Identifying Efficient Queries for Black-Box Model Classification

Merrick Ohata [1] **Carey Priebe** [1] **Hayden Helm** [2]

Abstract

Modern generative model use is often constrained to black box access. Recent work demonstrates that black box outputs for a set of input queries yield vector representations that are effective for inference of model-level properties. The quality of these representations depends heavily on the selection of input queries. In the context of model classification, we formalize this effect using the *discriminative factorization*, a decomposition of query-based model interaction into independent statistical "directions". Under this framework, it is possible to separate informative and uninformative queries using parameters that can be estimated from the spectral structure of a query-model matrix. For a real model auditing task, we demonstrate that query sets selected using the estimated discriminative factorization reproduce oracle ordering and improve classification efficiency without requiring task-specific knowledge.

1. Introduction

A black-box model is a system where users submit queries and receive responses with no access to model states, generation parameters, or (pre-)training procedures. These models may have a property of interest that is not directly observable, such as the presence of specific training data (Shokri et al., 2017; Carlini et al., 2021) or a behavioral bias (Caliskan et al., 2017; Bai et al., 2024). As models continue growing and providers increasingly restrict access to API endpoints (OpenAI, 2025; Anthropic, 2025; Google DeepMind, 2025), the black-box setting reflects practical model access and computational constraints. For a collection of models where the property of interest is known, our goal is to predict this property for a new model.

[1]Department of Applied Mathetmatics and Statistics, Johns Hopkins Universiy, Baltimore, USA [2]Helvian, California, USA. Correspondence to: Merrick Ohata <mohata1@jh.edu>, Hayden Helm <hayden@helvian.io>.

Accepted to the 1st Workshop on Combining Theory and Benchmarks, CTB@ICML 2026, Seoul, South Korea. Copyright 2026 by the author(s).

We approach this by constructing low-dimensional representations of black-box models. Other approaches such as weight-space methods construct model representations from internal activations (Duderstadt et al., 2023; Huh et al., 2024; Horwitz et al., 2025) or weights directly (Chen et al., 2025), but require access to model internals. We work in a true black-box setting, where model internals are inaccessible.

Other black-box analyses focus on individual models. Membership inference (Shokri et al., 2017; Duan et al., 2024), training data extraction (Carlini et al., 2021), and pretraining data detection (Shi et al., 2024; Maini et al., 2024) test whether specific examples influenced a single model, typically requiring token-level log-probabilities. Distribution testing (Canonne, 2020) studies the query complexity of distinguishing distributions from samples of a single source. We address a *population* of models.

Recent work has shown that black-box models can be mapped into Euclidean space via multidimensional scaling by sketching their behavior with respect to a set of queries (Helm et al., 2024). These representations come with desirable statistical properties, such as use in consistent (Acharyya et al., 2024; 2025a; Helm et al., 2025) and efficient (Helm et al., 2026) model-level inference.

Current theoretical results treat all queries as equally informative for prediction. In practice, the information contributed varies widely. In particular, query sets "aligned" with the inference task produce representations with substantially more discriminative power for a fixed query budget than those "not aligned" (Helm et al., 2025; 2024).

Contributions. In this paper, we introduce the *discriminative factorization* to formalize the notion of an (un)informative query, recover estimates of the discriminative factorization from the spectral properties of the *query-model matrix*, and demonstrate that query sets selected using this framework successfully improve model-level classification efficiency without task-specific knowledge.

2. Motivation and Setting

Let $\mathcal{F} := \{f, f : \mathcal{Q} \to \mathcal{X}\}$ be the space of black-box generative models. Each $f \in \mathcal{F}$ is a random mapping from finite input space $\mathcal{Q}$ with $|\mathcal{Q}| = M < \infty$ to 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". For a fixed embedding function $g : \mathcal{X} \to \mathbb{R}^p$ and model f, we let $P_f(q)$ denote the embedded response distribution associated with q. The query distribution Π_Q over $\mathcal{Q}$ is sampled independently m times to create the query multiset Q that we use to access models.

2.1. Constructing Euclidean Model Representations

In the black box setting, we say $f \neq f'$ precisely when there exists $q \in \mathcal{Q}$ such that $P_f(q) \neq P_{f'}(q)$. Thus, to capture dissimilarity between models we define

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

Where d_P^2 is a metric on $\mathcal{P}_1(\mathbb{R}^p)$ of negative type (the *energy distance* (Székely & Rizzo, 2013; Rizzo & Székely, 2016) is one such metric). d_Q is then a negative type metric on $\mathcal{P}_1(\mathbb{R}^p)^m$ (Schoenberg, 1938).

For a collection of n black box models $f_1, \ldots, f_n$ and multiset of m queries Q, we define the $n \times n$ pairwise distance matrix $D := D_{i,i'} = d_Q(f_i, f_{i'})$. We then apply classical multidimensional scaling (MDS) (Torgerson, 1952) 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 Q. We let Ψ_Q denote the matrix with ith row equal to $\psi_Q(f_i)$.

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 is negative type, MDS achieves zero stress for any $n < \infty$ with $d \geq \min\{n-1, m(V-1)\}$ (Székely & Rizzo, 2004; Schoenberg, 1938).

2.2. Motivating Example: Sensitive Data

To demonstrate the effect of query set on inference, we begin by examining a task inspired by Helm et al. (2025): Can we detect whether a particular topic appeared in a model's fine-tuning data, using only its generated outputs?

To investigate, we first fine-tune 100 LoRA adapters for `Qwen2.5-1.5B-Instruct` on subsets of Yahoo Answers (Zhang et al., 2015): 50 on non-sensitive topics (class 0) and 50 on mixtures containing "Politics & Government" data (class 1). We then consider model representations for three different input sets $\mathcal{Q}$: *Signal* queries directly concerning "Politics & Government"; *Orthogonal* queries on topics in none of the fine-tuning datasets; and the union of both (*Uniform*).

For each of the Signal, Orthogonal, and Uniform, we sample m queries uniformly at random and construct Euclidean model representations from temperature-zero responses. Following Chen et al. (2022), we use the second elbow

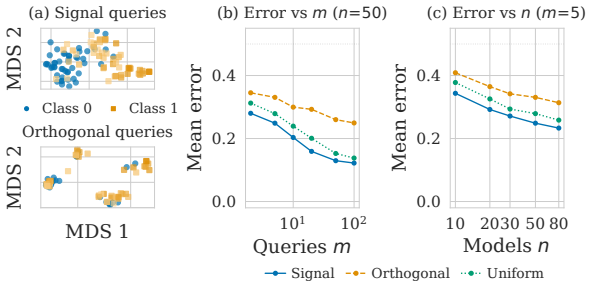

Figure 1. Classification of models by the presence of particular fine-tuning data. An oracle knows which queries probe the distinguishing axis (Signal queries) and which do not (Orthogonal queries). **(a)** Demonstration of the effect of query selection on model representations. MDS embeddings in $\mathbb{R}^2$ 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 in (b) and (c) are the average of 500 different random samples.

of the scree plot of singular values of D as the embedding dimension d. We then train a random forest classifier with default parameters (Pedregosa et al., 2011) on a sample of n training representations and evaluate on the remainder. This process is repeated 500 times for each configuration of (m, n). Additional details can be found in Appendix A.

Figure 1 illustrates the phenomena of interest. Panel (a) clearly shows the discrepancy in embedding quality: Signal queries yield strong class separation, while Orthogonal queries show much weaker separation. Panel (b) quantifies this observation via the classification error on the test models. The relative ordering of the sets matches intuition: Signal query sets outperform Uniform query sets, which Outperform orthogonal query sets, regardless of the number of queries included. Orthogonal sets still enable better-than-chance classification on average, however. Panel (c) shows that this ordering is stable across the number of models for a fixed query budget, with a natural decrease in error as the size of the training sample increases.

In practice, the collection of Signal queries for a classification task is not known *a priori*. We therefore introduce a framework where we can describe the Signal/Orthogonal distinction and which can be represented by a query-model matrix. The spectral properties of this matrix, which can be estimated from the training data, identify Signal queries to improve query efficiency for inference on new models.

3. Framework and Estimation

Figure 1 demonstrates that queries do not contribute equally to the model representations Ψ_Q for a classification task. To

formalize this, we decompose the discriminative content of a query into independent directions.

Definition 3.1. Consider a model space $\mathcal{F}$, query set $\mathcal{Q}$, and appropriate metric d_P. 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 α and ϕ admit a *discriminative factorization* of rank r for $(\mathcal{F}, \mathcal{Q}, d_P)$. 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. Each *zero set* of direction ℓ is $\mathcal{Z}_\ell = \{q \in \mathcal{Q} : \alpha_\ell(q) = 0\}$ and corresponds to queries Orthogonal to direction ℓ. The $\rho_\ell = \Pi_Q(\mathcal{Z}_\ell)$ are the *zero-set probabilities*. Queries $q \in \bigcap_\ell \mathcal{Z}_\ell$ are Orthogonal to the task.

Note that a discriminative factorization always exists with $r \leq M$: For any enumeration $q_1, \ldots, q_M$ of $\mathcal{Q}$, we can take $\alpha_\ell(q) = \mathbb{1}[q = q_\ell]$ and $\phi_\ell(f, f') = d_P^2(P_f(q_\ell), P_{f'}(q_\ell))$. For finite models $f_1, \ldots, f_n$,

3.1. Estimating Orthogonal Queries

In practice, the discriminative factorization must be estimated from the n labeled training models $(f_1, y_1), \ldots, (f_n, y_n)$ and a pilot collection of m queries Q. This is a one-time cost for all future classification. We can do this by constructing 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 $q_j \in Q$. Because of the definition of black-box models in Section 2, each $P_f(q)$ corresponds to a "feature" of the model f. Finding informative queries then amounts to finding informative "features" in this sense.

We can thus formulate estimates $\hat{r}, \hat{\alpha}$ of the discriminative factorization using the singular value decomposition of E. We can estimate r using any singular value based rank-estimation technique (e.g. using the spectral gap: $\hat{r} = \operatorname{argmax}_s \ \sigma_s/\sigma_{s+1}$ for the ordered singular values). Once we determine $\hat{r}$, we let $\hat{\boldsymbol{\alpha}}_Q$ be the first $\hat{r}$ left singular vectors of E, with each $\hat{\alpha}_\ell(q_j)$ corresponding to the jth entry of the ℓth column. We can estimate $\hat{\mathcal{Z}}_\ell = \{q_j \in Q : [|\hat{\alpha}_\ell(q_j)| < \epsilon]\}$ with $\hat{\rho}_\ell = \frac{1}{m}\sum_{j=1}^{m} \mathbb{1}\left[q_j \in \hat{\mathcal{Z}}_\ell\right]$ for a given ϵ. Once we have obtained $\hat{\mathcal{Z}}_\ell$, the estimated Orthogonal set are the queries contained in $\bigcap_\ell \hat{\mathcal{Z}}_\ell$.

Because the discriminative factorization is unknown, a "good" threshold ϵ can be difficult to determine with moderate pools of available queries. In the following experiment,

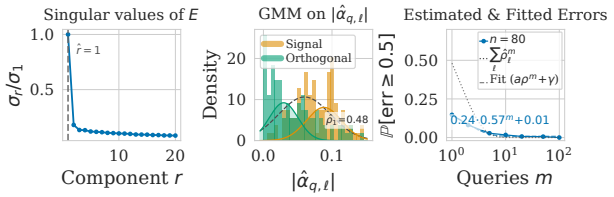

Figure 2. Estimation and validation of the discriminative factorization on the fine-tuning data task. **Left**: Singular value ratios σ_k/σ_{k+1} of the query-model matrix E, with the spectral gap identifying $\hat{r} = 1$ **Center**: GMMs fitted to the $|\hat{\alpha}_\ell(q)|$. Queries are colored according to their estimated GMM component. **Right**: Empirical failure-to-classify over all queries for for $n = 80$ (solid), estimated bound $\hat{\rho}^m$ from the GMM (dotted), and a fitted empirical reference curve of the form $a\rho^m + \gamma$ (dashed).

we opt for a simple Gaussian mixture model (GMM) estimation for $\hat{\mathcal{Z}}_\ell$. For each ℓ, we fit a GMM with $K = \{1, 2\}$ components on $|\hat{\alpha}_\ell(q)|$. We say $K = 2$ whenever $\mathcal{Z}_\ell$ is non-empty and take $\hat{\rho}_\ell$ as the mixture weight of the component close to zero. $\hat{\mathcal{Z}}_\ell$ is the set of of q_j where $\hat{\alpha}_\ell(q_j)$ is in the near-zero component. This estimation is always conservative, since it assigns weak-signal queries to Orthogonal sets and draws efficiency towards that of the full pool. A conservative estimation is not a drawback and generally improves the quality of the Signal set.

Towards a Bound on Failure to Classify The probability of failing to classify models can be naturally formulated as the probability of drawing *only* Orthogonal queries at inference time. A rough bound on the failure probability is thus $\sum_\ell \rho_\ell^m$. [1] As successful classification also depends on the training set, this bound is low for the practical case. In reality, the bound will be of the form $\sum_\ell \rho_\ell^m + \gamma(n)$ for some $\gamma(n) \to 0$ as $n \to \infty$, with $\gamma(n)$ accounting for the probability that the training sample is insufficient for separation. We do not characterize these properties beyond this informal bound in this short paper.

4. Experiments

We evaluate the discriminative factorization framework on the same fine-tuning task as in Section 2.2. Instead of separating *Signal* and *Orthogonal* queries *a priori*, we begin with the *Uniform* set of all available queries. We apply the the spectral gap and Gaussian mixture procedures from Section 3.1 to obtain $\hat{r}$ and construct *estimated Orthogonal* and *estimated Signal* sets from the $\hat{\mathcal{Z}}_\ell$. We then repeat the same procedure as in Section 2.2, sampling m queries uniformly from each of the three groups to construct Euclidean model representations, training a classifier on n models, and

[1] All queries Orthogonal is equivalent to $\alpha_\ell(q) = 0$ for all ℓ and 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 the r directions yields $\sum_\ell \rho_\ell^m$

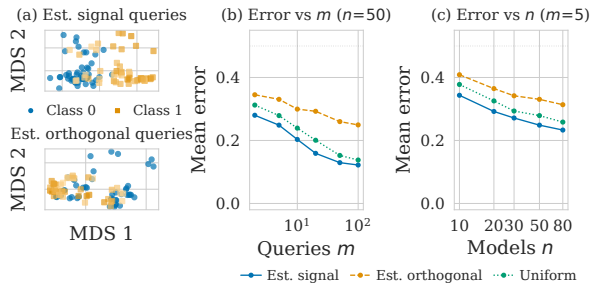

Figure 3. Classification of models by the presence of particular fine-tuning data. Instead of oracle knowledge, we construct *estimated* sets from the pool of all available queries (Uniform) using the discriminative factorization. Results using the estimated sets mirror the corresponding oracle results from Figure 1. **(a)** Demonstration of query selection effect for estimated sets. MDS embeddings in $\mathbb{R}^2$ using $m = 5$ *estimated* Signal queries (top) and *estimated* Orthogonal queries (bottom). Each dot is a model. **(b)** Classification error as a function of query budget m with $n = 50$ for estimated Signal, estimated 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 in (b) and (c) are the average of 500 different random samples.

evaluating on the remainder over 500 repetitions.

Figure 2 demonstrates the estimation process (Left and Center) and the empirical failure probability for better-than-chance classification (Right) on uniformly sampled queries. The process finds a single discriminative direction with $\hat{\rho}_1 = 0.48$. The probability of failing to perform classification is low, even with a relatively small number of queries (e.g. 5). The heuristic bound from Section 3.1 proves accurate for moderate m.

Figure 3 shows the results of the same experiment as in Section 2.2 using the estimated Orthogonal and Signal query sets as proxies. Panel (a) reveals strong separation for the estimated Signal queries in embedding space, while the estimated Orthogonal queries produce much weaker separation. The classification error in panels (b) and (c) preserve the ordering of the sets, which is stable across both model and query count. Like the oracle counterpart, estimated Orthogonal queries also enable better-than-chance classification.

This shows the efficacy of the discriminative factorization framework. Although the full query pool was constructed specifically to include Signal queries for demonstration purposes, it is important to note that the preservation of the ordering means that the notion of more- or less-informative queries is indeed captured. That is, the selection process successfully identifies efficient queries. We thus expect similar results for a more general query pool, though the differences may be less striking depending on the overall composition.

5. Discussion

We introduced the discriminative factorization, a decomposition of the query-model interaction into independent directions, and used it to improve query efficiency in classifying black-box generative models. The framework yields a tentative bound $\sum_\ell \rho_\ell^m$ on the probability of failing to achieve classification, and a practical estimation procedure for recovering the discriminative rank, per-direction zero-set probabilities, and a set of Signal queries from available data. Experiments on a real fine-tuning auditing task confirms that the estimated parameters confer oracle-like Signal vs Orthogonal separation, with the classification error maintaining stable relative ordering across both training sample and query set size.

The finite-space assumptions $|\mathcal{Q}| = M$ and $|\mathcal{X}| = V$ hold in practice – the query space is bounded by the context window length and the response space by the maximum generation length. While V can be large, the effective dimensionality is controlled by the discriminative rank r, not by V.

5.1. Limitations and future work

As this short paper only addresses a single classification setting, we hope to extend this work by examining other base models and auditing tasks. Additionally, the tentative bound on failure to classify from Section 3.1 is informal. We hope to address this in future work and provide a theoretically rigorous version. Such a bound would be able to provide a concrete query budget m^* for a given set of n training models and failure probability under random query sampling. Combined with Signal query set estimation, this would allow practitioners to obtain better-than-chance classification with high probability with higher efficiency.

The experiments use temperature-zero generation, so each model's response to a query is deterministic. The theory natively handles stochastic responses, but empirical validation carries increases cost: nonzero temperature requires multiple responses per query in order to estimate the corresponding response distributions. Understanding the effects of distribution estimation in the black-box setting, as studied for other tasks (Acharyya et al., 2025a;b), would improve the current work's generality. Extending beyond model classification would further probe the limitations of the theoretical framework.

The ability of the estimated Orthogonal set to achieve better-than-chance classification reflects the conservativeness of the GMM zero-set estimation method. However, this behavior was also observed in the oracle case. This implies that there may be very few queries *truly* Orthogonal to the task; a true Orthogonal set may not exist in the sense of the framework. While our setup designates particular Signal and Orthogonal sets, the capture of query utility means

that query selection can be further refined. For example, given the $\hat{\alpha}_\ell(q)$, selecting queries covering the r directions while avoiding redundancy is a D-optimal experimental design problem (Pukelsheim, 2006), for which greedy algorithms with approximation guarantees exist via submodularity (Krause et al., 2008). Formalizing this connection would further reduce query count and translate directly to cost savings at metered API endpoints. Caching responses from previously evaluated models further amortizes cost, since only the model of interest requires new generation (Helm et al., 2026)

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Acharyya, A., Trosset, M. W., Priebe, C. E., and Helm, H. S. Consistent estimation of generative model representations in the data kernel perspective space. *arXiv preprint arXiv:2409.17308*, 2024.

Acharyya, A., Agterberg, J., Park, Y., and Priebe, C. E. Concentration bounds on response-based vector embeddings of black-box generative models. *arXiv preprint arXiv:2511.08307*, 2025a.

Acharyya, A., Priebe, C. E., and Helm, H. S. Testing for LLM response differences: the case of a composite null consisting of semantically irrelevant query perturbations. *arXiv preprint arXiv:2509.10963*, 2025b.

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

Bai, X., Wang, A., Sucholutsky, I., and Griffiths, T. L. Measuring implicit bias in explicitly unbiased large language models. *arXiv preprint arXiv:2402.04105*, 2024.

Caliskan, A., Bryson, J. J., and Narayanan, A. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017.

Canonne, C. L. A survey on distribution testing: your data is big. but is it blue? *Theory of Computing*, 2020. Graduate Surveys, No. 9.

Carlini, N., Tramèr, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, Ú., Oprea, A., and Raffel, C. Extracting training data from large language models. In *30th USENIX Security Symposium*, pp. 2633–2650. USENIX Association, 2021.

Chen, G., Helm, H. S., Lytvynets, K., Yang, W., and Priebe, C. E. Mental state classification using multi-graph features. *Frontiers in Human Neuroscience*, Volume 16 - 2022, 2022. ISSN 1662-5161. doi: 10.3389/fnhum.2022.930291. URL https://www.frontiersin.org/journals/human-neuroscience/articles/10.3389/fnhum.2022.930291.

Chen, N., Helm, H., Park, Y., Priebe, C., and Villar, S. Extracting information from fine-tuned weights. In *Non-Euclidean Foundation Models: Advancing AI Beyond Euclidean Frameworks*, 2025.

Duan, M., Suri, A., Mireshghallah, N., Min, S., Shi, W., Zettlemoyer, L., Tsvetkov, Y., Choi, Y., Evans, D., and Hajishirzi, H. Do membership inference attacks work on large language models? In *First Conference on Language Modeling (COLM)*, 2024.

Duderstadt, B., Helm, H. S., and Priebe, C. E. Comparing foundation models using data kernels. *arXiv preprint arXiv:2305.05126*, 2023.

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

Helm, H., Duderstadt, B., Park, Y., and Priebe, C. E. Tracking the perspectives of interacting language models. *arXiv preprint arXiv:2406.11938*, 2024.

Helm, H., Acharyya, A., Park, Y., Duderstadt, B., and Priebe, C. Statistical inference on black-box generative models in the data kernel perspective space. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 3955–3970, Vienna, Austria, 2025. Association for Computational Linguistics.

Helm, H., Johnson, B., and Priebe, C. Query-efficient model evaluation using cached responses, 2026.

Horwitz, E., Kurer, N., Kahana, J., Amar, L., and Hoshen, Y. We should chart an atlas of all the world's models. *arXiv preprint arXiv:2503.10633*, 2025.

Huh, M., Cheung, B., Wang, T., and Isola, P. Position: The platonic representation hypothesis. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 20617–20642. PMLR, 2024.

Joshi, M., Choi, E., Weld, D. S., and Zettlemoyer, L. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pp. 1601–1611, 2017.

Krause, A., Singh, A., and Guestrin, C. Near-optimal sensor placements in Gaussian processes: Theory, efficient algorithms and empirical studies. *Journal of Machine Learning Research*, 9(8):235–284, 2008.

Maini, P., Jia, H., Papernot, N., and Dziedzic, A. LLM dataset inference: Did you train on my dataset? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Nussbaum, Z., Morris, J. X., Duderstadt, B., and Mulyar, A. Nomic embed: Training a reproducible long context text embedder. *arXiv preprint arXiv:2402.01613*, 2024.

OpenAI. GPT-5 system card. Technical report, 2025. URL https://cdn.openai.com/gpt-5-system-card.pdf.

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

Pukelsheim, F. *Optimal Design of Experiments*. Number 50 in Classics in Applied Mathematics. Society for Industrial and Applied Mathematics, 2006. Originally published: New York, J. Wiley, 1993.

Rizzo, M. L. and Székely, G. J. Energy distance. *Wiley Interdisciplinary Reviews: Computational Statistics*, 8(1): 27–38, 2016.

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

Shi, W., Ajith, A., Xia, M., Huang, Y., Liu, D., Blevins, T., Chen, D., and Zettlemoyer, L. Detecting pretraining data from large language models. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.

Shokri, R., Stronati, M., Song, C., and Shmatikov, V. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 3–18. IEEE, 2017. doi: 10.1109/SP.2017.41.

Székely, G. J. and Rizzo, M. L. Testing for equal distributions in high dimension. *InterStat*, 5(16.10):1249–1272, 2004.

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

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

Zhang, X., Zhao, J., and LeCun, Y. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

A. Experimental Details

A.1. Data

All training data is drawn from the Yahoo Answers Topics dataset (Zhang et al., 2015), which contains approximately 1.4M questions with best answers across 10 topic categories. We designate *Politics & Government* as the sensitive category. Five other 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.

To reduce inter-adapter variance unrelated to the sensitive content, all adapters draw training examples from a shared pool of 2,500 documents per topic group, sampled once at the start. Class 0 adapters draw 500 training examples entirely from the non-sensitive pool. Class 1 adapters 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.

A.2. 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).

A.3. 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 directly probe the sensitive data distinguishing class 1 adapters from class 0.

Orthogonal queries. Questions drawn from TriviaQA (Joshi et al., 2017) (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.

A.4. 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.

A.5. Embedding

Raw text responses are embedded using `nomic-embed-text-v1.5` (Nussbaum et al., 2024) (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.

A.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 $d = 8$ 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.