identical in the two ecosystems for any sample sizes $\{n_j\}$.

Assume, for contradiction, that there exists a procedure that maps the joint law of observational logs to the true value of U_t. Since the joint law is the same in both ecosystems, the output of such a procedure must be the same under $k = 0$ and $k = 1$, hence it cannot simultaneously equal $U_t^{(0)} = 0$ and $U_t^{(1)} = \mathbb{P}^\star_{\text{design}}(\mathsf{S} \setminus \mathsf{S}_0) > 0$. This contradiction shows that U_t is not identifiable from observational logs generated under unmatched designs, as claimed. □

Appendix D Uniqueness, Robustness and Scalar Summaries

We now formalise the observation that the scalar uniqueness functional cannot distinguish between stable and highly oscillatory forms of uniqueness along the dose axis.

Proposition 12 (Ambiguity between uniqueness and robustness). Fix a context and design distributions $(\mathbb{P}_{\text{fit}}, \mathbb{P}_{\text{eval}})$ on $\mathcal{X} \times \Theta$. Assume that there exists an input $x_0 \in \mathcal{X}$ and that the support of $\mathbb{P}_{\text{eval}}$ contains points of the form (x_0, θ) for all θ in the unit interval $[0, 1]$. Let $0 < L_{\text{low}} < L_{\text{high}}$ be arbitrary constants. Then there exist two ecosystems sharing the same peer set and two targets A and B such that

$$\mathcal{U}_{tA} = \mathbb{E}_{\mathbb{P}_{\text{eval}}}[|R_A(X, \vartheta)|] = \mathbb{E}_{\mathbb{P}_{\text{eval}}}[|R_B(X, \vartheta)|] = \mathcal{U}_{tB},$$

while the PIER functions R_A and R_B satisfy

$$\sup_{\theta \neq \theta'} \frac{|R_A(x_0, \theta) - R_A(x_0, \theta')|}{|\theta - \theta'|} \leq L_{\text{low}}, \qquad \sup_{\theta \neq \theta'} \frac{|R_B(x_0, \theta) - R_B(x_0, \theta')|}{|\theta - \theta'|} \geq L_{\text{high}}.$$

Proof. The idea is to construct two scalar functions of the dose θ with identical expected absolute values under the marginal of $\mathbb{P}_{\text{eval}}$ on $[0, 1]$ but with different Lipschitz constants, and then embed them as PIER functions at an input x_0.

Consider the marginal distribution of (X, ϑ) under $\mathbb{P}_{\text{eval}}$. By assumption, the event $\{X = x_0\}$ has positive probability and, conditional on $X = x_0$, the dose ϑ has support containing $[0, 1]$. For simplicity, we argue as if (X, ϑ) under $\mathbb{P}_{\text{eval}}$ were uniform on $\{x_0\} \times [0, 1]$; any other absolutely continuous distribution with full support on this set can be handled by adjusting constants.

Define a function $r_A : [0, 1] \to \mathbb{R}$ by setting $r_A(\theta) = c_A$ for all θ, where c_A is a constant. This function is constant and hence has Lipschitz constant zero. Choosing any $L_{\text{low}} > 0$, it holds that

$$\sup_{\theta \neq \theta'} \frac{|r_A(\theta) - r_A(\theta')|}{|\theta - \theta'|} = 0 \leq L_{\text{low}}.$$

The expected absolute value of $r_A(\vartheta)$ under the uniform distribution on $[0, 1]$ is $|c_A|$.

Next, define a function $r_B : [0, 1] \to \mathbb{R}$ by $r_B(\theta) = c \sin(2\pi K\theta)$ for some amplitude $c > 0$ and integer frequency $K \geq 1$. This function is continuously differentiable with derivative $r_B'(\theta) = 2\pi Kc \cos(2\pi K\theta)$. The mean value theorem implies that the Lipschitz constant of r_B on $[0, 1]$ equals the supremum of $|r_B'(\theta)|$ over $\theta \in [0, 1]$, which is $2\pi Kc$. By choosing K sufficiently large, specifically $K \geq L_{\text{high}}/(2\pi c)$, one ensures that the Lipschitz constant is at least L_{high}.

To match expected absolute values, compute

$$\mathbb{E}[|r_B(\vartheta)|] = c\, \mathbb{E}[|\sin(2\pi K\vartheta)|].$$

If ϑ is uniform on $[0, 1]$, then $U = 2\pi K\vartheta$ is uniform on $[0, 2\pi K]$, and the periodicity of $|\sin(\cdot)|$ implies

$$\mathbb{E}[|\sin(2\pi K\vartheta)|] = \frac{1}{2\pi K} \int_0^{2\pi K} |\sin u|\, \mathrm{d}u = \frac{1}{2\pi} \int_0^{2\pi} |\sin u|\, \mathrm{d}u = \frac{2}{\pi},$$

so that $\mathbb{E}[|r_B(\vartheta)|] = 2c/\pi$. Setting $c_A = 2c/\pi$ yields $\mathbb{E}[|r_A(\vartheta)|] = |c_A| = 2c/\pi = \mathbb{E}[|r_B(\vartheta)|]$.

To embed these functions as PIER, consider an ecosystem with a single peer that has identically zero scalarised response, $Y_j(x, \theta) = 0$ for all (x, θ). For any target in this ecosystem, the convex peer expressivity class consists only of the zero function, so the PIER coincides with the target's scalarised response. For target A, define $Y_A(x_0, \theta) = r_A(\theta)$

for $\theta \in [0,1]$ and extend arbitrarily outside this set. For target B, define $Y_B(x_0, \theta) = r_B(\theta)$ on $[0,1]$ and extend arbitrarily elsewhere. Under $\mathbb{P}_{\text{eval}}$, the marginal distribution of ϑ conditional on $X = x_0$ yields identical expectations $\mathbb{E}_{\mathbb{P}_{\text{eval}}}[|R_A(X, \vartheta)|] = \mathbb{E}_{\mathbb{P}_{\text{eval}}}[|R_B(X, \vartheta)|]$, while the Lipschitz constants of the PIER functions along θ at x_0 differ as constructed. This establishes the claimed ambiguity. $\qquad\square$