How Many Probes Identify What a Network Computes?

Anonymous authors
Paper under double-blind review

Abstract

Two networks can attain the same benchmark performance while implementing meaningfully different input–output mechanisms. We formalize when conventional evaluation can distinguish such mechanisms and when designed probes are needed. If two mechanisms disagree on an ϵ fraction of the evaluation distribution, any score obtained by averaging a bounded loss differs by at most $2B\epsilon$, independent of benchmark sample size. We then study identification from queries chosen to separate candidate mechanisms. For a probe set with separation Δ, the query budget scales inversely with Δ, while a sufficient budget grows only logarithmically with the number k of candidates; per-query unreliability is controlled by a Bhattacharyya exponent $D(\rho)$. Simulations recover the predicted inverse-separation scaling over a 25-fold range of Δ. A greedy multicover identifies one of 256 candidate mechanisms using 24 probes, whereas random draws from the same pool still misidentify 31% of models after 256 queries. On sixteen trained transformers, one designed probe reaches the 95% identification target that requires five random probes, and the random-probe curve is predicted without a fitted constant. Finally, across seven Pythia-410m checkpoints, two mechanisms derived from the prompt and corpus statistics recover the checkpoint at which the dominant behavior changes. Across these settings, the benefit of probe design is governed by a separation quantity that can be computed before querying the model.

1 Introduction

Agreement on a benchmark does not imply computational equivalence. Two models can achieve the same aggregate score while relying on different cues or rules, and those differences can become visible precisely on inputs that are rare or absent under the evaluation distribution. This distinction matters whenever the goal is not only to estimate performance, but also to determine which behavior a deployed model will follow under targeted distribution shift.

Increasing benchmark size addresses sampling error, but it does not by itself resolve mechanisms that rarely disagree under the benchmark distribution. Section 3 makes this limitation explicit. If two mechanisms differ on only an ϵ fraction of the evaluation distribution, then every score formed by averaging a bounded loss differs by at most $2B\epsilon$. Collecting more samples estimates that same expectation more precisely; it does not change ϵ.

A different strategy is to choose inputs on which the candidate mechanisms are expected to disagree. We call the disagreement rate of a probe set its *separation*, Δ, and analyze how it controls identification cost. The resulting bounds show inverse dependence on Δ, while the sufficient query budget grows logarithmically with the number of candidate mechanisms. In this formulation, mechanism identification becomes an experimental-design problem: the relevant budget can be estimated from the probes before the model is queried.

The experiments illustrate the practical consequence. Carefully selected probes distinguish 256 candidate mechanisms with 24 inputs. In a controlled training experiment, a single designed probe reaches the same 95% identification target that requires five random in-distribution probes. Across checkpoints of Pythia-410m, the same framework identifies a change in the dominant behavior using candidates derived from the task rather than assigned in advance.

Contributions.

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

- We give a blindness bound (Theorem 2) that applies uniformly to metrics obtained by averaging a bounded loss over the evaluation distribution, including accuracy, Brier score, clipped cross-entropy, and bounded edit-distance variants.
- We derive sufficient and necessary query bounds for mechanism identification (Theorems 4 and 5), exposing the inverse dependence on probe-set separation and the effect of per-query unreliability.
- We formulate probe selection as a design problem, identify the degeneracy of the direct max–min objective, and use a submodular multicover surrogate with a standard $(1 - 1/e)$ greedy guarantee.
- We validate the framework on synthetic candidate families, trained transformers whose mechanism is determined after training, and Pythia-410m checkpoints with candidates derived from prompt and corpus statistics.

2 Related work

What a benchmark can resolve. Schaeffer et al. (2023) showed that apparent emergent abilities (Wei et al., 2022; Srivastava et al., 2023) can arise from discontinuous metrics and may disappear under continuous alternatives. That result concerns how a reported score changes with model scale. Theorem 2 asks a different question: how much any bounded observational score can change when the underlying computation changes at fixed scale. Reparameterizing emergence against pretraining loss (Du et al., 2024) changes the horizontal axis but not this resolution limit.

Evaluation as measurement. Prior work quantifies uncertainty in benchmark scores (Miller, 2024), documents sensitivity to random seeds and data order (Dodge et al., 2020; Bouthillier et al., 2021; Picard, 2021; Henderson et al., 2018), reduces evaluation cost (Polo et al., 2024; Perlitz et al., 2024), and studies benchmark generalization (Recht et al., 2019). These approaches improve the reliability or efficiency of estimating performance. Our focus is complementary: given competing hypotheses about a model's computation, which black-box inputs distinguish them?

Interpretability. Circuit analysis localizes mechanisms by inspecting or intervening on internal components (Elhage et al., 2021; Wang et al., 2023; Conmy et al., 2023; Meng et al., 2022), whereas probing classifiers decode information from representations (Belinkov, 2022; Hewitt & Liang, 2019; Elazar et al., 2021). We instead study a black-box identification problem: how many queries are needed to distinguish candidate mechanisms, and which queries should be selected?

Group testing and experimental design. Identifying one of k hypotheses from binary outcomes is closely related to group testing (Dorfman, 1943; Du & Hwang, 2000), with classical constructions based on superimposed codes and separating systems (Kautz & Singleton, 1964; Katona, 1966). Our design objective is a submodular multicover and therefore inherits standard greedy guarantees (Nemhauser et al., 1978; Wolsey, 1982; Chvátal, 1979; Feige, 1998). The setting is also related to optimal experimental design (Fedorov, 1972; Pukelsheim, 1993; Chaloner & Verdinelli, 1995). Unlike active learning (Settles, 2009; MacKay, 1992; Houlsby et al., 2011), the model is already trained; the objective is to discriminate among hypotheses about its behavior.

Why multiple mechanisms arise. Neural networks can favor inexpensive or spurious features (Geirhos et al., 2020; Shah et al., 2020; Ilyas et al., 2019), distinct subnetworks can compete (Merrill et al., 2023), and the training distribution can steer a transformer between a positional shortcut and an induction-based solution (Kawata et al., 2025). This multiplicity motivates mechanism identification as a separate evaluation problem.

3 Benchmarks are blind to mechanism

Let $\mathcal{X}$ denote the input space and $\mathcal{D}$ the evaluation distribution. We call a map $f : \mathcal{X} \to \mathcal{Y}$ a *mechanism*.

Definition 1. *A score M is* observational *on $\mathcal{D}$ with range B if $M(f) = \mathbb{E}_{x\sim\mathcal{D}}[\varphi(f(x), y(x))]$ for some φ satisfying $\|\varphi\|_\infty \le B$.*

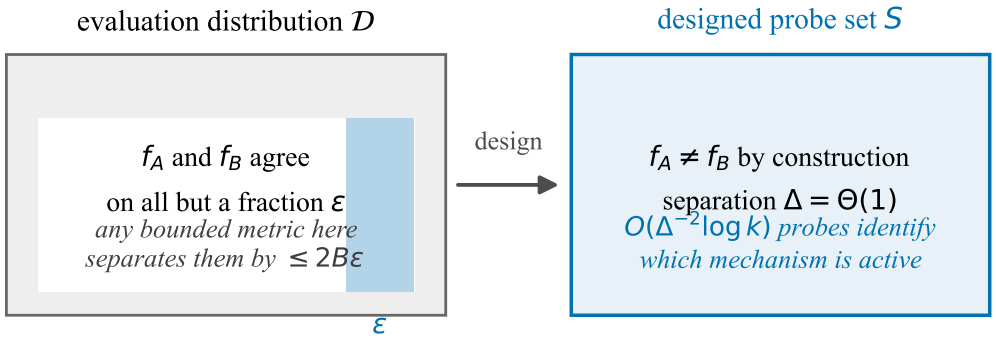

Figure 1: Left: two mechanisms agree on all but an ϵ fraction of the evaluation distribution, so Theorem 2 bounds the difference of any observational score by $2B\epsilon$. Right: a probe set is chosen to make the candidates disagree, yielding separation $\Delta = \Theta(1)$ and a much smaller identification budget.

Accuracy, Brier score, clipped cross-entropy, and bounded token edit distance fit this form. The defining feature is that the score averages over the same evaluation distribution; smoothness of φ is not required.

Theorem 2 (Blindness). *Let $\epsilon = \Pr_{x\sim\mathcal{D}}[f_A(x) \neq f_B(x)]$. Then $|M(f_A) - M(f_B)| \leq 2B\epsilon$ for every observational M on $\mathcal{D}$ with range B, while on $\mathcal{D}' = \mathcal{D}(\cdot \mid f_A \neq f_B)$ the two mechanisms disagree on every input.*

Proof. Couple f_A and f_B under $\mathcal{D}$. Outside the disagreement event the integrands coincide. On the disagreement event, which has probability ϵ, their difference is at most $2B$. □

Corollary 3. *For any $\delta > 0$, there exist mechanisms and an evaluation distribution for which every observational score changes by less than δ, while the two mechanisms disagree on every input from $\mathcal{D}'$.*

The bound is independent of the particular bounded loss used to construct the score. Consequently, changing the metric cannot recover a mechanism difference that has negligible mass under the evaluation distribution. Figure 1 illustrates the contrast with a probe distribution chosen to concentrate on disagreement inputs.

4 The Cost of Identification

Let $\mathcal{F} = \{f_1, \ldots, f_k\}$ be the candidate mechanisms and $S = (x_1, \ldots, x_n)$ a probe set. Define its *separation* by

$$\Delta(S) = \min_{i\neq j} \frac{1}{n} \sum_{t=1}^{n} \mathbf{1}[f_i(x_t) \neq f_j(x_t)]\,,$$

the smallest fraction of probes on which any pair of candidates disagrees. A queried model follows its mechanism with probability $1 - \rho$ on each query, where ρ absorbs stochastic decoding and scoring slack. Define the Bhattacharyya exponent

$$D(\rho) = -\log\big(2\sqrt{\rho(1-\rho)}\big), \qquad D(0) = \infty, \quad D(\tfrac{1}{2}) = 0. \tag{1}$$

Theorem 4 (Sufficiency). *Plurality agreement identifies the active mechanism with probability at least $1 - \delta$ once*

$$n \geq \frac{1}{\Delta} \cdot \frac{\log\big(k(k-1)/2\delta\big)}{\min\{D(\rho),\, \log 2\}}. \tag{2}$$

Proof sketch. Fix a competitor j. Only the at least $n\Delta$ probes separating $f_\star$ from f_j carry information; on each, the observation favors $f_\star$ with probability $1 - \rho$. A Chernoff bound for the resulting ± 1 walk

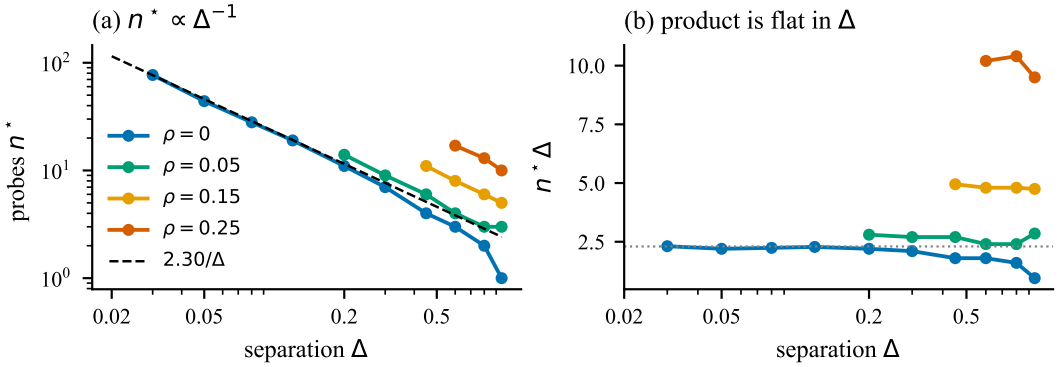

Figure 2: (a) Number of probes required as a function of separation for four levels of per-query unreliability, with the $2.30/\Delta$ reference curve shown for comparison. (b) The product $n^\star\Delta$, which is approximately flat in Δ within each tested reliability regime.

gives pairwise failure probability at most $\exp(-n\Delta D(\rho))$. In the noiseless branch, the probability of observing no separating query is bounded by $(1-\Delta)^n \le e^{-n\Delta}$. A union bound over candidate pairs and rearrangement yield (2). Appendix A gives the full argument. □

Theorem 5 (Necessity). *Every decision rule requires* $n = \Omega\big(\Delta^{-1}\log(1/\delta)/\mathrm{KL}(1-\rho\,\|\,\rho)\big)$ *probes, and* $n = \Omega\big(\Delta^{-1}\log(1/\delta)\big)$ *when* $\rho = 0$.

Proof sketch. Apply a two-point argument (Le Cam, 1973; Tsybakov, 2009) to the closest pair in $\mathcal{F}$. Their observation laws differ only on the $n\Delta$ separating probes, each of which contributes $\mathrm{KL}(1-\rho\|\rho)$. □

The upper and lower bounds both expose the inverse dependence on separation. At fixed reliability, the sufficient budget grows only logarithmically with the number of candidates, whereas the two-point lower bound isolates the unavoidable $1/\Delta$ dependence. Thus the quantity with the greatest leverage is the separation induced by the selected inputs. For a continuous candidate class, the same covering argument replaces the finite-cardinality term by the metric entropy $\log N(\Delta)$ at the relevant resolution.

Measuring the separation law. Figure 2 tests the dependence on Δ directly. Across Δ from 0.02 to 0.95, the fitted log–log slope of $n^\star$ is -1.05 with $R^2 = 0.99$, and $n^\star\Delta$ is approximately constant. In the noiseless regime, the measured values of $n^\star\Delta$ lie between 2.20 and 2.31 for $\Delta \le 0.3$, close to the 2.30 reference value used in the experiment. Increasing per-query unreliability changes the constant while preserving the empirical inverse-separation trend over the tested regime: $n^\star\Delta$ is approximately 2.64, 4.85, and 10.0 for $\rho = 0.05$, 0.15, and 0.25, respectively.

5 Designing the Probe Set

For a fixed budget n, the direct max–min design objective is

$$\max_{|S|=n} \min_{i\neq j} \sum_{x\in S} \mathbf{1}[f_i(x) \neq f_j(x)]\,.$$

This objective is unsuitable for naive greedy optimization. Until every pair of candidates has been separated at least once, the minimum remains zero, so many candidate probes receive identical value and the greedy choice is effectively uninformative. In our experiments, this direct variant performed no better than random selection.

We instead introduce a saturation level. Let $\Phi_T(S) = \sum_p \min(\mathrm{cov}_p(S), T)$, where p indexes candidate pairs, and increase T by one whenever every pair has reached the current level. For fixed T, Φ_T is monotone and submodular, so greedy maximization under a cardinality constraint obtains the

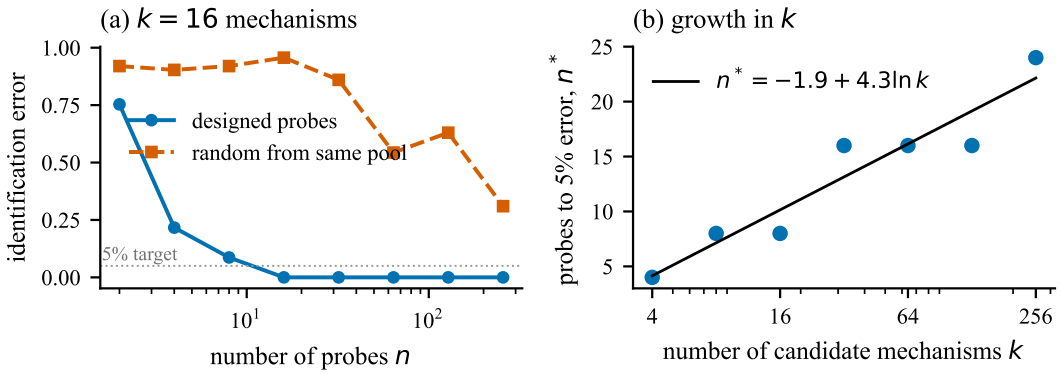

Figure 3: (a) Identification error versus probe count at $k = 16$ for designed and random probe sets drawn from the same candidate pool. (b) Number of probes required to reach 5% error as the number of candidate mechanisms increases.

n	1	2	3	5	8	12	20
designed probes	0.988	0.987	0.998	**1.000**	1.000	1.000	1.000
random in-distribution	0.701	0.819	0.881	0.952	0.987	0.998	1.000
predicted $1 - \frac{1}{2}(1-\eta)^n$	0.700	0.820	0.892	0.961	0.992	0.999	1.000

Table 1: Fraction of the sixteen trained networks whose mechanism is identified correctly. The prediction in the third row contains no fitted parameter.

standard $(1 - 1/e)$ approximation guarantee (Nemhauser et al., 1978; Wolsey, 1982). Increasing the saturation level progressively balances coverage across pairs and approximates the desired max–min objective. When each candidate reads out one of k features, the optimal construction reduces to a separating system of size $\Theta(\log k)$ (Katona, 1966; Kautz & Singleton, 1964).

6 Experiments

The synthetic experiments focus on a regime where informative inputs are rare. Each pool contains M candidate inputs, with only 2–5% of the pool separating a given pair of candidate mechanisms. Unless stated otherwise, per-query unreliability is $\rho = 0.05$ and the target identification error is 5%.

Designed probes versus random collection. Figure 3(a) compares designed probes with random draws from the same pool for $k = 16$. Sixteen designed probes reach the target error, whereas random selection still misidentifies 31% of models after 256 queries. Panel (b) varies the number of candidates: the required designed-probe counts are $4, 8, 8, 16, 16, 16, 24$ for $k = 4, \ldots, 256$. Increasing the hypothesis set by a factor of 64 therefore increases the observed probe budget by a factor of six; a log-linear fit gives $R^2 = 0.904$, consistent with the logarithmic dependence in the sufficient bound.

Networks whose mechanism is learned. The preceding candidate functions are specified directly. We next train sixteen two-layer transformers on a task with two competing solutions: reading a cue token or computing the parity of three content tokens. At the chosen training setting, both solutions occur across random seeds. A held-out set of 4000 disagreement inputs is used to assign the resulting networks to one of the two mechanisms, yielding seven rule networks and nine cue networks with no ambiguous cases. We then ask how accurately the mechanism can be identified from n probes.

Identification is nearly deterministic on the designed probes: one probe is already correct for 98.8% of networks. For the random in-distribution probes, the corresponding calculation gives $n^\star = \lceil \log(2\delta)/\log(1-\eta) \rceil = 5$ at the 95% target and predicts accuracy 0.700 at $n = 1$. These quantities were recorded before the experiment; the measured values are 5 and 0.701, respectively (Table 1, Figure 4), and the full predicted curve is within 0.011 of the measured one. The gap

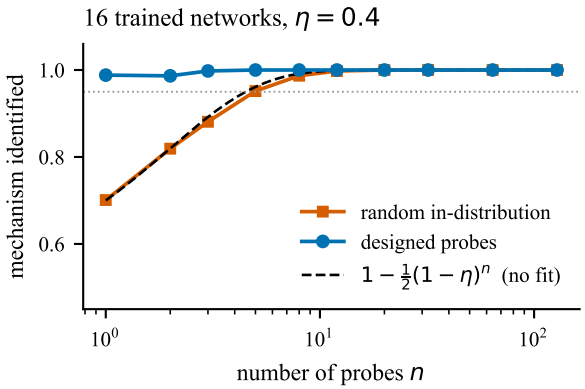

Figure 4: Mechanism identification on trained networks. One designed probe reaches the 95% target that requires five random in-distribution probes, and the random-probe curve is predicted without fitting.

increases as disagreement inputs become rarer: at $\eta = 10^{-2}$, the same calculation requires roughly 300 random probes, whereas a single perfectly separating designed probe remains sufficient in this two-candidate setting.

6.1 Mechanism identification across language-model checkpoints

A practical limitation of the preceding experiments is that the candidate set is specified in advance. We therefore construct a setting in which both candidates are derived from the task: one from the current prompt and one from corpus statistics.

Setup. Prompts are WikiText prefixes of 64 tokens containing one planted pair $(t,\ \text{after})$ and ending with a repeat of t. The two candidates are INDUCTION, $f_{\text{ind}}(x)$, defined as the token that followed the earlier occurrence of t in the prompt, and BIGRAM, $f_{\text{big}}(x) = \arg\max_w \text{count}(t, w)$ over WikiText. Both are defined on every constructed input and neither is fitted to the evaluated model. Designed probes enforce after $\neq f_{\text{big}}$, giving $\Delta = 1$. Natural probes sample after from the corpus bigram distribution, giving measured separation $\Delta = 0.746$.

We evaluate seven checkpoints of Pythia-410m along a single training trajectory (Biderman et al., 2023). Ground-truth labels are assigned using 1,200 designed probes per checkpoint. At step 512, the model agrees more often with BIGRAM than with INDUCTION (0.164 versus 0.008). From step 1000 onward, the ordering reverses and remains reversed, with induction agreement ranging from 0.254 to 0.324 and bigram agreement from 0.138 to 0.102. The change between steps 512 and 1000 is consistent with the formation of induction behavior described in prior work (Olsson et al., 2022) (Figure 5a).

Identification results. Pooling the seven checkpoints, designed probes reach 95% identification accuracy at $n = 32$, whereas natural probes reach the same threshold at $n = 64$ (Figure 5b). At step 512, the only checkpoint assigned to the bigram-dominant class, designed probes are exact by $n = 32$, while natural probes reach 0.973.

Observed gains track separation. The natural probe set has $\Delta = 0.746$, so its inverse separation is 1.34 relative to the designed set with $\Delta = 1$. The measured threshold ratio is $2\times$, a modest gain compared with the much larger ratios in Section 6, where informative inputs occupy only 2–5% of the pool. This contrast is consistent with the central design principle: when an existing evaluation set already has $\Delta = \Theta(1)$ for the candidate mechanisms of interest, deliberate probe design offers limited additional benefit; when separating inputs are rare, the potential gain is much larger (Figure 5c).

Scope of the certificate. The two derived candidates together account for approximately 0.43 of the model's predictions on designed probes. The certificate therefore identifies which of these two

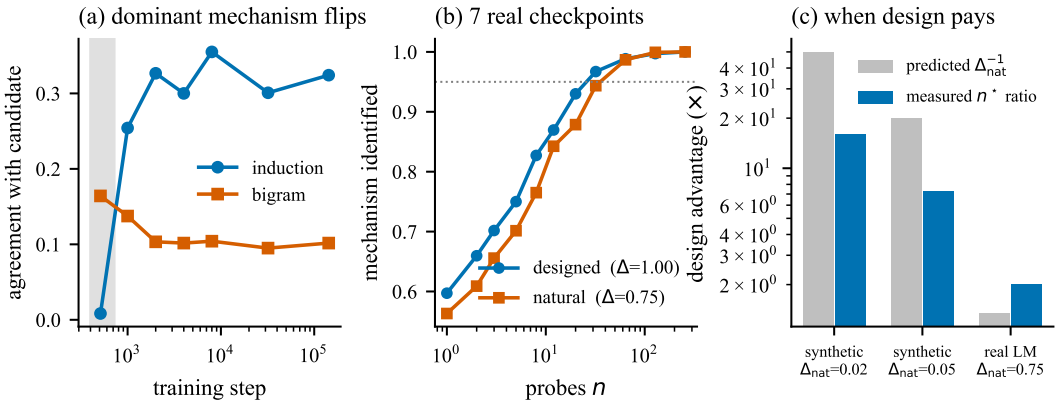

Figure 5: (a) Agreement of Pythia-410m with the two derived candidates across training; the dominant candidate changes between steps 512 and 1000. (b) Identification accuracy pooled over seven checkpoints. (c) Design advantage predicted from $\Delta_{\mathrm{nat}}^{-1}$ compared with the measured $n^{\star}$ ratio across two regimes with substantially different separation.

specified mechanisms is dominant; it does not assert that the pair exhausts the model's behavior. More generally, any identification certificate is conditional on the candidate set against which it is computed.

7 Discussion

Performance estimation and mechanism identification require different kinds of data. Increasing the number of samples drawn from a fixed benchmark distribution reduces uncertainty in the estimated score, but it does not change the mass on which competing mechanisms disagree. In contrast, mechanism identification can benefit directly from selecting inputs that increase pairwise separation.

For evaluations intended to distinguish specific failure modes or computational strategies, separation provides a useful additional diagnostic. It can be computed from the candidate mechanisms before the evaluated model is queried, and the bounds in Section 4 convert it into a query budget. A small value of Δ indicates that the evaluation set has little power to distinguish the specified candidates, even if the aggregate score itself is estimated precisely.

The analysis also clarifies where additional effort is most useful. Expanding a finite candidate set affects the sufficient budget only logarithmically, whereas improving separation changes the leading query cost directly. In the experiments here, that distinction ranges from a modest factor-of-two gain when natural separation is already high to much larger savings when informative inputs are rare.

Reproducibility Statement

The appendix specifies the theoretical assumptions, proof details, training configuration, and simulation protocol. The supplementary material contains the numerical outputs underlying the reported tables and figures so that the plotted quantities and summary statistics can be checked directly.

AI Use Statement

GPT was used only to polish the writing of this manuscript.

References

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, et al. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, 2023.

Xavier Bouthillier, Pierre Delaunay, Mirko Bronzi, et al. Accounting for variance in machine learning benchmarks. In *Proceedings of Machine Learning and Systems*, 2021.

Kathryn Chaloner and Isabella Verdinelli. Bayesian experimental design: A review. *Statistical Science*, 10(3):273–304, 1995.

Vasek Chvátal. A greedy heuristic for the set-covering problem. *Mathematics of Operations Research*, 4(3):233–235, 1979.

Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. In *Advances in Neural Information Processing Systems*, 2023.

Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. *arXiv preprint arXiv:2002.06305*, 2020.

Robert Dorfman. The detection of defective members of large populations. *The Annals of Mathematical Statistics*, 14(4):436–440, 1943.

Ding-Zhu Du and Frank K. Hwang. *Combinatorial Group Testing and Its Applications*. World Scientific, 2000.

Zhengxiao Du, Aohan Zeng, Yuxiao Dong, and Jie Tang. Understanding emergent abilities of language models from the loss perspective. In *Advances in Neural Information Processing Systems*, 2024.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9, 2021.

Nelson Elhage, Neel Nanda, Catherine Olsson, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021.

Valerii V. Fedorov. *Theory of Optimal Experiments*. Academic Press, 1972.

Uriel Feige. A threshold of $\ln n$ for approximating set cover. *Journal of the ACM*, 45(4):634–652, 1998.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep reinforcement learning that matters. In *AAAI Conference on Artificial Intelligence*, 2018.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Conference on Empirical Methods in Natural Language Processing*, 2019.

Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*, 2011.

Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. In *Advances in Neural Information Processing Systems*, 2019.

Gyula Katona. On separating systems of a finite set. *Journal of Combinatorial Theory*, 1(2):174–194, 1966.

William H. Kautz and Richard C. Singleton. Nonrandom binary superimposed codes. *IEEE Transactions on Information Theory*, 10(4):363–377, 1964.

Ryotaro Kawata, Yujin Song, Alberto Bietti, Naoki Nishikawa, Taiji Suzuki, Samuel Vaiter, and Denny Wu. From shortcut to induction head: How data diversity shapes algorithm selection in transformers. In *Advances in Neural Information Processing Systems*, 2025.

Lucien Le Cam. Convergence of estimates under dimensionality restrictions. In *The Annals of Statistics*, volume 1, pp. 38–53. 1973.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

David J. C. MacKay. Information-based objective functions for active data selection. *Neural Computation*, 4(4):590–604, 1992.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. In *Advances in Neural Information Processing Systems*, 2022.

William Merrill, Nikolaos Tsilivis, and Aman Shukla. A tale of two circuits: Grokking as competition of sparse and dense subnetworks. In *ICLR Workshop on Mathematical and Empirical Understanding of Foundation Models*, 2023.

Evan Miller. Adding error bars to evals: A statistical approach to language model evaluations. *arXiv preprint arXiv:2411.00640*, 2024.

George L. Nemhauser, Laurence A. Wolsey, and Marshall L. Fisher. An analysis of approximations for maximizing submodular set functions. *Mathematical Programming*, 14(1):265–294, 1978.

Catherine Olsson, Nelson Elhage, Neel Nanda, et al. In-context learning and induction heads. *Transformer Circuits Thread*, 2022.

Yotam Perlitz, Elron Bandel, Ariel Gera, Ofir Arviv, Liat Ein-Dor, Eyal Shnarch, Noam Slonim, Michal Shmueli-Scheuer, and Leshem Choshen. Efficient benchmarking of language models. In *NAACL*, 2024.

David Picard. torch.manual_seed(3407) is all you need: On the influence of random seeds in deep learning architectures for computer vision. *arXiv preprint arXiv:2109.08203*, 2021.

Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. tinybenchmarks: Evaluating llms with fewer examples. In *International Conference on Machine Learning*, 2024.

Friedrich Pukelsheim. *Optimal Design of Experiments*. Wiley, 1993.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, 2019.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a mirage? In *Advances in Neural Information Processing Systems*, 2023.

Burr Settles. Active learning literature survey. Technical report, University of Wisconsin–Madison, 2009.

Harshay Shah, Kaustav Tamuly, Aditi Raghunathan, Prateek Jain, and Praneeth Netrapalli. The pitfalls of simplicity bias in neural networks. In *Advances in Neural Information Processing Systems*, 2020.

Aarohi Srivastava et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*, 2023.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2009.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: A circuit for indirect object identification in gpt-2 small. In *International Conference on Learning Representations*, 2023.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models. *Transactions on Machine Learning Research*, 2022.

Laurence A. Wolsey. An analysis of the greedy algorithm for the submodular set covering problem. *Combinatorica*, 2(4):385–393, 1982.

A PROOFS

Theorem 4. Let A_j be the event that candidate j attains at least as many agreements as the true mechanism $f_\star$, and let $S_{\star j}$ denote the probes on which they differ, so $|S_{\star j}| \geq n\Delta$. Outside $S_{\star j}$ the two candidates contribute equally, hence A_j depends only on the separating probes. Set $Z_t = +1$ when the observation at $t \in S_{\star j}$ matches $f_\star$ and $Z_t = -1$ otherwise. The variables Z_t are independent with $\Pr[Z_t = 1] = 1 - \rho$. Therefore $\Pr[A_j] = \Pr[\sum_t Z_t \leq 0]$, and optimizing the Chernoff parameter for a ± 1 walk gives $\Pr[A_j] \leq \left(2\sqrt{\rho(1-\rho)}\right)^{|S_{\star j}|} = e^{-|S_{\star j}| D(\rho)}$, with D as in (1). In the noiseless branch, the probability of observing no separating query is $(1-\Delta)^n \leq e^{-n\Delta}$. Combining the two cases yields $\Pr[A_j] \leq \exp(-n\Delta \min\{D(\rho), \log 2\})$, and a union bound over $\binom{k}{2}$ pairs gives the stated result.

Theorem 5. Consider the two closest candidates and the hypotheses under which either one is active. Their observation laws coincide outside the $n\Delta$ separating probes and differ by $\mathrm{KL}(1-\rho\|\rho)$ on each separating probe. The total divergence is therefore at most $n\Delta\,\mathrm{KL}(1-\rho\|\rho)$. Le Cam's method lower-bounds the sum of type-I and type-II errors by a constant unless this quantity is at least of order $\log(1/\delta)$. For $\rho = 0$, the analogous bound follows from the probability of observing a separating probe.

B EXPERIMENTAL DETAILS

The learned-mechanism experiment uses two-layer pre-layernorm transformers of width 32 with four attention heads and a fourfold feedforward expansion. Each model is trained with AdamW (Loshchilov & Hutter, 2019) for 2500 steps at batch size 128 on 20,000 examples, using learning rate 3×10^{-3} under a one-cycle schedule, weight decay 10^{-2}, and gradient clipping at norm one. Sequences contain a marker token at position zero drawn from a two-symbol alphabet disjoint from the content vocabulary, followed by eight content tokens drawn uniformly from sixteen symbols. The label is the parity of whether the first three content tokens lie in the upper half of the alphabet, and the marker equals the label except on a fraction η of examples.

The sixteen models are trained in a batched ensemble implementation with a leading ensemble axis. To preserve independence across members, the objective is the sum of each member's mean batch loss, rather than a mean over the entire ensemble, and gradient clipping is applied separately to each member. Each member also receives an independently sampled dataset.

C SIMULATION DETAILS

The pool used in Section 6 contains $M = 4000$ candidate inputs, with 2–5% of the pool separating a given pair of candidates. The remaining inputs are points on which the candidates agree, modeling the regime in which informative inputs are uncommon under ordinary collection. Greedy probe selection evaluates the saturating multicover objective over all remaining pool items at each step. For Figure 2, identification error at each (Δ, ρ, n) is estimated from 4000 trials, and $n^\star$ is located by exponential search followed by bisection.