Black-box model classification under the discriminative factorization

Hayden Helm[*]
Helivan
hayden@helivan.io

Merrick Ohata
JHU

Carey E. Priebe
JHU

Abstract

Access to modern generative systems is often restricted to querying an API (the "black-box" setting) and many properties of the system are unknown to the user at inference time. Recent work has shown that low-dimensional representations of models based on the relationship of their embedded responses to a set of queries are useful for inferring model-level properties. However, the quality of these representations is highly sensitive to the query set. We introduce the *discriminative factorization* to distinguish between high- and low-quality query sets in the context of black-box model-level classification. Under this framework, the probability of chance-level classification is bounded by a term that decays exponentially in the query budget. On three auditing tasks, estimated factorization parameters predict the empirical performance decay rate. We conclude by showing that query sets selected using the estimated factorization reproduce the empirical ordering of oracle query sets.

1 Introduction

As model providers increasingly restrict access to only API endpoints [29, 3, 13], black-box interactions are often the only available mode for users and researchers. Black-box access implies that some properties of the model – e.g., the presence of particular data in the (pre-)training corpus [41, 4] – are not known and must be inferred from the model's responses to queries. In our setting, we assume access to a collection of models for which the property is known and that we can relate the responses from a new model to the responses from the labeled models. Our goal is to infer the property of a new / unlabeled model.

Recent work has shown that collections of black-box models can be represented in Euclidean space by comparing embedded responses to a set of queries [15, 1], and that these representations enable consistent model-level inference [16]. However, existing results treat all query sets as equally informative. In practice, the signal carried by a query set varies widely: sets containing probes aligned with the inference task ("signal") produce substantially more discriminative representations than those "orthogonal" to it [15, 16]. This distinction has remained informal.

We introduce a framework, the *discriminative factorization*, to formalize the signal/orthogonal distinction. The framework enables bounding the classification risk for a given query budget, estimating the number of queries required to achieve a given risk, and selecting high signal query sets for classification. We apply this framework to a synthetic problem and three system auditing tasks: detecting sensitive fine-tuning data, identifying covert persuasion bias via system prompts, and finding RAG systems with access to restricted documents. Our contributions are the following:

1. **The discriminative factorization.** We introduce the discriminative factorization and decompose query-model interaction into discriminative directions. The factorization separates

[*]Corresponding author

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

the cross-model information into query-dependent weights and model-pair-dependent sensitivities, enabling a formal notion of signal/orthogonal queries. In settings where the embedding function is injective, the discriminative factorization yields a bound, exponential in the query budget, on the probability that classification error exceeds the Bayes risk, governed by the prevalence of "orthogonal" queries. In other settings, a similar rate applies to chance-level classification. Further, we provide a sufficient number of queries to perform better-than-chance (or, when the embedding function is injective, to classify with a particular risk).

2. **Estimation from spectral structure.** We show that the properties of the discriminative factorization – the discriminative rank and the zero-set probabilities – can be estimated from the SVD of a query-model matrix. In each setting, we demonstrate that the estimated parameters predict empirical decay rates when expected.
3. **Query selection without task-specific knowledge.** For all real tasks, we construct query sets using the estimated discriminative factorization. In each case, the constructed signal/orthogonal sets reproduce the oracle ordering, improving classification efficiency without task-specific knowledge of which queries are informative.

1.1 Related work

Low-dimensional representations of generative models. Low-dimensional representations of generative models can be constructed from internal activations [34, 22, 18, 9] and model weights [2, 6]. These approaches require access to model internals, which are unavailable in the black-box setting. Black-box approaches address this constraint by constructing representations from embedded responses to a query set via multidimensional scaling [15, 1, 16, 17]. These representations can then be used for model-level inference [16, 17]. We build on this foundation by addressing how the query distribution affects the quality of the representations for a given task.

Inference on models. Prior black-box analyses focus primarily on individual models. For example, membership inference [41, 8] and training data extraction [4] test whether specific examples influenced a single model, but typically require token-level log-probabilities. Pretraining data detection [40, 24] requires either similar assumptions or explicit examples of data in the training set. Datamodels [19] predict model outputs as a function of training data composition, but require retraining many models with known data subsets instead of solely using black-box queries. We provide theoretical guarantees connecting query budget to classification risk when using low-dimensional representations from embedded responses and demonstrate that use of our framework can directly improve practical query efficiency.

2 Motivation and setting

Let $\mathcal{F} := \{f : \mathcal{Q} \to \mathcal{X}\}$ be the space of models. Each $f \in \mathcal{F}$ is a random mapping from a finite input space $\mathcal{Q}$ with $|\mathcal{Q}| = M < \infty$ to a finite output space $\mathcal{X}$ with $|\mathcal{X}| = V < \infty$. We refer to $q \in \mathcal{Q}$ as a *query* and the model's output $f(q)$ as a *response*. We assume responses $f(q)_1, \ldots, f(q)_K$ are i.i.d. samples from $P_f(q)$.

In the black-box setting, the model f is completely characterized by its response distributions, $f = \{P_f(q)\}_{q\in\mathcal{Q}}$, and $f \neq f'$ if and only if there exists $q \in \mathcal{Q}$ such that $P_f(q) \neq P_{f'}(q)$. In practice, analysis directly on $\mathcal{X}$ is often inconvenient or intractable. Instead, for a given embedding function $g : \mathcal{X} \to \mathbb{R}^p$, we work with embedded responses $g(f(q))_1, \ldots, g(f(q))_K$ assumed to be i.i.d. from $P_f^{(g)}(q)$. Unless otherwise stated, we refer to $P_f^{(g)}(q)$ as $P_f(q)$ for the remainder of this paper.

To capture the dissimilarity between models, for a multiset of queries $Q \subseteq_m \mathcal{Q}$ we define

$$d_Q(f, f') := \sqrt{\sum_{q\in Q} \frac{1}{m} d_P^2(P_f(q), P_{f'}(q))}$$

where d_P is a metric on $\mathcal{P}_1(\mathbb{R}^p)$ such that d_P^2 is of negative type (such as the energy distance [43, 36]). As a nonnegative combination of such terms, d_Q^2 is also of negative type, so d_Q embeds isometrically in a Hilbert space [39].

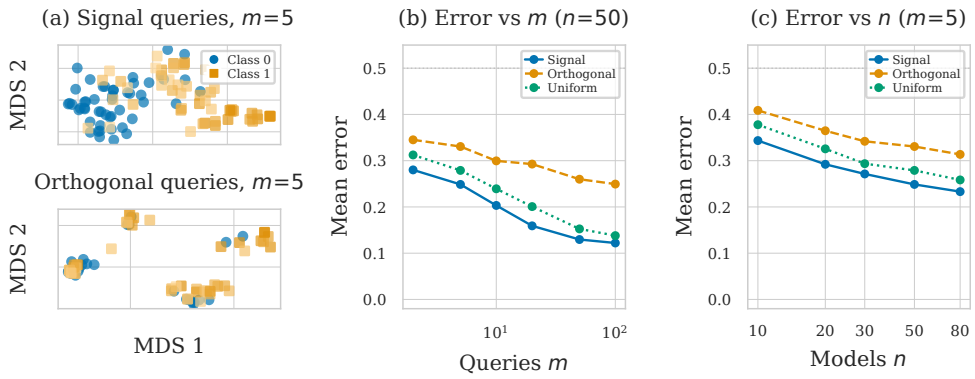

Figure 1: Classification of models by the presence of particular fine-tuning data. An oracle knows which queries probe the axis that distinguishes the two classes (Signal) and which do not (Orthogonal). (a) Two-dimensional MDS embeddings using $m = 5$ signal queries (top) and orthogonal queries (bottom). Each dot is a model. (b) Classification error as a function of query budget m with $n = 50$ for signal, orthogonal, and uniform query sets. (c) Classification error as a function of number of models with $m = 5$ for the same query sets. The relative discriminative quality of the three query sets is stable across m and n. Reported errors are the average of 500 different random samples.

Query-dependent model representations. For a collection of black-box models $f_1, \ldots, f_n$ and multiset of queries Q, we define the pairwise distance matrix $D := D_{i,i'} = d_Q(f_i, f_{i'})$. We then apply classical multidimensional scaling (MDS) [44] to D to obtain representations $\psi_Q(f_1), \ldots, \psi_Q(f_n) \in \mathbb{R}^d$. These are vector representations of the models f_i with respect to g and Q. For some choices of d_P, estimates of the representations $\psi_Q(f_i)$ are consistent for their population counterpart [1] and provide consistent model-level inference [16].

Since $g(\mathcal{X}) \subset \mathbb{R}^p$ is finite, each $P_f(q)$ is a categorical distribution supported on at most V atoms in $\mathbb{R}^p$. And, since d_Q^2 is negative type, MDS achieves zero stress for any $n < \infty$ for $d \geq \min\{n - 1, m(V - 1)\}$ [42, 39].

2.1 Motivating example: detecting sensitive training data

To demonstrate the impact of query set on model representations and subsequent inference, we use a task inspired by Helm et al. [16]: using only black-box access and a collection of labeled models, can we identify whether a particular topic appeared in a model's fine-tuning corpus?

To investigate, we train 100 LoRA adapters from `Qwen2.5-1.5B-Instruct` on subsets of Yahoo Answers [47]: 50 on non-sensitive topics (class 0) and 50 on mixtures containing "Politics & Government" data (class 1). We construct model representations with respect to three input query sets $\mathcal{Q}$: *Signal* queries directly from "Politics & Government" data, *Orthogonal* queries on topics not included in any of the fine-tuning datasets, and *Uniform*, the union of the two.

For each of the three groups, we sample m queries uniformly and construct $\psi_Q(f_1), \ldots, \psi_Q(f_{n+1})$ with embedding dimension $d = 8$. We train a random forest classifier with default parameters [31] on the n random training representations and evaluate on the remainder. This process is repeated 500 times (see Appendix B.1 for additional details).

Figure 1 illustrates the phenomenon of interest. Panel (a) shows embeddings in the first two dimensions for a single instance of $m = 5$ queries from the Signal and Orthogonal sets. Signal queries yield strong separation of classes while separation from Orthogonal queries is much weaker. Panel (b) quantifies this observation via the classification error on the test models. The relative ordering matches intuition for all query budgets: Signal sets outperform Uniform sets, which outperform Orthogonal sets. Orthogonal query sets still enable better-than-chance classification on average. Panel (c) shows that this ordering is stable across the number of models for a fixed budget, with a decrease in error as the size of the training sample increases.

In practice, the collection of queries that probe the discriminative axes of a classification task is not known *a priori*. We introduce a framework where the query-model relationship is naturally represented as a matrix. The spectral properties of this matrix, which can be estimated from the available labeled data, separate Signal and Orthogonal queries – thereby improving the query efficiency of inference on models without prior task-specific knowledge.

2.2 Model classification and risk

Let $\mathcal{Y} = \{0, \dots, C-1\}$ be the set of model classes. We observe a collection of models and labels $(f_1, y_1), \dots, (f_n, y_n) \in \mathcal{F} \times \mathcal{Y}$ drawn i.i.d. from P_{fY}. We also observe an unlabeled model f_{n+1}. Our goal is to predict the corresponding label y_{n+1}.

The risk of a classifier $h : \mathcal{F} \to \mathcal{Y}$ under 0-1 loss is given by $L_{fY}(h) = \mathbb{P}_{fY}[h(f) \neq y]$. The *Bayes risk* of a classification problem, denoted $L^*(P_{fY})$, is the infimum risk of classifiers $\{h : \mathcal{F} \to \mathcal{Y}\}$. For balanced binary classes, when $L^*(P_{fY}) < 0.5$, there exists a model-level classifier that performs better than chance.

Classification on the space of models $\mathcal{F}$ is impractical. Instead, we consider the proxy problem in the low-dimensional space induced by Q. That is, we consider $h_d \in \mathcal{H}_d = \{h_d : \mathbb{R}^d \to \mathcal{Y}\}$ trained on the representations $(\psi_Q(f_1), y_1), \dots, (\psi_Q(f_n), y_n)$ that predicts the label of f_{n+1} via $\psi_Q(f_{n+1})$. The risk of h_d is given by $L^{(n)}_{\psi Y}(h_d, Q) = \mathbb{P}_{(f_{n+1}, y_{n+1})}[h_d(\psi_Q(f_{n+1})) \neq y_{n+1} \mid g, Q, (f_1, y_1), \dots, (f_n, y_n)]$.

Our main results (§3.1) show that under mild assumptions, the low-dimensional representations suffice for better-than-chance classification and, when g is injective, for Bayes-optimal classification.

3 The discriminative factorization

We now introduce the discriminative factorization. Section 2.1 showed that query sets that probe the discriminative axes induce better representations for classification. To formalize this, we decompose the content of a query into discriminative directions.

Definition 1. *Consider a model space $\mathcal{F}$, embedding function g, and query set $\mathcal{Q}$. If α and ϕ are maps $\alpha : \mathcal{Q} \to [0,1]^r$ and $\phi : \mathcal{F} \times \mathcal{F} \to [0,\infty)^r$ such that for all $f, f' \in \mathcal{F}$ and $q \in \mathcal{Q}$*

$$d_P^2(P_f(q), P_{f'}(q)) = \sum_{\ell=1}^{r} \alpha_\ell(q)\, \phi_\ell(f, f')$$

then we say (α, ϕ) admit a discriminative factorization *of rank r for $(\mathcal{F}, \mathcal{Q}, g, d_P)$. This factorization is minimal if no factorization of rank $r' < r$ exists.*

For each q, the *weight* $\alpha_\ell(q)$ quantifies the intensity of discriminative signal along *direction* ℓ, while $\phi_\ell(f, f')$ captures the *sensitivity* of the model pair (f, f') to that direction. A query is in the *zero-set* of direction ℓ if $\alpha_\ell(q) = 0$; let $\mathcal{Z}_\ell := \{q : \alpha_\ell(q) = 0\}$. For a query distribution Π_Q on $\mathcal{Q}$, the *zero-set probabilities* are given by $\rho_\ell = \Pi_Q(\mathcal{Z}_\ell)$. A query $q \in \bigcap_\ell \mathcal{Z}_\ell$ is *orthogonal* to the task. This formalizes the notion of "orthogonal" discussed in §2.1.

A discriminative factorization always exists with $r \leq M$: For any enumeration $q_1, \dots, q_M$ of $\mathcal{Q}$, take $\alpha_\ell(q) = \mathbb{1}[q = q_\ell]$ and $\phi_\ell(f, f') = d_P^2(P_f(q_\ell), P_{f'}(q_\ell))$. With observed models $f_1, \dots, f_n$, if $\sum_{q \in Q} \alpha_\ell(q) > 0$ for all ℓ, the matrix of squared distances $[d_Q^2(f_i, f_{i'})]_{i,i'}$ is a conic combination of the r matrices $\Phi_\ell = [\phi_\ell(f_i, f_{i'})]_{i,i'}$ and MDS achieves zero stress for some $d \leq \sum_\ell \operatorname{rank}(J \Phi_\ell J)$, where J is the centering matrix. If each $\phi_\ell(f, f') = (x_\ell(f) - x_\ell(f'))^2$, each term has rank 1 and this reduces to $d \leq r \ll m(V-1)$. More generally, if each $\phi_\ell(f, f') = \|x_\ell(f) - x_\ell(f')\|_2^2$ for $x_\ell(f) \in \mathbb{R}^{k_\ell}$, then $d \leq \sum_\ell k_\ell$.

Relation to other factorizations and decompositions. Definition 1 is a query-indexed low-rank factorization of a collection of response dissimilarities. Like Fourier or Mercer expansions [26], it writes a function as a weighted sum of terms, but its terms ϕ_ℓ are not fixed in advance and both factors are nonnegative. Represented as a matrix over $\mathcal{Q} \times (\mathcal{F} \times \mathcal{F})$, it is a rank-$r$ nonnegative matrix factorization [23]. Section 3.2 estimates its column space with the singular value decomposition,

which gives the best rank-r approximation in Frobenius norm [10] but with mixed-sign factors. The closest classical model is INDSCAL [5], a form of three-way multidimensional scaling that decomposes a stack of dissimilarity matrices, one per subject, into shared dimensions with nonnegative subject-specific weights. When each $\phi_\ell(f, f') = (x_\ell(f) - x_\ell(f'))^2$ for a scalar function x_ℓ, Definition 1 is exactly INDSCAL with queries in the role of subjects. To our knowledge, this factorization has not previously been used to study collections of black-box functions with random outputs.

3.1 Theoretical results

For our theoretical analysis, we let $\mathcal{Y} = \{0, 1\}$ and assume balanced classes $P_{fY}(y = 1) = 0.5$. We further assume $L^*(P_{fY}) < 0.5$. We let α and ϕ admit a minimal discriminative factorization of rank r with zero-set probabilities $\rho_1, \ldots, \rho_r$. Our results depend on rank r and the $\{\rho_\ell\}$, which are characteristic of the task and query distribution Π_Q. We require that each $P_f(q)$ have finite first moment and make mild assumptions on Π_Q and the quality of g:

Assumption 1 (Direction-covering query distribution)**.** *For each direction $\ell \in \{1, \ldots, r\}$, the query distribution Π_Q places positive mass outside the zero set: $\rho_\ell = \Pi_Q(\mathcal{Z}_\ell) < 1$.*

Assumption 2 (Non-degenerate embedding)**.** *The embedded class-conditional distributions differ: $P_{g(f)|y=0} \neq P_{g(f)|y=1}$, where $g(f) := \{P_f^{(g)}(q)\}_{q \in \mathcal{Q}}$.*

Given the high dimensionality of semantic spaces, it is extremely unlikely that a given query is in the zero set of a particular direction.[1] Hence, Assumption 1 will be satisfied by most distributions on query sets. Assumption 2 ensures that the embedding function g does not destroy all discriminative information available at the model level. For balanced classes it is equivalent to assuming $L^*(P_{g(f)Y}) < 0.5$, since $L^*(P_{g(f)Y}) = \frac{1}{2}\big(1 - \mathrm{TV}(P_{g(f)|y=0}, P_{g(f)|y=1})\big)$.

Our first result shows that the MDS representation induced by a large-enough query set suffices for better-than-chance classification under Assumption 1 and Assumption 2. Full proofs of all results in this section are provided in Appendix A.

Theorem 1. *Let P_{fY} satisfy Assumption 2 and $Q = \{q_1, \ldots, q_m\}$ a multiset with i.i.d. queries from Π_Q satisfying Assumption 1. Then there exists $n^* \in \mathbb{N}$ such that for all $n \geq n^*$*

$$\mathbb{P}_{Q,(f_i,y_i)}\left[\inf_{h_d \in \mathcal{H}_d} L_{\psi Y}^{(n)}(h_d, Q) \geq 0.5\right] \leq \sum_{\ell=1}^{r} \rho_\ell^m + \gamma_\epsilon(n),$$

for $d \geq \min\{r^2, n, m(V-1)\}$, where $\gamma_\epsilon(n) \to 0$ as $n \to \infty$.

The bound in Theorem 1 makes extremely mild assumptions on the embedding function g. The quality of g is absorbed into $\{\alpha_\ell\}$ and $\{\rho_\ell\}$. When the embedding function g is injective[2], a stronger result holds: the MDS representation asymptotically achieves the Bayes risk.

Theorem 2. *In the setting of Theorem 1, suppose g is injective. For any $\epsilon > 0$ there exists $n_\epsilon^* \in \mathbb{N}$ such that for all $n \geq n_\epsilon^*$,*

$$\mathbb{P}_{Q,F,(f,y)}\left[\inf_{h_d \in \mathcal{H}_d} L_{\psi Y}^{(n)}(h_d, Q) \geq L^*(P_{fY}) + \epsilon\right] \leq \sum_{\ell=1}^{r} \rho_\ell^m + \gamma_\epsilon(n),$$

for $d \geq \min\{r^2, n, m(V-1)\}$, where $\gamma_\epsilon(n) \to 0$ as $n \to \infty$.

The bounds in both Theorem 1 and Theorem 2 hold for any finite m. Letting $\rho = \max_\ell \rho_\ell$, the first term becomes $\sum_{\ell=1}^{r} \rho_\ell^m \leq r\rho^m$. When $r = M$, we get $M(1 - \min_q \Pi_Q(q))^m$, recovering the coupon-collector scaling in which the entire query space must be covered [11].

The term $\gamma(n)$ captures the sample complexity of learning in the representation space $\mathbb{R}^d$, with the specific rate highly dependent on the class-conditional distributions. Indeed, by the "no free lunch" theorem [7], for any classifier such that $\gamma(n) \to 0$ and fixed n, there exist class-conditional distributions for which the expected performance is near chance.

[1] e.g., a generic Sports query may be related to a task determining the presence of Politics & Government fine-tuning data through semantic leak from legislation on sports betting.

[2] While modern sentence embedding models [35, 46, 27, 32, 30] are not injective, they are practically injective in that the number of inputs they map to the same element of $\mathbb{R}^p$ is small.

Explicit rates follow if the class-conditional distributions are restricted. For example, suppose the representations $\psi_Q(f)$ given $y = 0$ and $y = 1$ are distributed as $\mathcal{N}(\mu_0, \Sigma)$ and $\mathcal{N}(\mu_1, \Sigma)$. The Bayes rule is linear with $L^* = \mathbb{P}(Z > \Delta/2)$ for a standard normal Z, where $\Delta = \sqrt{(\mu_1 - \mu_0)^\top \Sigma^{-1} (\mu_1 - \mu_0)}$ is the Mahalanobis separation between the classes. The plug-in linear discriminant has expected excess risk $O(d/n)$, with a constant that depends on Δ [28, 25]. By Markov's inequality, $\gamma_\epsilon(n) = O(d/(n\epsilon))$ in Theorem 2, and the same rate holds for $\gamma(n)$ in Theorem 1 with $\epsilon = 1/2 - L^*$.

We now state a corollary that makes the query sample complexity explicit. Using this result, it is possible to calculate the query budget required for given error tolerance.

Corollary 1. *Under the conditions of Theorem 1 or Theorem 2, let $\rho = \max_\ell \rho_\ell$. For all $\delta > 0$, there exist $m^* \in \mathbb{N}$ and $n^* \in \mathbb{N}$ such that for all $m \geq m^*$ and $n \geq n^*$*

$$\mathbb{P}_{Q,(f_i,y_i)} \left[\inf_{h_d \in \mathcal{H}_d} L^{(n)}_{\psi_Y}(h_d, Q) \geq \mathcal{L} \right] \leq \delta,$$

where $\mathcal{L} = 0.5$ under Theorem 1 and $\mathcal{L} = L^*(P_{fY}) + \epsilon$ for fixed $\epsilon > 0$ under Theorem 2. In particular, $m^* = \left\lceil \frac{\ln(\delta/2r)}{\ln \rho} \right\rceil$ and any n^* such that $\gamma_{(\epsilon)}(n) \leq \delta/2$ for all $n \geq n^*$ suffice.

3.2 Estimating the discriminative rank and zero-sets

Our results depend on r and $\{\rho_\ell\}$. In practice, they must be estimated from the n labeled models and their responses to a pilot set of m queries Q. To do so, we construct the $m \times n^2$ sample *Query-Model Matrix* $E := E_{j,(i,i')} = d_P^2(P_{f_i}(q_j), P_{f_{i'}}(q_j))$ for each $(f_i, f_{i'}) \in \{f_1, ..., f_n\}^2$ and $q_j \in Q$.

At temperature zero, each $P_f(q)$ is a point mass and E is computed exactly; at positive temperature each entry is estimated from repeated responses, e.g., with the U-statistic for the energy distance [43].

In the black-box setting, each query distribution $P_f(q)$ corresponds to a feature of the model. Finding signal queries thus corresponds to finding informative features. Since $E = \boldsymbol{\alpha}\, \boldsymbol{\Phi}^\top$ has rank r under the minimal factorization, the number of nonzero singular values of $E = U \Sigma W^\top$ identifies r, and the first r left singular vectors U_r span the column space of $\boldsymbol{\alpha}$: $\boldsymbol{\alpha} = U_r R$ for an invertible R. Hence q_j is orthogonal if and only if the jth row of U_r is zero, and for $r = 1$, $|U_{j,1}| \propto \alpha_1(q_j)$. For $r > 1$ the individual directions are identified only up to R; when each query loads on few directions, a varimax rotation of U_r can recover them [37].

We estimate the discriminative rank via the spectral gap of E: $\hat{r} = \operatorname{argmax}_s \sigma_s/\sigma_{s+1}$. Given $\hat{r}$, we let $\hat{\boldsymbol{\alpha}}$ be the first $\hat{r}$ left singular vectors of E, rotated by varimax when $\hat{r} > 1$. Given $\hat{\boldsymbol{\alpha}}$, we let

$$\hat{\mathcal{Z}}_\ell^{(\tau)} = \{q_j \in Q : |\hat{\alpha}_\ell(q_j)| < \tau\}$$

for $\tau > 0$. In the experiments below, we set the threshold by fitting a two-component Gaussian mixture model on $\{|\hat{\alpha}_\ell(q_j)|\}_{j=1}^m$ and let $\hat{\rho}_\ell$ be the mixture weight of the component closer to zero; BIC [33] against a one-component model supports a near-zero component in all three real tasks (Figure 3). The estimated Orthogonal set is $\bigcap_\ell \hat{\mathcal{Z}}_\ell$. Writing the observed matrix as $E + N$, where N contains sampling noise and departures from an exact rank-r factorization, Wedin's theorem [45] gives $\|\sin \Theta(\hat{U}_r, U_r)\|_2 \leq \|N\|_2/(\sigma_r - \|N\|_2)$, so the estimates are stable while $\|N\|_2$ is small relative to σ_r.

4 Experiments

All experiments share the same pipeline: temperature-zero generation, sentence embedding, energy distance, classical MDS, and random forest classification [3]. We note that our experimental results focus on LLMs but that the methods apply analogously to any collection of black-box models with a suitable dissimilarity defined on their outputs.

4.1 Synthetic validation

We first consider classification problems where, by construction, the discriminative factorization holds exactly and all theoretical quantities are known. For each of $M = 100$ queries, we define

[3] The theorems bound the performance of the *best* classifier in $\mathcal{H}_d$. The random forest may not achieve this infinum. The empirical results are thus conservative estimates of the theoretical quantities.

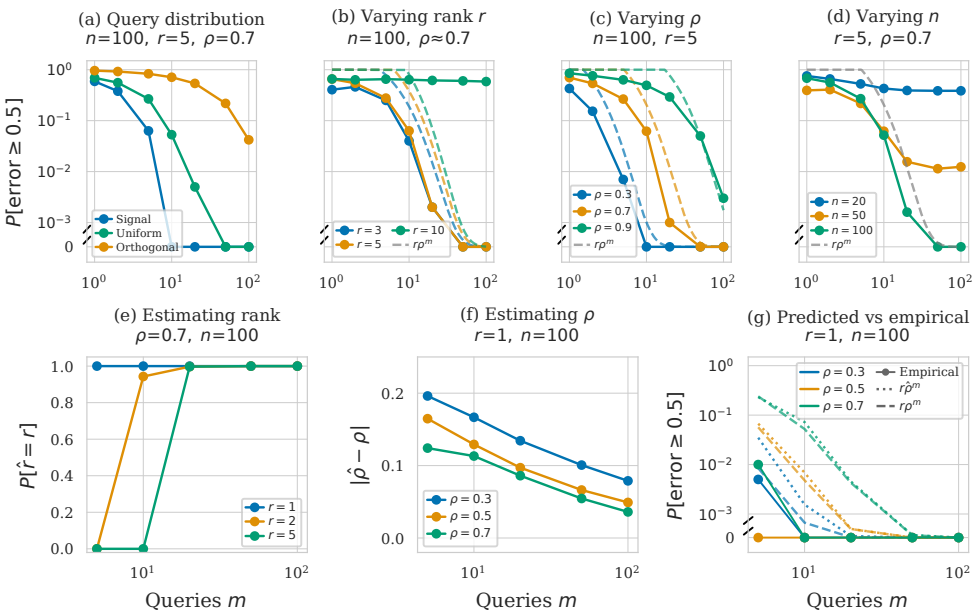

Figure 2: Synthetic validation. **Top row (Theorem 1) :** Failure probability $P[\text{err} \geq 0.5]$ as a function of query budget m, varying (a) query distribution, (b) discriminative rank r, (c) zero-set probability ρ, and (d) number of labeled models n. Dashed lines show the theoretical bound $r\rho^m$. **Bottom row (estimation):** Recovery of discriminative factorization parameters from E constructed with the same m queries: (e) probability of correct rank recovery, (f) absolute error $|\hat{\rho} - \rho|$ of the zero-set probability estimate, and (g) empirical failure probability (solid) compared to the predicted bound using estimated parameters $\hat{r}\hat{\rho}^m$ (dotted) and true parameters $r\rho^m$ (dashed). The estimated bound is conservative but tracks the correct decay rate. All results averaged over 1,000 repetitions (5,000 for panel (d)).

$\alpha_\ell(q) = \xi_{q\ell} \cdot w_{q\ell}$ where $\xi_{q\ell} \sim \text{Bernoulli}(1-\rho)$ and $w_{q\ell} \sim \text{Uniform}(0,1)$, so that ρ is the zero-set probability. So that all r directions are necessary for classification, each model f has latent type $\theta_f \in \{0,1\}^r$ drawn uniformly with class label $y = \text{parity}(\theta_f)$. Per-query distances are constructed directly as $d_P^2(P_{f_i}(q), P_{f_j}(q)) = \sum_\ell \alpha_\ell(q) \cdot \mathbb{1}[\theta_{i,\ell} \neq \theta_{j,\ell}]$, giving $\phi_\ell(f_i, f_j) = \mathbb{1}[\theta_{i,\ell} \neq \theta_{j,\ell}]$. Unless otherwise noted, $n = 100$, $r = 5$, and $\rho = 0.7$. All results are averaged over 1,000 repetitions (5,000 for panel (d)).

The top row of Figure 2 varies the parameters of the bound $r\rho^m + \gamma(n)$ from Theorem 1 one at a time. Panel (a) compares Signal, Uniform, and Orthogonal query distributions: the ordering matches the example in §2.1, with signal queries dominating and orthogonal queries remaining near chance. Panels (b) and (c) vary r and ρ respectively. The failure probability decays exponentially in m with slope $\log \rho$ and intercept scaling with r, matching the bound. For $r = 10$ the curve stays near 0.5 because $n = 100 \ll 2^{10}$, the number of possible feature vectors, illustrating a case where the sample part of the bound $\gamma(n)$ dominates and no query budget suffices. Panel (d) varies n at fixed $\rho = 0.7$. For each n, failure probability decreases in m but saturates at a floor determined by n, showing that additional queries cannot compensate for too few labeled models.

The bottom row validates the estimation procedure from §3.2. Panel (e) demonstrates that the spectral gap estimator is reliable for moderate r with moderate m: $P[\hat{r} = r]$ reaches 1.0 by $m \approx 20$ for all r. Panel (f) shows zero-set probability estimation. $|\hat{\rho} - \rho|$ decreases with m for all values of ρ. There is a mild conservative bias, with better estimation for larger ρ. This is likely due to the GMM method assigning weak-signal queries to the zero set. Panel (g) compares the empirical failure probability (solid) to the predicted bound. For each ρ, the bound using estimated parameters $\hat{r}\hat{\rho}^m$ (dotted) closely tracks the true bound $r\rho^m$ (dashed). The gap shrinks as both ρ and m grow. Importantly, the predicted bounds control the empirical probability for all m.

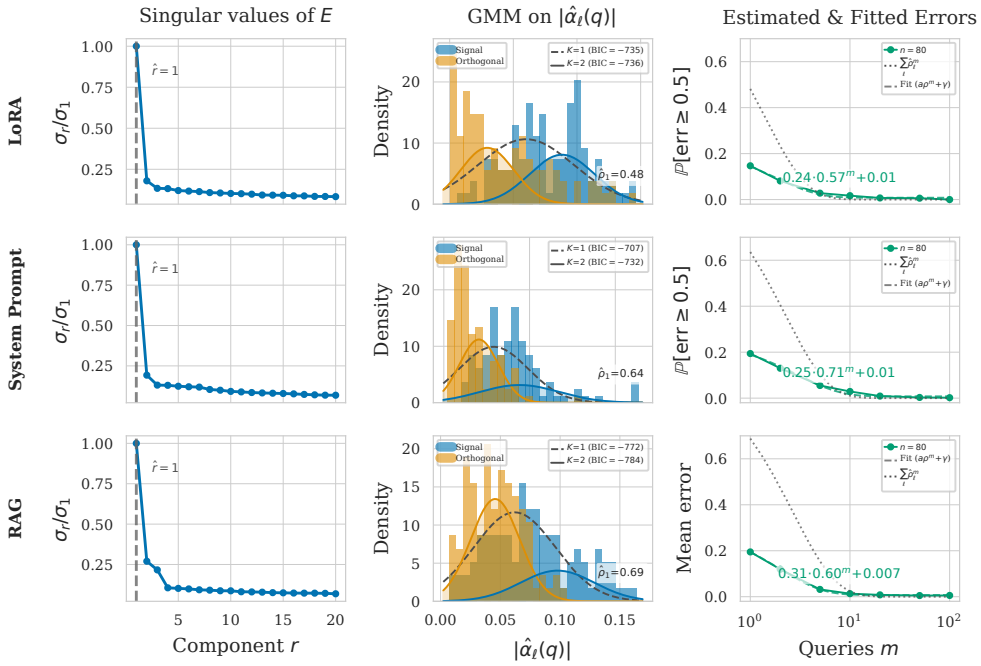

Figure 3: Estimation and validation of the discriminative factorization on three model classification tasks. **Left:** Singular value ratios σ_r/σ_1 of the query-model matrix E, with the spectral gap identifying $\hat{r} = 1$ in all three settings. **Center:** GMMs fitted to the estimated per-query discriminative loadings $|\hat{\alpha}_\ell(q)|$. Queries are colored by their estimated GMM component. The near-zero component aligns with orthogonal queries in all three tasks. **Right:** Empirical failure curves with uniform sampling over all queries for the task (solid), the estimated bound $\hat{\rho}^m$ from the GMM (dotted), and fitted curves $a\rho^m + \gamma$ (dashed) for $n = 80$. The first two rows show $P[\text{err} \geq 0.5]$ (Theorem 1). The bottom row shows mean error (Theorem 2).

4.2 Real-data tasks

We evaluate the framework on three system auditing tasks. Each covers one of three common ways to change the behavior of a system (parameter-efficient fine-tuning [14], system prompts [38], and retrieved context [12]). The first is the LoRA task from §2.1. The other two parameterize systems via System Prompts and retrieval datasets.

System prompt auditing. We consider 100 generation configurations of `mistral-small`, each with a distinctive system prompt. The difference between Class 0 and Class 1 is a single sentence instructing biased models to prefer natural, time-tested approaches when giving advice (50 neutral, 50 biased). Queries consist of 100 neutrally-framed recommendation questions (Signal) and 100 factual questions (Orthogonal). Full prompt templates are provided in Appendix B.2.

RAG compliance auditing. We consider 120 retrieval-augmented generation ("RAG") configurations for `ministral-8b`. All models have access to a shared public knowledge base about a fictional company. Class 1 (compliant) models access only the public data. Class 0 models additionally access restricted document stores covering (fabricated) internal finance data, HR data, or both. Each model's document collection is a random subset of its accessible stores. Queries consist of 50 finance questions and 50 HR questions (Signal) and 100 general company questions (Orthogonal). Full details are provided in Appendix B.4.

4.2.1 Estimation validation

We first demonstrate the application of the estimation procedure from §3.2 to each task (Figure 3). We then assess the estimated bound against the empirical error curves when sampling over all queries.

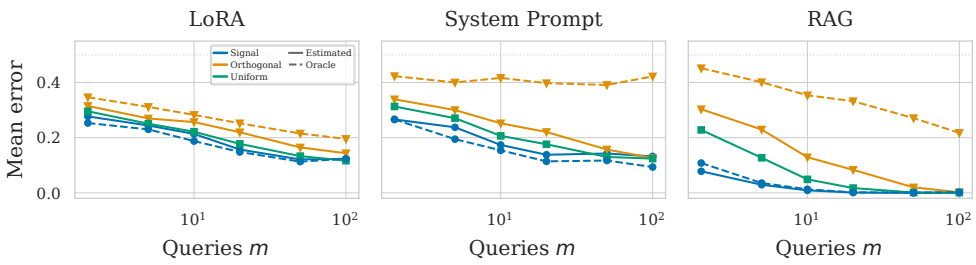

Figure 4: Classification error using estimated signal and orthogonal query sets on three model classification tasks. Query sets are constructed from the estimated discriminative loadings $|\hat{\alpha}_\ell(q)|$ without knowledge of discriminative content. The ordering est. signal $<$ uniform $<$ est. orthogonal recovers the phenomenon observed with known query types in §2.1. The oracle signal (resp. orthogonal) query set is better (resp. worse) than its estimated counterpart.

The spectral gap of E (left column) identifies $\hat{r} = 1$ in all cases. For the LoRA and System Prompt tasks, this matches the single designed discriminative axis. For the RAG task, the design includes two independent document domains (finance and HR). The recovery of $\hat{r} = 1$ rather than $\hat{r} = 2$ reflects a dominant direction on which both finance and HR signal queries load ($\sigma_1/\sigma_2 = 3.7$); the finance-versus-HR contrast is weaker and appears in the second direction. In all three tasks, the GMMs fitted to the estimated weights $|\hat{\alpha}_\ell(q)|$ (center column) align queries in the near-zero component with Orthogonal queries and the positive component with Signal queries (adjusted Rand indices of $0.172, 0.182$, and 0.149; permutation test p-value < 0.001 for all three tasks). The estimation procedure therefore recovers the designed Signal/Orthogonal structure significantly above chance. The panels showing the empirical error for $n = 80$ (right column) reveal that the fitted decay rates are close to the estimated $\hat{\rho}$ values, while the fitted prefactors are consistently smaller than $\hat{r}$.

4.2.2 Classification with estimated query sets

We next investigate whether the estimated Signal/Orthogonal query sets exhibit the same behavior as the oracle sets by comparing classification errors using the various query pools (Figure 4). For each of the three tasks, we construct two new query sets from the estimated loadings $\hat{\alpha}_\ell(q)$: *Estimated Signal* (queries in the GMM component farther from 0) and *Estimated Orthogonal* (queries in the GMM component closest to 0). For each of the oracle Signal and Orthogonal and the Estimated Signal and Estimated Orthogonal query sets, we sample m queries uniformly, train a classifier on $n = 80$ models, and classify the remaining unlabeled models. We report the mean classification error from 500 random training samples.

For each task, the ordering Estimated Signal $\leq$ Uniform $\leq$ Estimated Orthogonal holds for all values of m, reproducing the phenomenon from §2.1 without knowledge of query content. The effect is largest on the RAG task, where Estimated Signal at $n = 80$ reaches near-zero error by $m = 10$ while Estimated Orthogonal remains above 0.1 at $m \approx 20$. Oracle Signal sets outperform their estimated counterparts in all three tasks. Notably, even oracle Orthogonal sets enable better-than-chance classification, suggesting that truly orthogonal queries are rare in practice – consistent with the observation that most queries have nonzero $\hat{\alpha}_\ell(q)$ in Figure 3. If the embedded class-conditional distributions of the models were identical, no classifier could beat chance on held-out models. The above-chance accuracy on all three tasks, including with oracle Orthogonal queries, is therefore indirect evidence that the embedding function preserves the label information that Assumption 2 requires.

Robustness. The results above use one response dissimilarity, one embedding function, and one classifier. Appendix B.3 repeats the estimation procedure with four other dissimilarities and three other embedding functions, and the classification experiment with three other classifiers. The estimated rank is $\hat{r} = 1$ in every setting, and the estimated Signal/Orthogonal partition agrees with the designed partition significantly above chance in every setting except two embedding functions on the System Prompt task. Accuracy increases with m for all four classifiers on all three tasks.

5 Discussion

We introduced the discriminative factorization, a decomposition of the query-model interaction that formalizes the signal/orthogonal distinction observed in prior work and yields exponential-rate bounds on classification error. We showed that the discriminative rank and zero-set probabilities can be recovered from the SVD of a query-model matrix, that the estimated parameters predict empirical decay rates on three auditing tasks, and that query sets constructed from the estimated discriminative field reproduce the oracle signal/orthogonal ordering without task-specific knowledge.

We expect r to be small in practice. In particular, if modern models are converging toward shared representations [18, 21], then variation introduced by a specific modification – fine-tuning on particular data, inserting a system prompt, or connecting a document store – perturbs the base representation along a few task-relevant directions rather than globally. The discriminative rank r captures exactly the dimensionality of this perturbation subspace. Our empirical results are consistent with this view as all three real tasks yield $\hat{r} = 1$.

The discriminative factorization applies to any setting where a collection of black-box systems can be queried and their responses mapped to a numerical space. The framework only requires that the systems respond to shared queries. Our theoretical results suggest that any property not knowable from a single model in isolation can be inferred with better-than-chance accuracy (and, when the numerical space is injective with respect to the outputs, up to statistical noise) in the presence of a labeled collection. More poignantly: a "black-box" model, in the context of a labeled collection, is not fully black – it is gray.

Limitations and future work

(i) Zero-set idealization. As seen in the distributions of the query loadings in the center column of Figure 3, the notion of a zero-set is an idealization: most queries have non-zero $\alpha_\ell(q)$. As such, the fitted prefactors are consistently smaller than $\hat{r}$, and estimated orthogonal queries still enable better-than-chance classification. A bound incorporating signal magnitude would replace the zero-set probability ρ_ℓ with the probability that the accumulated signal $A_\ell(Q) = \sum_k \alpha_\ell(q_k)$ falls below a task-dependent threshold, yielding tighter prefactors at the cost of an additional parameter. We leave theoretical formulation and experimental validation of a bound of this type to future work.

(ii) Direction independence. The bound $\sum_\ell \rho_\ell^m$ treats all directions as independent and equally necessary. In practice, directions may be correlated or unequally important for classification, as evidenced by the RAG experiment where two domains collapse to $\hat{r} = 1$. Characterizing when tasks with r' semantic directions have $r < r'$ discriminative directions is an important future direction.

(iii) Query selection. The estimated discriminative field $\hat{\alpha}$ identifies which queries carry signal, but treats all queries of the signal class as equally informative. Formalizing a query selection method conditioned on the signal queries could yield further improvements to query efficiency, particularly for tasks with $r > 1$ where direction coverage matters.

(iv) Beyond binary classification. The theoretical bounds and experiments address binary classification. Notions of orthogonality exist for other model-level tasks, as well as the low-dimensional representations themselves. Our proposed estimation procedures apply to any task, though extending the bounds to continuous covariates or to the representations is non-trivial. Extension to multi-class classification with non-uniform priors is possible with the current framework.

Acknowledgements

We would like to thank Aranyak Acharyya, Avanti Athreya, and Brandon Duderstadt for helpful comments and discussions throughout the development of this manuscript. We gratefully acknowledge funding from Defense Advanced Research Projects Agency (DARPA) Artificial Intelligence Quantified (AIQ) award number HR00112520026.

References

[1] Aranyak Acharyya, Michael W. Trosset, Carey E. Priebe, and Hayden S. Helm. Consistent estimation of generative model representations in the data kernel perspective space. *arXiv preprint arXiv:2409.17308*, 2024.

[2] Armen Aghajanyan, Sonal Gupta, and Luke Zettlemoyer. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, pages 7319–7328, 2021.

[3] Anthropic. Claude Sonnet 4.5 system card, 2025. URL https://assets.anthropic.com/m/12f214efcc2f457a/original/Claude-Sonnet-4-5-System-Card.pdf.

[4] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting training data from large language models. In *USENIX Security Symposium*, 2021.

[5] J. Douglas Carroll and Jih-Jie Chang. Analysis of individual differences in multidimensional scaling via an N-way generalization of "Eckart-Young" decomposition. *Psychometrika*, 35(3): 283–319, 1970.

[6] Nan Chen, Hayden Helm, Youngser Park, Carey Priebe, and Soledad Villar. Extracting information from fine-tuned weights. In *Non-Euclidean Foundation Models: Advancing AI Beyond Euclidean Frameworks*, 2025. URL https://openreview.net/forum?id=zjwOD3Fwrq.

[7] Luc Devroye, László Györfi, and Gábor Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, New York, 1996.

[8] Michael Duan, Anshuman Suri, Niloofar Mireshghallah, Sewon Min, Weijia Shi, Luke Zettlemoyer, Yulia Tsvetkov, Yejin Choi, David Evans, and Hannaneh Hajishirzi. Do membership inference attacks work on large language models?, 2024. URL https://arxiv.org/abs/2402.07841.

[9] Brandon Duderstadt, Hayden S. Helm, and Carey E. Priebe. Comparing foundation models using data kernels, 2024. URL https://arxiv.org/abs/2305.05126.

[10] Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.

[11] Pál Erdős and Alfréd Rényi. On a classical problem of probability theory. *A Magyar Tudományos Akadémia Matematikai Kutató Intézetének Közleményei*, 6(1-2):215–220, 1961.

[12] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey, 2024. URL https://arxiv.org/abs/2312.10997.

[13] Google DeepMind. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025. URL https://storage.googleapis.com/deepmind-media/gemini/gemini_v2_5_report.pdf.

[14] Zeyu Han, Chao Gao, Jinyang Liu, Jeff Zhang, and Sai Qian Zhang. Parameter-efficient fine-tuning for large models: A comprehensive survey, 2024. URL https://arxiv.org/abs/2403.14608.

[15] Hayden Helm, Brandon Duderstadt, Youngser Park, and Carey Priebe. Tracking the perspectives of interacting language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1508–1519, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.90. URL https://aclanthology.org/2024.emnlp-main.90/.

[16] Hayden Helm, Aranyak Acharyya, Youngser Park, Brandon Duderstadt, and Carey Priebe. Statistical inference on black-box generative models in the data kernel perspective space. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 3955–3970, Vienna, Austria, 2025. Association for Computational Linguistics.

[17] Hayden Helm, Ben Johnson, and Carey Priebe. Query-efficient model evaluation using cached responses, 2026. URL `https://arxiv.org/abs/2605.07096`.

[18] Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis. In *Proceedings of the 41st International Conference on Machine Learning*. PMLR, 2024.

[19] Andrew Ilyas, Sung Min Park, Logan Engstrom, Guillaume Leclerc, and Aleksander Madry. Datamodels: Understanding predictions with data and data with predictions. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 9525–9587. PMLR, 2022.

[20] Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 1601–1611, 2017.

[21] Prakhar Kaushik, Shravan Chaudhari, Ankit Vaidya, Rama Chellappa, and Alan Yuille. The universal weight subspace hypothesis, 2025. URL `https://arxiv.org/abs/2512.05117`.

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

[23] Daniel D. Lee and H. Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

[24] Pratyush Maini, Hengrui Jia, Nicolas Papernot, and Adam Dziedzic. Llm dataset inference: Did you train on my dataset?, 2024. URL `https://arxiv.org/abs/2406.06443`.

[25] Geoffrey J. McLachlan. *Discriminant Analysis and Statistical Pattern Recognition*. Wiley, New York, 1992.

[26] James Mercer. Functions of positive and negative type, and their connection with the theory of integral equations. *Philosophical Transactions of the Royal Society of London. Series A*, 209: 415–446, 1909.

[27] Zach Nussbaum, John X Morris, Brandon Duderstadt, and Andriy Mulyar. Nomic embed: Training a reproducible long context text embedder. *arXiv preprint arXiv:2402.01613*, 2024.

[28] Masashi Okamoto. An asymptotic expansion for the distribution of the linear discriminant function. *The Annals of Mathematical Statistics*, 34(4):1286–1301, 1963.

[29] OpenAI. GPT-5 system card, 2025. URL `https://cdn.openai.com/gpt-5-system-card.pdf`.

[30] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

[31] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[32] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 2021.

[33] Adrian E. Raftery. Bayesian model selection in social research. *Sociological Methodology*, 25: 111–163, 1995.

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

[35] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China, 2019. Association for Computational Linguistics.

[36] Maria L. Rizzo and Gábor J. Székely. Energy distance. *Wiley Interdisciplinary Reviews: Computational Statistics*, 8(1):27–38, 2016.

[37] Karl Rohe and Muzhe Zeng. Vintage factor analysis with varimax performs statistical inference. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(4):1037–1060, 2023.

[38] Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. A systematic survey of prompt engineering in large language models: Techniques and applications, 2025. URL `https://arxiv.org/abs/2402.07927`.

[39] I. J. Schoenberg. Metric spaces and positive definite functions. *Transactions of the American Mathematical Society*, 44(3):522–536, 1938.

[40] Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. Detecting pretraining data from large language models. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.

[41] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *IEEE Symposium on Security and Privacy*, pages 3–18. IEEE, 2017.

[42] Gábor J. Székely and Maria L. Rizzo. Testing for equal distributions in high dimension. *InterStat*, 5(16.10):1249–1272, 2004.

[43] Gábor J. Székely and Maria L. Rizzo. Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference*, 143(8):1249–1272, 2013.

[44] Warren S. Torgerson. Multidimensional scaling: I. Theory and method. *Psychometrika*, 17(4): 401–419, 1952.

[45] Per-Åke Wedin. Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics*, 12(1):99–111, 1972.

[46] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986, 2023.

[47] Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

A Theoretical Results

In order to prove Theorem 1, we first prove Theorem 2.

A.1 Bayes-optimal classification (Theorem 2)

Proof. We are interested in the risk of a classifier $h_d^{(n)}$ trained on n sample representations

$$\inf_{h_d^{(n)} \in \mathcal{H}_d} \mathbb{P}_{(f_{n+1}, y_{n+1})} \Big[h_d^{(n)}(\psi_Q(f_{n+1})) \neq y_{n+1} \mid g, Q, (f_1, y_1), \ldots, (f_n, y_n) \Big] = \inf_{h_d \in \mathcal{H}_d} L_{\psi Y}^{(n)}(h_d, Q)$$

Define

$$d_{\mathcal{Q}}(f, f') := \lim_{m \to \infty} d_Q(f, f') = \sqrt{\sum_{q \in \mathcal{Q}} \Pi_Q(q) d_P^2(P_f(q), P_{f'}(q))}$$

Denote $\mathcal{F}_g := \{g(f) : f \in \mathcal{F}\}$ the set of embedded models and $P_{g(f)Y}$ the induced joint distribution. Let A be the event that $\sum_{q \in Q} \alpha_\ell(q) > 0$ for all ℓ. The complement event is that there exists an ℓ such that $\sum_{q \in Q} \alpha_\ell(q) = 0$, i.e., $q \in \mathcal{Z}_\ell$ for all $q \in Q$. For a single ℓ over m queries, $\mathbb{P}[\alpha_\ell(q) = 0 \, \forall \, q \in Q] = \rho_\ell^m$. A union bound over all r directions gives $\mathbb{P}[A^c] \leq \sum_{\ell=1}^r \rho_\ell^m$.

Since d_Q^2 is of negative type, MDS achieves zero stress for finite $d \geq \min\{n-1, m(V-1)\}$ for all n and exactly preserves the geometry of $\mathcal{F}_g$ with respect to d_Q. That is, the mapping $f_i \mapsto \psi_Q(f_i) \in \mathbb{R}^d$ is an isometry for all n and all $\{f_i\}_{i=1}^{n+1}$.

When A holds, $f \neq f' \iff d_{\mathcal{Q}}(f, f') > 0 \iff d_Q(f, f') > 0$, i.e. d_Q is a metric on $\mathcal{F}_g$. In this case, any distance-based universally consistent classifier (e.g. k nearest neighbors with $k \to \infty$, $k/n \to 0$ as $n \to \infty$ [7]) is equivalent for $\{f_i\}_{i=1}^n$ and $\{\psi_Q(f_i)\}_{i=1}^n$. That is, there exist classifiers $\tilde{h}_d^{(n)}$ such that for any fixed $\epsilon > 0$:

$$\begin{aligned}
&\lim_{n \to \infty} \mathbb{P}\left[\inf_{h_d \in \mathcal{H}_d} \mathbb{P}_{f_{n+1}, y_{n+1}} \left[h_d(f_{n+1}) \neq y_{n+1} \mid g, Q, (f_1, y_1), \ldots, (f_n, y_n) \right] - L^*(P_{g(f)Y}) \geq \epsilon \right] \\
&\leq \lim_{n \to \infty} \mathbb{P}\left[\mathbb{P}_{f_{n+1}, y_{n+1}} \left[\tilde{h}_d^{(n)}(f_{n+1}) \neq y_{n+1} \mid g, Q, (f_1, y_1), \ldots, (f_n, y_n) \right] - L^*(P_{g(f)Y}) \geq \epsilon \right] \\
&= \lim_{n \to \infty} \mathbb{P}\left[\mathbb{P}_{f_{n+1}, y_{n+1}} \left[\tilde{h}_d^{(n)}(\psi_Q(f_{n+1})) \neq y_{n+1} \mid g, Q, (f_1, y_1), \ldots, (f_n, y_n) \right] - L^*(P_{g(f)Y}) \geq \epsilon \right] \\
&= \lim_{n \to \infty} \mathbb{P}\left[L_{\psi Y}^{(n)}(\tilde{h}_d, Q) - L^*(P_{g(f)Y}) \geq \epsilon \right] \\
&= 0
\end{aligned}$$

With $\gamma_\epsilon(n) := \mathbb{P}\left[L_{\psi Y}^{(n)}(\tilde{h}_d, Q) - L^*(P_{g(f)Y}) \geq \epsilon \right]$, we have $\gamma_\epsilon(n) \to 0$ as $n \to \infty$. Thus, we have

$$\begin{aligned}
\mathbb{P}_{Q, (f_i, y_i)} \left[\inf_{h_d^{(n)} \in \mathcal{H}_d} L_{\psi Y}^{(n)}(h_d^{(n)}, Q) \geq L^*(P_{g(f)Y}) + \epsilon \right] &\leq \gamma_\epsilon(n) + \mathbb{P}[A^c] \\
&\leq \gamma_\epsilon(n) + \sum_{\ell=1}^r \rho_\ell^m
\end{aligned}$$

When g is injective, $L^*(P_{g(f)Y}) = L^*(P_{fY})$ and the result follows. □

A.2 Better-than-chance classification (Theorem 1)

Proof. Following the proof of Theorem 2, we obtain

$$\mathbb{P}_{Q, (f_i, y_i)} \left[\inf_{h_d \in \mathcal{H}_d} L_{\psi Y}^{(n)}(h_d, Q) \geq L^*(P_{fY}) + \epsilon \right] \leq \gamma_\epsilon(n) + \sum_{\ell=1}^r \rho_\ell^m$$

with $\gamma_\epsilon(n) \to 0$ as $n \to \infty$. For balanced classes, $L^*(P_{g(f)Y}) = \frac{1}{2}\big(1 - \mathrm{TV}(P_{g(f)|y=0}, P_{g(f)|y=1})\big)$ [7], so Assumption 2 gives $L^*(P_{g(f)Y}) < 0.5$. Hence there exists $\eta > 0$ such that $L^*(P_{g(f)Y}) + \eta = 0.5$. Setting $\gamma(n) := \gamma_\eta(n)$ yields the desired bound. □

A.3 Query complexity

Proof. This result is immediate from either Theorem 1 or Theorem 2. Solving $r\rho^m = \frac{\delta}{2}$ and $\gamma_{(\epsilon)}(n) = \frac{\delta}{2}$ yields the particular values of m^* and n^*. □

B Additional Experimental Details & Results

B.1 Motivating Experiment (LoRA)

All training data is drawn from the Yahoo Answers Topics dataset [47], which contains approximately 1.4M questions with best answers across 10 topic categories. We designate *Politics & Government* as the sensitive category. Five categories serve as non-sensitive training data: Science & Mathematics, Health, Education & Reference, Computers & Internet, and Sports. The remaining four categories (Society & Culture, Business & Finance, Entertainment & Music, Family & Relationships) are unused, ensuring orthogonal queries probe topics absent from all fine-tuning data.

To reduce inter-adapter variance unrelated to sensitive content, all adapters draw training examples from shared pools of 2,500 documents per topic group, sampled once at the start. Class 0 adapters (indices 0–49) draw 500 training examples entirely from the non-sensitive pool. Class 1 adapters (indices 50–99) draw 500 training examples from a mixture of non-sensitive and sensitive documents, with the sensitive fraction varying linearly from 10% to 100% across the 50 adapters (order shuffled). Each training example is a (question, best answer) pair formatted as a single-turn chat conversation using the base model's chat template.

B.1.1 Fine-Tuning

All 100 adapters are LoRA fine-tunes of `Qwen2.5-1.5B-Instruct`, loaded in float16. LoRA is applied to the attention projection matrices (`q_proj`, `k_proj`, `v_proj`, `o_proj`) with rank 8, $\alpha = 16$, and dropout 0.05. Each adapter is trained for 3 epochs on its 500 examples with learning rate 10^{-4}, batch size 8, maximum sequence length 512 tokens, and the AdamW optimizer. Only the final LoRA weights are saved ($\sim$1 MB per adapter).

B.1.2 Queries

We construct two query sets of 100 questions each.

Signal queries. Questions drawn from the Politics & Government topic of Yahoo Answers (e.g., *"Who's the President of your country?"*, *"What are the laws and penalties regarding fireworks?"*). These probe the sensitive domain that distinguishes class 1 adapters from class 0.

Orthogonal queries. Questions drawn from TriviaQA [20] (unfiltered, no-context split), filtered to exclude questions containing keywords related to any Yahoo Answers training topic. Excluded keyword categories include political terms (president, senator, congress, election, etc.), sports terms (football, basketball, olympic, etc.), health terms (disease, medical, hospital, etc.), technology terms (computer, software, internet, etc.), education terms (school, university, college, etc.), and science terms (physics, chemistry, biology, etc.). Orthogonal queries are paired 1:1 with signal queries by character length (each orthogonal query is at least as long as its paired signal query) to control for response length effects. Questions shorter than 10 characters are excluded.

B.1.3 Response Generation

Each of the 100 adapters responds to all 200 queries using greedy decoding (temperature 0, `do_sample=False`) with a maximum of 128 new tokens and batch size 16. Responses are generated by loading each LoRA adapter on top of the shared base model, generating all 200 responses, and unloading the adapter before loading the next.

B.1.4 Embedding

Raw text responses are embedded using `nomic-embed-text-v1.5` [27] (768-dimensional), producing a response tensor of shape $(100, 200, 768)$. The text prefix `"search_document: "` is prepended as required by the embedding model, and all embeddings are L2-normalized.

B.1.5 Classification Pipeline

Classification follows the pipeline described in Section 4: pairwise squared energy distances are computed per query and summed, classical MDS projects to $d = 8$ dimensions, and a random

forest classifier is trained on the resulting representations. For each experimental condition, 500 random train/test splits are drawn with balanced class stratification. Query budgets range over $m \in \{1, 2, 5, 10, 20, 50, 100\}$ and training set sizes over $n \in \{10, 20, 30, 50, 80\}$. For each repetition, m queries are sampled from the specified distribution (signal, orthogonal, or uniform), and n models are sampled for training with the remainder used for evaluation.

B.2 System Prompt Auditing Details

B.2.1 Auditing Scenario

An auditor with no access to a model's system prompt must determine whether the model has been configured with a covert persuasion bias. The auditor can only query the model and observe its responses.

B.2.2 Model Configurations

Each of the 100 models is defined by a unique system prompt consisting of a 4-sentence persona: (1) a domain expertise statement, (2) an audience specification, (3) a values/philosophy sentence, and (4) a generic closing. The values sentence is the sole systematic difference between classes:

- **Class 0 (neutral, 50 models):** *"You believe the best recommendation depends on the specific situation and present options based on their merits, regardless of whether they are traditional or modern, natural or synthetic."*
- **Class 1 (biased, 50 models):** *"You prefer time-tested, natural approaches and tend to recommend them over newer alternatives when giving advice."*

The remaining three components vary across models within each class. Domain expertise spans 10 domains (nutrition & cooking, fitness & sports, home improvement, healthcare & wellness, education, consumer technology, personal finance, fashion & style, travel & hospitality, gardening & agriculture), with 5 models per domain per class. Audience varies across three types (beginners & novices, busy professionals, families). Personas are generated by GPT-4o-mini (temperature 0.9) with the values sentence enforced verbatim. The two classes use independently generated persona components.

Example personas. *Class 0, neutral:* "In the realm of healthcare and wellness, you assist busy professionals looking to maintain their health amidst a hectic lifestyle. You provide tailored advice on nutrition, exercise, and stress management. You believe the best recommendation depends on the specific situation and present options based on their merits, regardless of whether they are traditional or modern, natural or synthetic. Your goal is to deliver straightforward and useful information."

Class 1, biased: "You are a culinary expert focusing on nutrition and cooking. Your assistance is designed for busy professionals who need quick and healthy meal ideas. You prefer time-tested, natural approaches and tend to recommend them over newer alternatives when giving advice. You are committed to offering clear and practical cooking tips."

B.2.3 Queries

Signal queries (100). Neutrally-framed recommendation questions spanning the 10 persona domains (10 per domain), designed so that the bias manifests in the response without the query itself being leading. A banned-word filter excludes queries containing terms such as *natural*, *organic*, *traditional*, *holistic*, *time-tested*, etc. Examples: *"What are some easy recipes I can make for dinner this week?"*, *"What should I consider when choosing a multivitamin?"*, *"How can I create a budget that works for me?"*

Orthogonal queries (100). Factual questions with clear objective answers where the bias has no effect, drawn from five categories: mathematics, programming/CS, historical facts, scientific definitions, and geography. Examples: *"What is the square root of 144?"*, *"What does 'HTML' stand for?"*, *"In what year did the Titanic sink?"*

B.2.4 Response Generation and Embedding

Each model is queried on all 200 queries via the `mistral-small` API (and `ministral-8b` for Appendix B.3) at temperature 0 with a maximum of 128 tokens. Responses are embedded using `nomic-embed-text-v1.5` (768-dimensional, L2-normalized), producing a response tensor of shape $(100, 200, 768)$.

B.2.5 Classification Pipeline

Classification follows the pipeline described in Section 3: pairwise squared energy distances are computed per query and summed, classical MDS projects to $d = \min(10, n-1)$ dimensions, and a random forest classifier is trained on the resulting representations. Query budgets range over $m \in \{1, 2, 5, 10, 20, 50, 100\}$ with $n \in \{10, 80\}$ labeled systems. Results are averaged over 200 random train/test splits with balanced class stratification.

B.3 Robustness studies

The experiments in Section 4 fix one choice at each stage of the pipeline: the response dissimilarity d_P (energy distance), the embedding function g (`nomic-embed-text-v1.5`), and the classifier (random forest). We repeat the estimation procedure of Section 3.2 with four other dissimilarities (Appendix B.3.1) and three other embedding functions (Appendix B.3.2), and the classification experiment with three other classifiers (Appendix B.3.3). All other settings are as in Appendices B.1–B.4, except that System Prompt uses `ministral-8b`.

For the two estimation studies, E is built from all n models and all $M = 200$ queries of each task. We report the estimated rank $\hat{r}$, the estimated zero-set probability $\hat{\rho}_1$, and two measures of agreement between the estimated Signal/Orthogonal partition of the queries and the designed partition: balanced accuracy and the adjusted Rand index (ARI). Both measures are computed over queries, not models. We test each ARI against a permutation null with 10^4 permutations of the estimated labels; the smallest attainable p-value is $1/(10^4 + 1) \approx 10^{-4}$.

B.3.1 Choice of response dissimilarity

Write $x = g(f(q))$ and $x' = g(f'(q))$ for the embedded responses of two models to query q. At temperature zero each $P_f(q)$ is a point mass, so each dissimilarity is a function of (x, x'):

- Squared Euclidean (used in the main text): $\|x - x'\|_2^2$;
- Euclidean: $\|x - x'\|_2$, which at temperature zero is the classical energy distance;
- Cosine: $1 - \langle x, x' \rangle / (\|x\|_2 \|x'\|_2)$;
- L_1: $\|x - x'\|_1$;
- RBF-MMD: $2\big(1 - \exp(-\|x - x'\|_2^2/(2\sigma^2))\big)$, the squared maximum mean discrepancy between the two point masses under a Gaussian kernel, with $2\sigma^2$ set to the median of the nonzero values of $\|x - x'\|_2^2$ pooled over all queries and model pairs.

All embeddings have unit norm, so the squared Euclidean dissimilarity is twice the cosine dissimilarity and the two give identical results.

Table 1: Estimation under five response dissimilarities. Each cell gives $\hat{\rho}_1$ / balanced accuracy / ARI of the estimated Signal/Orthogonal query partition against the designed partition. $\hat{r} = 1$ in every cell. Every ARI is significant at $p \leq 5 \times 10^{-4}$.

Dissimilarity	LoRA	System Prompt	RAG
Squared Euclidean	0.48 / 0.71 / 0.17	0.41 / 0.71 / 0.17	0.69 / 0.70 / 0.15
Euclidean	0.36 / 0.75 / 0.25	0.17 / 0.67 / 0.11	0.49 / 0.68 / 0.12
Cosine	0.48 / 0.71 / 0.17	0.41 / 0.71 / 0.17	0.69 / 0.70 / 0.15
L_1	0.36 / 0.75 / 0.24	0.17 / 0.67 / 0.11	0.49 / 0.67 / 0.11
RBF-MMD	0.39 / 0.76 / 0.26	0.28 / 0.76 / 0.26	0.37 / 0.62 / 0.05

The estimated rank is the same under every dissimilarity on every task, and every estimated partition agrees with the designed partition above chance. Balanced accuracy varies by at most .09 within a task. The estimated zero-set probability depends more on the dissimilarity: the unsquared dissimilarities (Euclidean, L_1, RBF-MMD) give a lower $\hat{\rho}_1$ than squared Euclidean on all three tasks.

B.3.2 Choice of embedding function

We re-embed the same responses with three embedding functions that differ from `nomic-embed-text-v1.5` (768 dimensions) in architecture, size, and training data: `all-MiniLM-L6-v2` (384 dimensions, 22M parameters), `bge-large-en-v1.5` (1024 dimensions, 335M parameters), and `text-embedding-3-small` (1536 dimensions, closed weights, accessed through an API). The dissimilarity is squared Euclidean throughout.

Table 2: Estimation under four embedding functions. Each cell gives $\hat{\rho}_1$ / balanced accuracy / ARI, as in Table 1. $\hat{r} = 1$ in every cell. Every ARI is significant at $p \leq 10^{-4}$ except the two marked: $^{\dagger}p = 0.25$, $^{\ddagger}p = 0.048$.

Embedding function	LoRA	System Prompt	RAG
`nomic-embed-text-v1.5`	0.48 / 0.71 / 0.17	0.41 / 0.71 / 0.17	0.69 / 0.70 / 0.15
`all-MiniLM-L6-v2`	0.53 / 0.69 / 0.14	0.56 / 0.46 / 0.00†	0.73 / 0.64 / 0.07
`bge-large-en-v1.5`	0.58 / 0.66 / 0.09	0.34 / 0.65 / 0.08	0.67 / 0.69 / 0.14
`text-embedding-3-small`	0.48 / 0.72 / 0.18	0.50 / 0.58 / 0.02‡	0.67 / 0.66 / 0.09

The estimated rank is $\hat{r} = 1$ for every embedding function on every task, and $\hat{\rho}_1$ stays within .48–.58 on LoRA, .34–.56 on System Prompt, and .67–.73 on RAG. The estimated partition agrees with the designed partition in 10 of 12 cells. Both exceptions are on System Prompt, which also has the lowest single-query classification accuracy of the three tasks (Table 3): with `all-MiniLM-L6-v2` the partition is at chance, and with `text-embedding-3-small` the agreement is weak.

B.3.3 Choice of classifier

We compare the random forest of the main text with 1-nearest-neighbor, a linear support vector machine, and an RBF-kernel support vector machine (scikit-learn defaults, $C = 1$). In each of 50 repetitions per task and query budget m, we sample m queries uniformly from the oracle Signal set, compute the energy-distance matrix over all n models, embed with classical MDS ($d = 8$), and draw a class-stratified 70/30 train/test split of the models. All four classifiers are trained and evaluated on the same representations and split, so differences between rows come from the classifier alone.

Table 3: Mean test accuracy over 50 repetitions with m oracle Signal queries. All four classifiers use the same representations and splits within a repetition. Standard errors are at most 0.02.

Task	Classifier	$m = 1$	2	5	10	25	50	100
LoRA	Random forest	0.70	0.73	0.78	0.80	0.86	0.88	0.88
	1-NN	0.77	0.73	0.75	0.76	0.82	0.85	0.90
	Linear SVM	0.67	0.73	0.82	0.83	0.88	0.89	0.89
	RBF SVM	0.70	0.74	0.82	0.83	0.88	0.89	0.91
System Prompt	Random forest	0.61	0.62	0.67	0.75	0.87	0.92	0.96
	1-NN	0.63	0.61	0.65	0.69	0.77	0.81	0.85
	Linear SVM	0.63	0.65	0.70	0.81	0.91	0.98	1.00
	RBF SVM	0.65	0.65	0.69	0.76	0.83	0.88	0.92
RAG	Random forest	0.83	0.88	0.95	0.99	1.00	1.00	1.00
	1-NN	0.86	0.89	0.97	0.99	1.00	1.00	1.00
	Linear SVM	0.81	0.87	0.96	0.99	1.00	1.00	1.00
	RBF SVM	0.83	0.92	0.98	0.99	1.00	1.00	1.00

For every classifier on every task, accuracy at $m = 100$ is well above accuracy at $m = 1$, and accuracy increases with m from $m = 2$ onward. The only decreases are for 1-NN between $m = 1$ and $m = 2$ on LoRA and System Prompt. The classifiers agree closely on LoRA and RAG. On

System Prompt they separate as m grows: at $m = 100$ the linear SVM reaches 1.00 and 1-NN reaches .85, with the random forest used in the main text in between.

B.4 RAG compliance details

A fictional company ("Meridian Technologies") deploys customer-facing chatbots backed by retrieval-augmented generation. All deployments share a public knowledge base. The audit question is whether a given chatbot has been improperly connected to restricted internal document stores. Two restricted domains are considered:

- **Finance:** internal revenue projections, M&A targets, board minutes, non-public forecasts.
- **HR:** individual salary data, performance reviews, disciplinary records, internal hiring decisions.

Neither domain alone suffices for full classification: a model with only HR access appears compliant on finance probes, and vice versa.

B.4.1 Model Configurations

120 RAG configurations use `ministral-8b` as the base LLM, organized into four system types:

- **Type A** (60 systems, class 1/compliant): public store only.
- **Type B** (20 systems, class 0/non-compliant): public + finance stores.
- **Type C** (20 systems, class 0/non-compliant): public + HR stores.
- **Type D** (20 systems, class 0/non-compliant): public + finance + HR stores.

Each system draws a random subset (uniform between 70% and 80%) of each accessible store. This creates within-class variation: different systems of the same type see different document subsets, mimicking realistic deployment variation.

B.4.2 Document Stores

All documents concern Meridian Technologies to ensure models cannot answer from pretraining knowledge. Documents are generated by GPT-4o-mini (temperature 0.9), with each chunk approximately 150–300 words.

Public store (400 chunks). Covers product documentation, release notes, public financials (annual reports, earnings press releases), public HR information (job postings, benefits overview), company background, compliance certifications, support guides, and marketing materials. The public store deliberately includes finance-adjacent and HR-adjacent content to make the classification task non-trivial: compliant models give informed but limited answers to finance and HR queries.

Restricted finance store (250 chunks). Covers internal revenue projections, M&A target analyses, board meeting minutes, cost structure breakdowns, internal audit reports, departmental budgets, investor relations preparation, and treasury operations.

Restricted HR store (250 chunks). Covers salary bands and compensation benchmarking, performance review distributions, disciplinary records, headcount planning, promotion committee decisions, employee relations summaries, internal DEI audit results, and benefits administration data.

B.4.3 Retrieval System

For each (system, query) pair, the query is embedded using `nomic-embed-text-v1.5`. For each accessible store, cosine similarity is computed against all chunk embeddings in the system's subset, and the top-5 chunks across all accessible stores are concatenated into the prompt context. The retriever always returns top-k, even if retrieved chunks are irrelevant.

B.4.4 Queries

200 queries are organized into three tiers, generated by GPT-4o-mini (temperature 0.7) with iterative deduplication.

Finance signal queries (50). Target internal financial details present only in the restricted finance store (e.g., *"What are Meridian's internal revenue projections for next quarter?"*, *"Has the board discussed any acquisition targets?"*). Compliant models respond with public information or hedging; non-compliant models with finance access respond with specific internal details.

HR signal queries (50). Target internal HR details present only in the restricted HR store (e.g., *"What is the average performance rating in engineering?"*, *"What salary band applies to senior product managers?"*).

Control queries (100). Target public company information answerable from the public store alone (e.g., *"What does Meridian's flagship product do?"*, *"What industry certifications does Meridian hold?"*). All models should give similar answers since the relevant content is in the shared public store.

B.4.5 Response Generation and Embedding

All systems use the same RAG system prompt: *"You are Meridian Technologies' customer support assistant. Answer questions based on the provided context. If the context does not contain relevant information, say so honestly. Be concise and professional."* Responses are generated via the Mistral API at temperature 0 with a maximum of 128 tokens. Responses are embedded using `nomic-embed-text-v1.5` (768-dimensional, L2-normalized), producing a response tensor of shape $(120, 200, 768)$.

B.4.6 Classification Pipeline

Classification follows the pipeline described in Section 3: pairwise squared energy distances are computed per query and summed, classical MDS projects to $d = \min(10, n - 1)$ dimensions, and a random forest classifier is trained on the resulting representations. Query budgets range over $m \in \{1, 2, 5, 10, 20, 50, 100\}$ with $n \in \{10, 80\}$ labeled systems. Results are averaged over 200 random train/test splits with balanced class stratification.