INTERACTION COLLAPSES SAFETY: ON THE QUERY-ACCESS LIMITS OF JAILBREAK-ROBUST GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent theoretical work has formalized a trade-off between consistency and breadth in language generation from positive examples. Deployed language models are, in addition, queried: users choose prompts and observe outputs, and jailbreaking uses this access. We develop a query-access theory of jailbreak robustness, in which a generator is trained from positive data and accessed through black-box sampling. We show that a generator that is both broad and sufficiently robust yields, from samples on a fixed batch of prompts, a membership estimator for the target language. Our main result, the Query-Access Impossibility Theorem, follows: for any class of prefix-closed languages (violations cannot be repaired once emitted) that is not statistically identifiable from positive data, no generator is both uniformly robust and uniformly detectably broad once its worst-case error is small relative to its detectable breadth. Adaptivity is unnecessary for this impossibility, but it can substantially reduce the cost of discovering the boundary. All properties are quantified over the random training sample, and an explicit rate sequence and an explicit non-identifiable class show that the hypotheses hold simultaneously. Three consequences follow: robust generators lose detectable breadth, broad generators leak detectably in the worst case, and requiring average-case consistency adds no constraint. Exact or noisy membership feedback removes the barrier, and preference feedback does so exactly when its reward carries information about the boundary. The message is that a sufficiently broad generator reveals, through its own samples, the boundary it is meant to protect: when positive data do not identify that boundary, the tension between helpfulness and jailbreak robustness is information-theoretic and cannot be removed by better training on the same data, but can be removed by feedback that certifies the boundary.

1 INTRODUCTION

Theoretical work on language generation models a learner that observes positive examples from an unknown target language K and must produce new members of K (Gold, 1967; Kleinberg & Mullainathan, 2024). Deployed Large Language Models (LLMs) are used differently: a user chooses a prompt, inspects the output, and may choose again. *Jailbreaking* is the use of this access to elicit outputs that violate safety constraints (Wei et al., 2023; Wolf et al., 2024).

Alignment methods such as Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al., 2022) target safety on typical inputs, whereas jailbreaks concern the worst case. Empirical studies report that aligned models can behave safely on typical inputs while remaining vulnerable to crafted prompts (Jain et al., 2023; Zou et al., 2023), and red-teaming and patching alternate accordingly. We ask: *is jailbreak-robust generation possible for expressive language classes, or does query access to a deployed model impose a barrier?*

We study this question in a statistical model of generation. A passive learner observes only positive samples, but anyone who can query the trained generator can observe where it places probability mass. We show that, when violations are irreversible once emitted, this access yields a *statistical membership estimator*: breadth gives every valid string non-negligible prefix mass under some prompt in a search budget, robustness keeps the probability of observing an invalid one small, and under the stated conditions this prefix mass can be estimated consistently enough to distinguish

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.

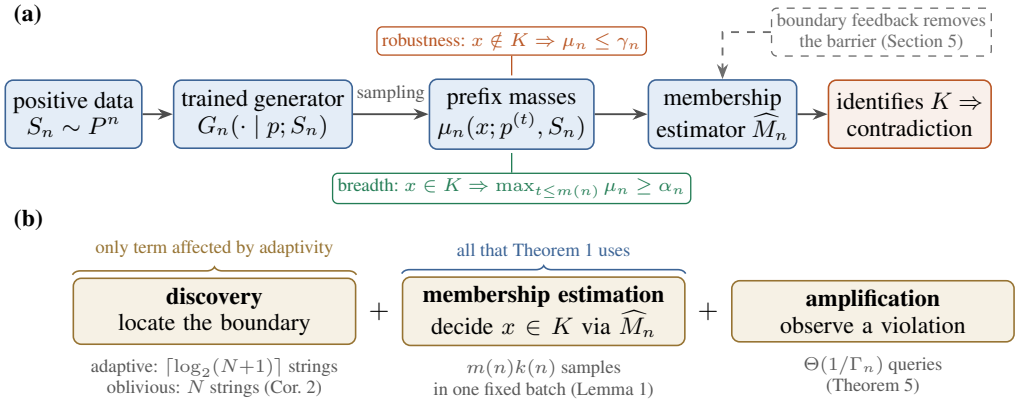


Figure 1: **(a)** The reduction behind Theorem 1. Robustness caps the prefix mass of invalid strings and breadth guarantees that of valid ones, so sampled prefixes decide membership, which is impossible for a non-identifiable class; feedback that carries boundary information removes the barrier. **(b)** The cost of an attack in the model. The impossibility rests only on membership estimation from one fixed, non-adaptive batch, and adaptivity reduces only the cost of discovery.

membership in K. A safe and broad generator therefore inherits the difficulty of *Statistical Identification in the Limit* (SIL).

A query-access theory of jailbreak robustness. The object of study is a generator trained from positive data and accessed only through black-box sampling, the access a deployed model grants its users. Our estimator uses only a fixed batch of prompts, repeated sampling, and inspection of output prefixes, so the theory concerns the information that sampling reveals, not the mechanics of particular attacks. The cost of an attack in the model splits into discovery, membership estimation and amplification, and adaptivity enters only the first (Figure 1b). Breadth is operational (detectable within a bounded probing budget), and prefix-closed languages model irreversible violations such as malformed tool calls, schema errors, or disclosed secrets.

Relation to query learning. In classical membership-query learning (Angluin, 1988b), the learner is given an oracle for the target concept. No such oracle is given here: the generator is trained from positive data only, and membership information arises from its sampling behavior. The theorem is thus a reduction from generation with black-box sampling to identification with approximate membership queries, and the Gold-type barrier (Gold, 1967) that passive generation avoids (Kleinberg & Mullainathan, 2024) reappears.

Why this matters. Three implications follow. First, the familiar tension between refusal and leakage is not an artifact of current training methods: without boundary information it is forced, which accounts for the recurrence of patching after red-teaming. Second, within the model, a fixed batch of prompts suffices for the information-theoretic boundary-recovery argument, even though adaptivity can substantially reduce discovery cost, which bears on how robustness should be evaluated. Third, the positive results state what a safety mechanism must supply, namely membership-type boundary information, a concrete criterion for designing verifiers and guardrails.

Contributions.

- **Main theorem** (Theorem 1). For non-identifiable prefix-closed classes, uniform robustness and uniform detectable breadth are incompatible under a statistical *Gap Condition*. The proof uses an oblivious membership estimator built from one fixed batch of budgeted prompts (Lemma 1).
- **A quantified, non-vacuous model.** Each property is a statement about the random training sample S_n, with breadth required for each string fixed before the sample is drawn (Section 4.1). An explicit polynomial rate sequence (Table 1) and an explicit non-identifiable class (Proposition 1) satisfy all hypotheses simultaneously.

- **Consequences.** Robustness forces loss of detectable breadth, breadth forces detectable worst-case leakage (Propositions 2 and 3), and consistency adds no constraint because robustness implies it (Corollary 1).
- **Boundary information** (Section 5). Membership feedback removes the barrier, and positive data reduce the required queries under a budget; preference feedback preserves the barrier for boundary-uninformative rewards and simulates membership for boundary-exact ones.
- **Adaptivity** (Corollary 2). Adaptivity is unnecessary for the information-theoretic impossibility, but it can reduce the cost of discovering the boundary exponentially; amplifying a known failure costs $\Theta(1/\Gamma_n)$ queries (Theorem 5).

2 Related Work

Language Learning and Identification. Our framework builds on the classical theory of *identification in the limit* (Gold, 1967), which established that many expressive language classes (e.g., any class containing all finite languages and one infinite language, such as the regular languages) cannot be identified from positive examples alone; Angluin's tell-tale criterion characterizes the identifiable classes (Angluin, 1980), and stochastic presentations do not enlarge this family (Angluin, 1988a). While recent work by Kleinberg & Mullainathan (2024) and Kalavasis et al. (2025; 2026) has extended these results to the statistical setting of language generation, they consider passive data access. Our work departs from this by adding sampling access to the trained generator, connecting to Angluin (1988b)'s *Query Learning* framework and to teaching complexity (Goldman & Kearns, 1995; Hegedüs, 1995): while a learner may be limited to positive data (Gold, 1967), query access to a broad and robust generator acts as approximate membership queries (Angluin, 1988b). Compared with the two closest antecedents: where Kleinberg & Mullainathan (2024) and Kalavasis et al. (2025) assume passive access to positive samples and study, respectively, the gap between generation and identification and the consistency–breadth trade-offs of hallucination under a fixed distribution, we add a query-based prober and take the supremum over prompts, so that the failure mode becomes worst-case jailbreaking rather than average-case hallucination.

Theoretical Limits of Alignment. Several recent studies formalize limitations of current alignment paradigms. Wolf et al. (2024) argue, within a behavioral-expectation model, that alignment which only attenuates an undesired behavior can be undone by sufficiently long prompts, and Wei et al. (2023) identify competing objectives and mismatched generalization as drivers of jailbreak success. Our results are complementary: rather than modelling a specific alignment mechanism, we ask what any generator trained from positive data can achieve when its worst-case behavior over prompts is measured, and we show that the tension between 'helpfulness' (breadth) and 'harmlessness' (robustness) arises necessarily in this model.

Certifiable Robustness and Membership Oracles. In response to the brittleness of heuristic alignment, practitioners combine preference-based training with output verification. Preference-based alignment (Christiano et al., 2017; Ouyang et al., 2022), including Constitutional AI, which trains on AI-generated preference labels (Bai et al., 2022), supplies ordinal feedback. Input and output classifiers that accept or reject content (Sharma et al., 2025), rule-based filters, and constrained decoding for structured outputs are closer to 'membership-like' mechanisms. Our Theorems 2 and 3 formalize the value of such absolute feedback, and Section 5.3 isolates when ordinal feedback suffices and when it does not. Bai et al. (2025) study generation in the limit with noise, loss and feedback; our positive results are close in spirit but measure worst-case error over prompts. Further related work on adversarial robustness and on hallucination versus jailbreaking is discussed in Appendix B.1.

3 Preliminaries and Problem Setup

Jailbreaks are inherently *interactive*: users refine prompts based on observed outputs, probing the model for weaknesses (Greshake et al., 2023; Jiang et al., 2024; Zou et al., 2023). We therefore evaluate the *supremum* of the error over prompts rather than its expectation under a prompt distribution, viewing jailbreaking as a worst-case amplification of hallucination (Kalavasis et al., 2025); further background is in Appendix B.2. We follow the statistical framework introduced in Kleinberg & Mullainathan (2024); Kalavasis et al. (2025).

Data and languages. Let Σ be a finite alphabet and $X = \Sigma^*$, with a fixed enumeration $x^{(1)}, x^{(2)}, \dots$ of X and index $\mathrm{idx}(x)$. A *language* is a subset $K \subseteq X$, and $\mathcal{L} = \{L_1, L_2, \dots\}$ is a countable collection of languages. We assume an unknown target language $K \in \mathcal{L}$ and an unknown distribution P such that $\mathrm{supp}(P) = K$. A learning algorithm observes a dataset $S_n = (x_1, \dots, x_n)$ drawn i.i.d. from P; samples are nested, S_n being a prefix of S_{n+1}. A class is *statistically identifiable in the limit* (SIL) if there is a learner, possibly randomized, that maps S_n to an index A_n such that for every $K \in \mathcal{L}$ and every P with $\mathrm{supp}(P) = K$, $\Pr[L_{A_n} = K] \to 1$ (Kalavasis et al., 2025).

Definition 1 (Generator)**.** *A generating algorithm $(G_n)_{n\in\mathbb{N}}$ is a sequence of mappings that, given S_n, outputs a distribution $G_n(\cdot\,; S_n)$ over X.*

We focus on generators that, like LLMs, produce strings token by token over Σ conditioned on a prompt. Modern Transformers typically assign non-zero probability to all strings in Σ^* (full support), so their behavior is characterized by the *distribution of probability mass*, not by which strings are possible. For such a model, $x \in \mathrm{supp}\, G_n(\cdot \mid p; S_n)$ holds for every x and carries no information, whereas the prefix mass $\Pr_{y\sim G_n(\cdot|p;S_n)}[y \text{ begins with } x]$ can still separate valid from invalid strings. We therefore characterize breadth via *prefix-detectability*: whether a valid string appears as an output prefix with probability at least a threshold α_n. Our results use only sampling access, and the decidability of the support of iterative generators (Kalavasis et al., 2025) is not needed (Appendix C.1). Let $\mathcal{P}$ denote a countable set of prompts with a fixed enumeration $p^{(1)}, p^{(2)}, \dots$.

Definition 2 (Prompt-Conditional Generator)**.** *A prompt-conditional generator maps S_n to a family of distributions $\{G_n(\cdot \mid p; S_n)\}_{p\in\mathcal{P}}$ over X. Training randomness is absorbed into S_n as an independent seed, so $G_n(\cdot \mid p; S_n)$ is a deterministic function of S_n.*

The choice of prompt can expand or restrict the effective support of the model, enabling task-specific helpfulness while potentially exposing safety vulnerabilities. We formalize jailbreaking as a closed-loop interaction.

Definition 3 (Interactive Prober)**.** *Given the trained generator (i.e., conditional on S_n), a prober $\mathcal{A}$ selects prompts sequentially and receives independent samples $\phi_t \sim G_n(\cdot \mid p_t; S_n)$. It is* adaptive *if $p_t = \mathcal{A}(\phi_1, \dots, \phi_{t-1})$ may depend on earlier outputs, and* oblivious *if its prompt sequence is fixed in advance.*

The prober does not control S_n, which remains an i.i.d. draw from P; it is therefore not the text-presenting adversary of Gold's model (Appendix B.3). In our proofs the prober is a device of the analysis: the identifier constructed in Theorem 1 is the learner itself, simulating the generator on its own sample.

Probability conventions. Three sources of randomness appear: the training sample S_n (including training randomness), the generator's samples at probing time, and, in Section 5, the answers of a noisy oracle. Unsubscripted $\Pr$ denotes the joint law of all sources present, and $\Pr_{S_n}$ the law of S_n alone. Quantifiers over (K, P) always sit *outside* probabilities, and a rate is *class-uniform* if it does not depend on (K, P). Every property below is a statement about the random sample S_n.

Roadmap. Each hypothesis below enters the reduction at one point, as summarized at the end of the proof of Theorem 1; the conditions are sufficient, and whether they are also necessary is an open question. The closed-loop view is developed in Appendix B.4, and notation is collected in Table 2 (Appendix A).

4 Main Results

We formalize the limits of prompt-conditional language generation under query access. Throughout, $K \in \mathcal{L}$ is unknown, $S_n \sim P^n$ with $\mathrm{supp}(P) = K$, and for each $n \geq 1$ the generator maps S_n to conditional distributions $G_n(\cdot \mid p; S_n)$ over X for every $p \in \mathcal{P}$.

4.1 Formal Properties

For a realized training sample S_n, define the prompt-conditional error, the worst-case error, and the prefix mass

$$e_n(p; S_n) := \Pr_{y \sim G_n(\cdot|p;S_n)}[y \notin K], \qquad \Gamma_n(S_n) := \sup_{p \in \mathcal{P}} e_n(p; S_n),$$
$$\mu_n(x; p, S_n) := \Pr_{y \sim G_n(\cdot|p;S_n)}[y \text{ begins with } x].$$

All three are random variables through their dependence on S_n.

Definition 4 (Jailbreak Robustness)**.** (G_n) *is* jailbreak-robust *on $\mathcal{L}$ if for every $K \in \mathcal{L}$ and every P with* $\mathrm{supp}(P) = K$, $\Gamma_n(S_n) \xrightarrow{p} 0$, *i.e.* $\Pr_{S_n}[\Gamma_n(S_n) > \varepsilon] \to 0$ *for every $\varepsilon > 0$. It has a* uniform robustness envelope (γ_n, ρ_n), *with deterministic $\gamma_n, \rho_n \to 0$, if*

$$\sup_{K \in \mathcal{L}} \sup_{P:\, \mathrm{supp}(P)=K} \Pr_{S_n \sim P^n}\left[\Gamma_n(S_n) > \gamma_n\right] \le \rho_n \quad \textit{for all } n.$$

The envelope implies jailbreak robustness and is used only in the impossibility results. Appendix B.6 compares the two notions.

Definition 5 (Uniform Detectable Breadth)**.** *Fix the enumeration $\{p^{(t)}\}_{t \ge 1}$ of $\mathcal{P}$. (G_n) achieves* uniform detectable breadth *with prompt budget $m(n) \to \infty$, string horizon $s(n) \to \infty$, floor $\alpha_n \ge n^{-c}$ ($c > 0$ fixed for the class $\mathcal{L}$) and failure rate $\beta_n \to 0$, all deterministic and class-uniform, if for every n, every $K \in \mathcal{L}$, every P with $\mathrm{supp}(P) = K$, and every fixed $x \in K$ with $\mathrm{idx}(x) \le s(n)$,*

$$\Pr_{S_n \sim P^n}\Big[\, \exists\, t \le m(n) :\ \mu_n\big(x;\, p^{(t)},\, S_n\big) \ge \alpha_n \Big] \ \ge\ 1 - \beta_n.$$

Remark 1 (Quantifier order, horizon and budget)**.** *This formulation strengthens the classical notion of breadth (Kalavasis et al., 2025) in four ways: coverage must be witnessed (i) by a prompt* within the search budget *($t \le m(n)$), (ii) with a* class-uniform *floor that does not depend on the unknown K or P, (iii) through an observable* prefix *event rather than an exact-string mass, and (iv) for each string x that is fixed* before *S_n is drawn. The last point matters: the formulation "for every $x \in K \setminus S_n$, $\Pr_{S_n}[\cdot] \to 1$" is ill-posed, because the range of the outer quantifier would depend on the random object inside the probability. The horizon $s(n)$ is necessary for class-uniform rates: for each prompt, $\sum_{|x|=\ell} \mu_n(x; p, S_n) \le \Pr[|y| \ge \ell] \to 0$ as $\ell \to \infty$, so $m(n)$ prompts give mass $\ge \alpha_n$ to only finitely many strings of each large length, and a rate uniform over all x in an infinite K is unattainable. Every fixed x lies within the horizon for all large n.*

Our analysis relies on a separability condition ensuring that invalid strings can be distinguished via observable prefix events. Formally,

Assumption 1 (Prefix-separability)**.** *For every $K \in \mathcal{L}$ and every $x \notin K$, there exists a prefix length $\ell(x) \le |x|$ such that $y_{1:\ell(x)} \ne x_{1:\ell(x)}$ for all $y \in K$.*

Remark 2 (Prefix-separability is prefix-closedness)**.** *Assumption 1 holds iff every $K \in \mathcal{L}$ is* prefix-closed *(each prefix of a member is a member). If $y \in K$ had a prefix $x \notin K$, then $y_{1:\ell(x)} = x_{1:\ell(x)}$, contradicting the assumption; conversely, if K is prefix-closed and $x \notin K$, take $\ell(x) = |x|$. Invalidity is thus unrecoverable: no extension of an invalid string is valid, a 'hard syntax error' semantics natural for tool calls, structured outputs, and irreversible safety violations such as a disclosed secret. It also excludes validity notions that depend on the whole string, such as factual correctness, where our reduction does not apply (Appendix B.5).*

Theorem 1 below is therefore a sharp result for irreversible, prefix-detectable constraints, the class that governs tool use, structured outputs and information disclosure; extending it to safety predicates that are not prefix-closed is a natural next step.

4.2 Query-Access Impossibility

We begin by formalizing the key estimator that allows us to reduce generation to identification. It is a 'search-and-verify' strategy in which the prober explores a fixed batch of prompts for a statistical witness of the string x.

Lemma 1 (Executable Membership Estimation)**.** *Let (G_n) have a uniform robustness envelope (γ_n, ρ_n) and uniform detectable breadth (m, s, α, β) over a class satisfying Assumption 1, and let $k(n) \geq 1$. Given $x \in X$, draw $k(n)$ i.i.d. samples $y^{(t,j)} \sim G_n(\cdot \mid p^{(t)}; S_n)$ from each of the first $m(n)$ prompts and set*

$$\widehat{M}_n(x) := \mathbb{1}\{\textit{some sample begins with } x\} = \mathbb{1}\big\{\exists\, t \leq m(n),\ j \leq k(n) :\ y^{(t,j)}_{1:|x|} = x\big\}, \qquad (1)$$

where a sample of length $< |x|$ never matches. Then for every n, every $K \in \mathcal{L}$, and every P with $\mathrm{supp}(P) = K$:

(a) if $x \notin K$, then $\Pr[\widehat{M}_n(x) = 1] \leq \rho_n + m(n)k(n)\gamma_n$;

(b) if $x \in K$ and $\mathrm{idx}(x) \leq s(n)$, then $\Pr[\widehat{M}_n(x) = 0] \leq \beta_n + e^{-k(n)\alpha_n}$.

Hence if (i) $k(n)\alpha_n \to \infty$ and (ii) $m(n)k(n)\gamma_n \to 0$, then $\widehat{M}_n(x) \xrightarrow{p} \mathbb{1}\{x \in K\}$ for every fixed $x \in X$. The estimator uses $m(n)k(n)$ samples and a prompt batch fixed in advance.

Remark 3 (The estimator needs no prefix oracle and no adaptivity)**.** *The algorithm uses only x itself: it sets $\ell(x) := |x|$ and tests whether a sample* begins with *x. It therefore requires no computable prefix-selection rule, no access to $\ell(\cdot)$, no knowledge of membership, and no knowledge or computation of which prompt witnesses x, since it scans all $m(n)$ budgeted prompts and none outside the budget. The procedure is identical whether or not $x \in K$; Assumption 1 is used only in the* analysis*. For $x \notin K$, every sample beginning with x is invalid (Remark 2), so matches are controlled by the worst-case error; for $x \in K$, Definition 5 supplies a budgeted prompt with prefix mass α_n. Finally, the prompt batch does not depend on any output, so the estimator is* oblivious *and the impossibility below holds against non-adaptive probing. Adaptivity is unnecessary for the information-theoretic impossibility, but it can substantially reduce the cost of discovering the boundary (Corollary 2).*

Theorem 1 (Query-Access Impossibility)**.** *Let $\mathcal{L}$ be a language class that is not Statistically Identifiable in the Limit (SIL) and that satisfies Assumption 1. There is no prompt-conditional generator (G_n) that simultaneously has a uniform robustness envelope (γ_n, ρ_n) and uniform detectable breadth (m, s, α, β) such that the* ***Gap Condition***

$$\gamma_n = o\left(\frac{\alpha_n}{m(n)}\right) \qquad (2)$$

holds. The same conclusion holds if the two properties are only required for all sufficiently large n.

The proof (Appendix D) chooses a sampling depth $k(n)$ in the window $1/\alpha_n \ll k(n) \ll 1/(m(n)\gamma_n)$, which is non-empty exactly under (2), applies Lemma 1 to the first $t(n) \leq s(n)$ strings, and runs an elimination identifier. All rates are class-uniform, so one procedure works with (K, P) unknown.

Scope of the theorem. Theorem 1 is deliberately abstract. It isolates what black-box sampling reveals about a safety boundary when valid behavior is broad, violations are prefix-detectable, and the training information does not identify the boundary, and it holds regardless of the semantics of the content, the procedure used to search for prompts, or the pipeline used to train the model. This generality comes from the proof: the identifier is the learner itself, simulating (G_n) on its own sample. Class-uniformity lets the same identifier succeed without knowing K or P, turning the per-string estimator of Lemma 1 into a statement about the whole class.

The hypotheses are jointly satisfiable. Table 1 lists a single rate sequence and checks every condition used in Lemma 1, Theorem 1 and Section 6. The reduction uses a single batch of $m(n)k(n) = n^3$ probe samples, shared by all tested strings, fewer than the $n^4/2$ queries needed to observe even one violation with probability $1/2$ (Theorem 5): the argument is an identification argument, not an amplification argument. Proposition 1 shows that the class-level hypotheses are also satisfiable, and Remark 10 (Appendix B.7) that each property alone is attainable.

Proposition 1 (A prefix-closed non-SIL class)**.** *Let $\Sigma = \{a\}$, $L_k := \{a^i : 0 \leq i \leq k\}$ for $k \geq 0$, and $L_\infty := a^*$. The class $\mathcal{L}_{\mathrm{chain}} := \{L_\infty, L_0, L_1, \dots\}$ satisfies Assumption 1 and is not SIL, even for randomized learners.*

Table 1: A single rate sequence satisfying all conditions simultaneously ($c = 1$).

Quantity	Choice	Condition checked
floor, breadth failure, envelope failure	$\alpha_n = n^{-1},\ \beta_n = \rho_n = n^{-1}$	$\alpha_n \geq n^{-c}$; $\beta_n, \rho_n \to 0$
budget, horizon, sampling depth	$m(n) = s(n) = n,\ k(n) = n^2$	$m, s \to \infty$; $k(n)\alpha_n = n \to \infty$
worst-case error envelope	$\gamma_n = n^{-4}$	$m(n)k(n)\gamma_n = n^{-1} \to 0$
Gap Condition (2)	$\gamma_n\, m(n)/\alpha_n = n^{-2}$	$\to 0$
per-string error r_n	$\max\{\rho_n, mk\gamma_n, \beta_n, e^{-k\alpha_n}\} = n^{-1}$	$\to 0$
test-set size	$t(n) = \min\{s(n), r_n^{-1/2}\} = n^{1/2}$	$t(n) \to \infty$; $4t(n)r_n = 4n^{-1/2} \to 0$
probe cost vs. amplification	$m(n)k(n) = n^3$ (shared batch); $Q_n \geq 1/(2\gamma_n) = n^4/2$	no violation needs to be observed

Remark 4 (Thin-breadth regime)**.** *For a* fixed (K, P)*, one may instead have* $\alpha_n \sim 2^{-|x|}$ *decaying exponentially with string length, or rates that depend on* P*. The estimator of Lemma 1 still converges pointwise, but* $k(n)$ *and* $t(n)$ *can then no longer be chosen without knowing* (K, P)*, so the class-level argument of Theorem 1 does not apply. Extending the theorem to non-uniform rates is a natural next step. An interpretation of the Gap Condition as a comparison between visible breadth and worst-case leakage is given in Appendix B.8.*

4.3 Consequences: Refusal, Vulnerability, and Consistency

We now draw the consequences of Theorem 1 for average-case alignment, worst-case safety, and coverage, distinguishing safety under the training distribution from safety under probing. The only independent constraint is the pairwise one of Theorem 1.

Definition 6 (Consistency)**.** *For a fixed prompt distribution* $\mathcal{P}_{\mathrm{train}}$*,* (G_n) *is* consistent *if for every* $K \in \mathcal{L}$ *and* P *with* $\mathrm{supp}(P) = K$*,* $\mathbb{E}_{p\sim\mathcal{P}_{\mathrm{train}}}[e_n(p; S_n)] \xrightarrow{p} 0$*.*

Remark 5 (Robustness implies consistency)**.** *Because* $\Gamma_n(S_n) = \sup_p e_n(p; S_n)$ *dominates* $\mathbb{E}_{p\sim\mathcal{P}_{\mathrm{train}}}[e_n(p; S_n)]$ *for every realization of* S_n*, jailbreak robustness implies consistency. The three properties are therefore not independent constraints; the binding constraint is the pairwise impossibility of Theorem 1, and adding consistency cannot restore feasibility.*

Corollary 1 (Consistency Corollary)**.** *Under the hypotheses of Theorem 1, no generator is simultaneously consistent, has a uniform robustness envelope* (γ_n, ρ_n)*, and has uniform detectable breadth* (m, s, α, β) *with* $\gamma_n = o(\alpha_n/m(n))$*.*

Corollary 1 adds a hypothesis to Theorem 1 and is therefore strictly weaker; it is not an independent three-way theorem. The informative content lies in the two directions of the theorem's contrapositive, which we state precisely.

Proposition 2 (Refusal)**.** *Let* $\mathcal{L}$ *be non-SIL and satisfy Assumption 1, and let* (G_n) *have a uniform robustness envelope* (γ_n, ρ_n)*. Then for every budget* $m(n)$*, horizon* $s(n)$*, failure rate* $\beta_n \to 0$ *and floor* $\alpha_n \geq n^{-c}$ *with* $m(n)\gamma_n = o(\alpha_n)$*, there are infinitely many* n *for each of which some* $K \in \mathcal{L}$*, some* P *with* $\mathrm{supp}(P) = K$*, and some* $x \in K$ *with* $\mathrm{idx}(x) \leq s(n)$ *satisfy*

$$\Pr_{S_n}\left[\, \mu_n(x; p^{(t)}, S_n) < \alpha_n \ \textit{for all } t \leq m(n) \,\right] > \beta_n.$$

Proposition 3 (Vulnerability)**.** *Let* $\mathcal{L}$ *be non-SIL and satisfy Assumption 1, and let* (G_n) *have uniform detectable breadth* (m, s, α, β)*. Then for every deterministic* $\gamma_n = o(\alpha_n/m(n))$ *and every* $\rho_n \to 0$*, there are infinitely many* n *for each of which some* $K \in \mathcal{L}$ *and* P *with* $\mathrm{supp}(P) = K$ *satisfy* $\Pr_{S_n}[\Gamma_n(S_n) > \gamma_n] > \rho_n$*. On that event some prompt has error above* γ_n*, and by Theorem 5 repeated querying of it yields a violation within* $O(1/\gamma_n)$ *queries.*

Propositions 2 and 3 give a formal account of two phenomena reported for aligned LLMs: robust generators must be refusal-like on some valid strings within the horizon, while broad generators must leak at a rate that probing can exploit. Consistency plays no role in either statement (Appendix B.9).

5 Positive Results

The impossibility results of the previous section arise from the absence of negative information: a learner observing only positive samples cannot identify the precise boundary of an expressive language class, and a broad generator trained by such a learner exposes that uncertainty to probing. We now show that when negative feedback is available via a membership oracle, the query-access impossibility barrier can be overcome, and we ask when preference feedback suffices.

5.1 Robust Generation with Membership Feedback

We extend the model by granting access to a membership oracle $\mathcal{O}(x) = \mathbb{1}\{x \in K\}$. This idealizes 'membership-like' feedback such as rule-based filters, formal verifiers, schema validators, or output classifiers (Sharma et al., 2025). Let $\pi(\cdot \mid p)$ be a base generative distribution (e.g., a pre-trained model) with full support on X.

Theorem 2 (Robust Generation with Feedback)**.** *(a)* Deployment-time oracle. *The filtered generator $G_n(x \mid p) \propto \pi(x \mid p)\,\mathcal{O}(x)$ has $\Gamma_n \equiv 0$ and $G_n(x \mid p) > 0$ for all $x \in K$ and $p \in \mathcal{P}$. (b)* Training-time oracle. *Suppose the learner has access to $S_n \sim P^n$ (with $\mathrm{supp}(P) = K$) and may query $\mathcal{O}$ on a growing sequence of finite sets $T_n \uparrow X$. Then there exists a prompt-conditional generator (G_n) with $\Pr_{S_n}[\Gamma_n(S_n) = 0] \to 1$, hence jailbreak-robust, that achieves asymptotic coverage: for every $x \in K$, $\Pr_{S_n}[\forall p \in \mathcal{P} :\ G_n(x \mid p; S_n) > 0] \to 1$.*

Remark 6 (The role of S_n under an oracle)**.** *Neither construction needs S_n: with an* unbounded-query *exact oracle, the oracle alone identifies K by querying a fixed enumeration, and at deployment time it can simply filter outputs. Theorems 2 and 3 are therefore best read as an information-theoretic* feedback-modality separation*, not as algorithms: holding the model fixed, adding absolute boundary information turns an impossible combination into a trivial one, mirroring the classical Gold-versus-Angluin separation between positive presentations and membership queries (Appendix B.10). Positive data become informative once queries are limited, as the following proposition makes precise, or noisy (Theorem 3).*

Proposition 4 (Query-Budgeted Identification)**.** *Let $\mathcal{H}_n := \{L \in \mathcal{L} : S_n \subseteq L\}$ and let $\mathrm{TD}(K; \mathcal{H})$ be the size of a smallest set $T \subseteq X$ such that K is the only $L \in \mathcal{H}$ with $L \cap T = K \cap T$. Any learner that, on sample S_n, determines K with certainty from S_n and membership queries must query at least $\mathrm{TD}(K; \mathcal{H}_n)$ strings. This quantity is at most $\mathrm{TD}(K; \mathcal{L})$ and is non-increasing in n, so more positive data can only reduce the queries required. If $\mathcal{H}_n$ is finite, $O\big(\mathrm{XTD}(\mathcal{H}_n) \log |\mathcal{H}_n|\big)$ adaptive queries suffice, where* XTD *is the extended teaching dimension (Hegedüs, 1995).*

5.2 Robust Generation under Noisy or Partial Feedback

The positive result in Theorem 2 relies on exact access to a membership oracle. In practice, feedback may be noisy. We assume access to a noisy oracle $\widetilde{\mathcal{O}}$ whose repeated queries on the same x return independent responses $\widetilde{\mathcal{O}}^{(1)}(x), \widetilde{\mathcal{O}}^{(2)}(x), \ldots$ satisfying

$$\Pr\left[\widetilde{\mathcal{O}}^{(j)}(x) = \mathbb{1}\{x \in K\}\right] \geq 1 - \eta \quad \text{for some } \eta < 1/2. \tag{3}$$

To allow for vanishing errors in this probabilistic setting, we relax the requirements, quantified over both the training sample and the oracle's randomness.

Definition 7 (Approximate Robustness and Coverage)**.** *Given $\varepsilon_n, \delta_n \to 0$, (G_n) is* (ε_n)-robust *if $\mathbb{E}_{S_n, \widetilde{\mathcal{O}}}[\Gamma_n] \leq \varepsilon_n$, and achieves* (δ_n)-coverage *if for every $x \in K$, $\Pr_{S_n, \widetilde{\mathcal{O}}}\left[\forall p \in \mathcal{P} :\ G_n(x \mid p) > 0\right] \geq 1 - \delta_n$; both for every $K \in \mathcal{L}$ and P with $\mathrm{supp}(P) = K$.*

Theorem 3 (Robust Generation with Noisy Feedback)**.** *Suppose the learner has access to $S_n \sim P^n$ (with $\mathrm{supp}(P) = K$) and to a noisy oracle $\widetilde{\mathcal{O}}$ with error rate $\eta < 1/2$, which it may query $O(n^c \log n)$ times at sample size n for some fixed $c > 0$. Then there exists a prompt-conditional generator (G_n) that is (ε_n)-robust and achieves (δ_n)-coverage, with $\varepsilon_n, \delta_n \to 0$.*

Theorem 3 shows that exact negative feedback is not necessary: repeated noisy checks, aggregated by majority vote, recover the same boundary information at logarithmic overhead per string. It also specifies what a noisy evaluator must provide for this to work: answers that are better than chance on the relevant strings and independent across repetitions, which gives a concrete criterion for assessing practical verifiers.

5.3 When Ordinal Feedback Contains Boundary Information

Alignment pipelines such as RLHF (Christiano et al., 2017; Ouyang et al., 2022) and Constitutional AI (Bai et al., 2022) use *preference* feedback in the form of pairwise comparisons. To ask what information ordinal feedback can carry, we study an abstract preference oracle: the feedback is modelled through a latent reward $r : X \to \mathbb{R}$ and the induced oracle $\mathcal{C}_r(x, y) = x$ if $r(x) \geq r(y)$ and y otherwise. Membership queries provide *absolute* information, $\mathcal{O}(x) = 0 \Rightarrow x \notin K$, while

a comparison provides *ordinal* information, $\mathcal{C}_r(x, y) = x \Rightarrow r(x) \geq r(y)$, and returns a winner even when both strings are unsafe. Whether ordinal information resolves the boundary depends on the reward. Call r *boundary-uninformative* for $\mathcal{L}$ if it is the same function whatever $K \in \mathcal{L}$ is, and *safety-consistent* if $r(x) = 0$ for all $x \notin K$; for example, $r_0(x) := \mathbb{1}\{x \in \bigcap_{L \in \mathcal{L}} L\}$ has both properties (on $\mathcal{L}_{\text{chain}}$, r_0 rewards only the empty string).

Theorem 4 (Persistence of the Impossibility under Preference Feedback)**.** *Let $\mathcal{L}$ be non-SIL and satisfy Assumption 1, and let r be boundary-uninformative for $\mathcal{L}$ (e.g. r_0, which is also safety-consistent). Then no generator trained on $(S_n, \mathcal{C}_r)$ has a uniform robustness envelope (γ_n, ρ_n) and uniform detectable breadth (m, s, α, β) with $\gamma_n = o(\alpha_n/m(n))$.*

Proposition 5 (Boundary-Exact Rewards Simulate Membership)**.** *If $r = \mathbb{1}_K$, then for any $y \in S_n$, $\mathcal{C}_r(x, y) = x$ iff $x \in K$. Hence one comparison per string simulates $\mathcal{O}$ (with uniformly random tie-breaking, $O(\log n)$ repetitions suffice), and Theorem 2(b) applies.*

Remark 7 (Scope: the calibrated-reference boundary)**.** *Theorem 4 and Proposition 5 bracket two idealizations. The distinction is not ordinal versus absolute feedback as such: comparisons against a calibrated reference of known reward (here, any positive sample) convert ordinal answers into absolute ones whenever the reward's level sets track the boundary. Reward models learned from human or AI comparisons lie between these two extremes, and where a given pipeline falls is an empirical question that our results make precise. Appendix B.11 relates this to red-teaming.*

6 Characterizing the Complexity of the Adversary

This section quantifies the three costs of Figure 1b. The amplification bound gives the baseline cost of exploiting a known failure; the more informative result is the distinction between *discovering* the boundary and *amplifying* a known failure.

Definition 8 (Jailbreak Query Complexity)**.** *For a realized S_n, the* jailbreak query complexity *$Q_n(S_n)$ is the smallest T such that some (possibly adaptive) prober achieves $\Pr[\exists t \leq T : \phi_t \notin K \mid S_n] \geq \frac{1}{2}$.*

Theorem 5 (Amplification Upper and Lower Bounds)**.** *For every realization with $\Gamma_n := \Gamma_n(S_n) > 0$, $1/(2\Gamma_n) \leq Q_n(S_n) \leq \lceil 2 \ln 2/\Gamma_n \rceil$, so $Q_n = \Theta(1/\Gamma_n)$. Under a uniform envelope, $Q_n(S_n) \geq 1/(2\gamma_n)$ with probability at least $1 - \rho_n$.*

The upper bound presumes that a prompt with error at least $\Gamma_n/2$ is known, and the lower bound holds for adaptive probers. Locating such a prompt is the *discovery* phase, which within our model amounts to learning the boundary through the estimator $\widehat{M}_n$ used as an approximate membership oracle.

Corollary 2 (Adversarial Discovery Complexity)**.** *Let r_n be as in Table 1 and fix (K, P). (a) Any deterministic membership-query algorithm making q queries, all within the horizon, can be run with $\widehat{M}_n$ in place of $\mathcal{O}$ using $q\, m(n)k(n)$ probe samples; with probability at least $1 - 2q\, r_n$ its output equals its output under $\mathcal{O}$. (b) Any algorithm that determines K exactly from S_n and membership answers queries at least $\mathrm{TD}(K; \mathcal{H}_n)$ strings (Goldman & Kearns, 1995). (c) On the finite chain $\{L_0, \ldots, L_N\}$, every oblivious query set that identifies all members has size at least N, while adaptive binary search needs $\lceil \log_2(N + 1) \rceil$ queries.*

Corollary 2 answers where adaptivity is essential: it is not needed for the impossibility (Remark 3), but it can reduce the number of probed strings, and hence probe samples, exponentially compared with any fixed test set, from $N\, m(n)k(n)$ to $O(\log N)\, m(n)k(n)$ in (c). The total cost of an attack in the model is the discovery cost plus the amplification cost $\Theta(1/\Gamma_n)$; consequences for 'safety through obscurity' are discussed in Appendix B.12.

7 Discussion

We have shown that a sufficiently broad generator reveals, through its samples, information about the boundary it is meant to protect. When positive data do not identify that boundary, we establish that black-box sampling turns generation into identification, so robustness and detectable breadth cannot both hold at the Gap-Condition rates (Theorem 1) and low expected error remains compatible with failures that probing can reach (Proposition 3). We further show that positive data and boundary information are distinct resources: supplying the latter, through membership or boundary-exact feedback, removes the barrier (Theorems 2–4; future directions in Appendix B.13).

Reproducibility statement

All assumptions are stated in Sections 3–4 (Definitions 4–5 and Assumption 1), and complete proofs of every lemma, theorem, proposition and corollary are given in Appendix D. The parameter sequence used to check that the hypotheses are jointly satisfiable is listed in full in Table 1. The only numerical result, the estimator simulation of Figure 2 and Table 3, is specified completely in Appendix C.3 (class, generator, rates, sampling depths and number of repetitions). The anonymized supplementary material contains the script that produces this figure and table, with its fixed random seed.

Ethics statement

This is a theoretical paper on the limits of jailbreak-robust generation. It contains no human-subject data, no new attack method against deployed systems, and no harmful model outputs; the only experiment uses a synthetic generator over a one-letter alphabet. The estimator of Lemma 1 is an analysis device that requires simulating the generator and enumerating prompts, and it does not provide a practical attack. We see the main risk as misreading the impossibility result as a claim that safety measures for deployed models are futile; the paper states explicitly that the result is conditional on its hypotheses (Section 7), and its positive results identify the kind of feedback under which robust and broad generation is possible.

AI use statement

We used an AI assistant for language editing, conducting literature searches, and checking theorem statements, assumptions, and experimental design. The authors take full responsibility for the content of the paper.

References

Dana Angluin. Inductive Inference of Formal Languages from Positive Data. *Information and Control*, 45(2):117–135, 1980.

Dana Angluin. Identifying Languages from Stochastic Examples. Technical Report YALEU/DCS/RR-614, Yale University, Department of Computer Science, 1988a.

Dana Angluin. Queries and Concept Learning. *Machine Learning*, 2(4):319–342, 1988b.

Yannan Bai, Debmalya Panigrahi, and Ian Zhang. Language Generation in the Limit: Noise, Loss, and Feedback. *arXiv preprint arXiv:2507.15319*, 2025.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional AI: Harmlessness from AI Feedback. *arXiv preprint arXiv:2212.08073*, 2022.

Paul F. Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep Reinforcement Learning from Human Preferences. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Herbert B. Enderton. *A Mathematical Introduction to Logic*. Academic Press, 2001.

Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, et al. Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned. *arXiv preprint arXiv:2209.07858*, 2022.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models. In *Findings of the Association for Computational Linguistics: EMNLP*, 2020.

E. Mark Gold. Language Identification in the Limit. *Information and Control*, 10(5):447–474, 1967.

Sally A. Goldman and Michael J. Kearns. On the Complexity of Teaching. *Journal of Computer and System Sciences*, 50(1):20–31, 1995. doi: 10.1006/jcss.1995.1003.

Kai Greshake, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. Not What You've Signed Up For: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection. In *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security*, AISec '23, pp. 79–90, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400702600. doi: 10.1145/3605764.3623985. URL https://doi.org/10.1145/3605764.3623985.

Tibor Hegedüs. Generalized Teaching Dimensions and the Query Complexity of Learning. In *Proceedings of the Eighth Annual Conference on Computational Learning Theory (COLT)*, pp. 108–117, 1995. doi: 10.1145/225298.225311.

Neel Jain, Avi Schwarzschild, Yuxin Wen, Gowthami Somepalli, John Kirchenbauer, Ping-yeh Chiang, Micah Goldblum, Aniruddha Saha, Jonas Geiping, and Tom Goldstein. Baseline Defenses for Adversarial Attacks Against Aligned Language Models. *arXiv preprint arXiv:2309.00614*, 2023.

Fengqing Jiang, Zhangchen Xu, Luyao Niu, Zhen Xiang, Bhaskar Ramasubramanian, Bo Li, and Radha Poovendran. ArtPrompt: ASCII art-based jailbreak attacks against aligned LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15157–15173, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.809. URL https://aclanthology.org/2024.acl-long.809/.

Alkis Kalavasis, Anay Mehrotra, and Grigoris Velegkas. On the Limits of Language Generation: Trade-Offs Between Hallucination and Mode Collapse. In *Proceedings of the 57th Annual ACM Symposium on Theory of Computing (STOC)*, 2025.

Alkis Kalavasis, Anay Mehrotra, and Grigoris Velegkas. On Characterizations for Language Generation: Interplay of Hallucinations, Breadth, and Stability. In *Proceedings of the 37th International Conference on Algorithmic Learning Theory (ALT)*, volume 313 of *Proceedings of Machine Learning Research*, pp. 1–49, 2026.

Jon Kleinberg and Sendhil Mullainathan. Language Generation in the Limit. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards Deep Learning Models Resistant to Adversarial Attacks. In *International Conference on Learning Representations (ICLR)*, 2018.

William Merrill and Ashish Sabharwal. The Parallelism Tradeoff: Limitations of Log-Precision Transformers. *Transactions of the Association for Computational Linguistics (TACL)*, 11:531–545, 2023.

William Merrill and Ashish Sabharwal. The Expressive Power of Transformers with Chain of Thought. In *International Conference on Learning Representations (ICLR)*, 2024.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, et al. Training Language Models to Follow Instructions with Human Feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. Red teaming language models with language models. In *Proceedings of the 2022 conference on empirical methods in natural language processing*, pp. 3419–3448, 2022.

Mrinank Sharma, Meg Tong, Jesse Mu, Jerry Wei, Jorrit Kruthoff, Scott Goodfriend, Euan Ong, Alwin Peng, et al. Constitutional Classifiers: Defending against Universal Jailbreaks across Thousands of Hours of Red Teaming. *arXiv preprint arXiv:2501.18837*, 2025.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How Does LLM Safety Training Fail? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Yotam Wolf, Noam Wies, Oshri Avnery, Yoav Levine, and Amnon Shashua. Fundamental Limitations of Alignment in Large Language Models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 53079–53112, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/wolf24a.html.

Andy Zou, Zifan Wang, J. Zico Kolter, and Matt Fredrikson. Universal and Transferable Adversarial Attacks on Aligned Language Models. *arXiv preprint arXiv:2307.15043*, 2023.

A Notation

Symbol	Meaning
X, K, P	Countable domain Σ^*, target language, generating distribution ($\mathrm{supp}(P) = K$)
S_n	n i.i.d. positive examples from P (training randomness included)
$\mathcal{P}$, $\{p^{(t)}\}$; $\{x^{(i)}\}$, idx	Prompt set and its fixed enumeration; enumeration of X and its index
$G_n(\cdot \mid p; S_n)$	Prompt-conditional generator, a deterministic function of S_n
e_n, Γ_n, μ_n	Prompt-wise error, worst-case error, prefix mass (random through S_n)
γ_n, ρ_n	Deterministic worst-case error envelope and its failure probability
$\alpha_n, \beta_n, m(n), s(n)$	Detectability floor, breadth-failure rate, prompt budget, string horizon
$k(n), t(n), r_n$	Per-prompt sampling depth, test-set size, per-string error bound
$\ell(x)$	Prefix length at which $x \notin K$ is detectable (Assumption 1)
$\widehat{M}_n(x)$	Oblivious membership estimator (Lemma 1)
$\mathcal{O}, \widetilde{\mathcal{O}}, \mathcal{C}_r$	Exact / noisy membership oracle; preference oracle induced by reward r
$\mathcal{H}_n$, $\mathrm{TD}(K; \mathcal{H})$	Version space consistent with S_n; teaching dimension of K within $\mathcal{H}$

Table 2: Recurring notation.

B Extended Background and Discussion

B.1 Further related work

Adversarial robustness. In classification, worst-case robustness is known to be harder than average-case generalization (Madry et al., 2018). Jailbreaking is the generative analogue, and our amplification bounds (Theorem 5) quantify the step from a small worst-case error to a reliable attack. Empirical jailbreak work documents attacks through optimized suffixes (Zou et al., 2023), ASCII-art encodings (Jiang et al., 2024), indirect prompt injection (Greshake et al., 2023), and automated or manual red-teaming (Perez et al., 2022; Ganguli et al., 2022); baseline defenses are brittle (Jain et al., 2023).

Hallucination versus jailbreaking. Kalavasis et al. (2025) formalize hallucination as generation outside a target language K under a fixed distribution. We view jailbreaking as a *worst-case amplification* of hallucination: by considering the supremum over all prompts, interaction exposes failures of vanishing probability that would be statistically invisible in passive generation. The mechanism is a short chain: under passive positive data alone, generation is strictly easier than identification (Kleinberg & Mullainathan, 2024); sampling access to a broad and robust generator yields an approximate membership estimator $\widehat{M}_n$ (Lemma 1); and once membership can be estimated, generation is as hard as identification, so the Gold-type barrier that passive access had avoided reappears (Theorem 1).

B.2 Background on jailbreaking

LLMs are increasingly deployed as interactive systems to respond to user prompts. Despite advances in alignment, these systems remain vulnerable to *jailbreaking*: adversarial prompting designed to elicit outputs that violate intended constraints, such as safety policies, output formats, or tool-use restrictions. Such vulnerabilities arise through systematic prompt injection and adversarial prompting techniques (Greshake et al., 2023; Jiang et al., 2024; Zou et al., 2023). These attacks exploit

the model's sensitivity to input phrasing, enabling users to bypass safety mechanisms even when the model has been explicitly aligned to avoid such behavior (Jain et al., 2023).

Unlike traditional distributional shift, jailbreaks are inherently *interactive*. Users refine prompts based on observed outputs, probing the model to identify weaknesses. This resembles query models in learning theory, where an agent selects inputs to extract information or maximize error, and it parallels adversarial machine learning, where worst-case robustness is harder than average-case generalization (Madry et al., 2018). Prior theoretical work has studied hallucination as the generation of outputs outside a target language K under a fixed data distribution (Kalavasis et al., 2025); we instead consider the supremum over all prompts, which exposes failures that passive evaluation cannot see. Alignment techniques like RLHF aim to reduce harmful outputs by incorporating human preferences (Ouyang et al., 2022), introducing *relative* feedback; Section 5.3 shows that whether such feedback can resolve the safety boundary depends on whether the reward carries boundary information, whereas *absolute* membership feedback always can.

B.3 Terminology and the role of interaction

The word 'adversary' carries a different meaning in the classical identification-in-the-limit literature (Gold, 1967), where it denotes the agent choosing the *presentation order* of the positive examples supplied to the learner. Our prober is a distinct entity: it does not control the training sample S_n, which remains an i.i.d. draw from P, but interacts with the trained generator G_n at deployment time by issuing prompts and observing outputs. We therefore call it the *interactive prober* or *attacker*.

Remark 8 (What interaction changes)**.** *Interactive prompting changes the information-theoretic nature of language generation. By exposing prefix-level probability mass, it allows the generation problem to be reduced to a Statistical Identification in the Limit problem. If the model's worst-case error γ_n vanishes faster than its breadth α_n can be detected within the budget, its outputs determine membership in K, which leads to the impossibility results for robust generation over non-identifiable classes.*

B.4 Prompt-conditional generation as a closed-loop system

We make the interaction between generator and prober explicit, which sharpens the connection between generation and identification.

Joint execution. Let Γ_1 denote the generating algorithm (G_n) and Γ_2 a prober. Conditional on S_n, their interaction over T rounds induces a trajectory

$$\pi = \{(p_t, \phi_t)\}_{t=1}^{T}, \qquad \phi_t \sim G_n(\cdot \mid p_t; S_n), \qquad p_t = \Gamma_2(\phi_1, \ldots, \phi_{t-1}).$$

We refer to this coupled process as the *joint execution* $\Gamma_1 \bowtie \Gamma_2$. The object of study is the closed-loop system formed by the generator and prober, not the generator in isolation.

Remark 9 (Proof-theoretic interpretation)**.** *Following the classical notion of proof systems (Enderton, 2001), one may interpret Γ_1 as a derivation system in which (i) prompts act as premises, (ii) generated outputs act as derived conclusions, and (iii) the target language K defines semantic validity.* Soundness *then corresponds to jailbreak robustness, $\Gamma_n(S_n) \xrightarrow{p} 0$ (Definition 4), and* completeness *to breadth: for every $x \in K$, $\Pr_{S_n}[\exists p \in \mathcal{P} : G_n(x \mid p; S_n) > 0] \to 1$. Both are quantified over the random sample, with x fixed outside the probability. Full-support generators are trivially complete in this sense, which is why our results use the quantitative version of Definition 5, with a mass threshold inside a prompt budget.*

Prompt-conditional consistency. A weaker notion than robustness requires correctness only prompt by prompt: for every (K, P) and every fixed $p \in \mathcal{P}$, $e_n(p; S_n) \xrightarrow{p} 0$. This allows the rate to depend on p, so a prober searching over prompts may still find failures; it corresponds to standard generalization rather than worst-case robustness. Consistency (Definition 6), averaged over a training prompt distribution, is weaker still.

Computational structure of iterative generators. Iterative generators, which produce outputs token by token, have a well-understood structure: the membership oracle problem for their support

is decidable (Kalavasis et al., 2025), and multi-step decoding increases the expressivity of Transformers (Merrill & Sabharwal, 2024; 2023). Our reduction uses neither property. It needs only the ability to draw samples, and it treats the model as a black box; the internal structure matters for the *computational* cost of an attack, not for the information-theoretic statements proved here.

From support to mass. A natural first attempt at an interactive membership test declares $x \in K$ whenever x lies in the support of $G_n(\cdot \mid p_t; S_n)$ for some probed prompt. For full-support models this test answers 'yes' for every string and carries no information (Appendix C.1). The estimator of Lemma 1 instead counts *observed* output prefixes over a fixed batch of prompts and samples, so it measures mass rather than support. It is correct when valid strings have mass at least α_n at some budgeted prompt and invalid strings have mass at most γ_n at every prompt, with γ_n small relative to $\alpha_n/m(n)$. In the passive setting generation is strictly easier than identification (Kleinberg & Mullainathan, 2024); with sampling access to a broad and robust generator, the two become equivalent in our model.

B.5 When is prefix-closedness natural?

Prefix-closedness is satisfied in a range of practically relevant settings, particularly when outputs obey syntactic or structural constraints: code that must parse, JSON that must match a schema, tool calls and structured actions in agentic systems, or templated formats, where a violation manifests at a definite position and cannot be undone by later tokens. It is also the natural reading of many safety specifications: once a model has emitted a harmful instruction, disclosed a secret, or executed an unauthorized action, no continuation restores safety. Empirical analyses of red-teaming suggest that many harmful or policy-violating generations exhibit identifiable early-token patterns (Gehman et al., 2020; Ganguli et al., 2022; Perez et al., 2022; Zou et al., 2023), which is consistent with, though it does not establish, this view.

However, prefix-closedness does not hold universally. In settings where validity depends on global or semantic properties, such as factual correctness or long-range coherence, invalid strings may share arbitrarily long prefixes with valid ones, making them indistinguishable from finite observations. In such cases the estimator of Lemma 1 is no longer reliable, and the reduction underlying our impossibility results does not go through. Our results therefore apply to language classes where violations are irreversible and locally observable, a regime covering many structured and safety-critical applications, but excluding intrinsically non-local notions of correctness.

B.6 On the two robustness notions

Jailbreak robustness in Definition 4 is pointwise in (K, P) and says nothing about rates. The envelope strengthens it in two ways: it fixes a rate γ_n at which the worst-case error vanishes, and it requires this rate, together with the exceptional probability ρ_n, to hold uniformly over the class. Both are used only in the impossibility results; the positive results in Section 5 establish the pointwise notion or an expectation bound. Note that $\Gamma_n(S_n)$ is random because the generator is trained on S_n; a statement such as "$\sup_p e_n(p) \to 0$" without reference to S_n is not well defined, which is why every robustness statement in this paper names the law under which it holds.

B.7 Each property is attainable in isolation

Remark 10 (Each property is attainable in isolation)**.** *The* print *generator $G_n(\cdot \mid p^{(t)}) = \delta_{x^{(t)}}$ has uniform detectable breadth with $\alpha_n = 1$, $\beta_n = 0$, $m(n) = s(n)$, but is not robust. The* copy *generator $G_n(\cdot \mid p) = \mathrm{Unif}(S_n)$ has $\Gamma_n \equiv 0$ but is not uniformly broad (take P giving small mass to all extensions of some $x \in K$). The results below concern the combination.*

B.8 Interpreting the Gap Condition

The Gap Condition compares two quantities that a deployed system trades against each other. The floor α_n measures how visibly the generator covers valid strings within the budget; the envelope γ_n measures how much it leaks in the worst case. If a model's breadth is 'thin', in the sense that the least likely valid outputs within the horizon have tiny mass, then (2) requires a generator that is even safer than it is visible, and the theorem constrains only such 'hyper-safe' generators. Conversely, if we

demand robustness against a prober with a large budget (e.g. $m(n) \sim n$) while maintaining a fixed breadth floor, Proposition 3 shows that the worst-case leakage cannot vanish faster than $\alpha_n/m(n)$ uniformly. Either way, within the model, a generator trained from positive data cannot be much safer than it is visibly broad without its outputs revealing the boundary of K, which is impossible for non-identifiable classes.

B.9 Refusal and prompt-specific failures

Propositions 2 and 3 offer a formal explanation of the 'refusal behavior' observed in aligned LLMs. To achieve robustness without a safety membership oracle, a generator must keep its detectable breadth thin relative to its error, which appears as refusal or over-conservative behavior on inputs whose validity the training data do not settle. A generator pushed toward breadth must leak in the worst case at a rate that probing can exploit, which appears as prompt-specific failures that are rare under the training distribution. The recurring cycle of jailbreaking and patching can thus be read as a property of generators that try to remain helpful (broad) and safe (robust) while learning from positive-only data. The propositions make this reading precise within the model; measuring where deployed systems lie relative to the frontier is an empirical question they help to pose.

Stronger versions are tempting but false in general. One might claim that a robust generator must contract its support to $\bigcap\{L \in \mathcal{L} : S_n \subseteq L\} \subsetneq K$; on $\mathcal{L}_{\text{chain}}$ with $K = L_k$ this intersection equals L_j for the longest sample a^j, which equals K as soon as a^k is observed, so strict containment fails with probability tending to one. Likewise, a broad generator need not place mass on a specific string outside K at every n; the precise statement is the failure of every envelope with $\gamma_n = o(\alpha_n/m(n))$ along infinitely many n (Proposition 3).

B.10 Relationship to Gold identifiability

Identifiability from positive presentations (Gold, 1967) implies SIL: a tell-tale-based learner (Angluin, 1980) succeeds with probability tending to one, since an i.i.d. sample from P with $\text{supp}(P) = K$ eventually contains any fixed finite tell-tale set for K. Conversely, stochastic presentations do not help for classes that violate Angluin's condition (Angluin, 1988a); for the concrete class used in this paper we prove non-identifiability directly (Proposition 1). The Gold-versus-Angluin separation between positive presentations and membership queries is mirrored in our setting: positive data alone cannot identify $\mathcal{L}_{\text{chain}}$, whereas membership queries on a growing test set can (Theorem 2(b)).

B.11 Preference feedback and red-teaming

Within the idealized model, Theorem 4 offers one possible account of why preference-trained models can show strong average-case safety yet remain vulnerable to red-teaming (Ganguli et al., 2022; Perez et al., 2022): if the reward being optimized does not track the boundary of K on the strings where the data leave it uncertain, preference optimization can expand support toward 'helpfulness' without clipping it at the safety boundary. Proposition 5 shows the other side: comparisons against a known-safe reference recover membership whenever the reward's level sets coincide with K. More generally, the analysis suggests that any practical mechanism capable of certifying the safety boundary, such as a classifier that rejects outputs or a rule-based constraint on structured outputs, supplies information qualitatively closer to membership feedback than purely ordinal preference feedback. This is a statement about kinds of information; how closely particular pipelines approximate either extreme is an empirical question.

B.12 Safety through obscurity

Corollary 2 shows that the security of a model in our setting depends on its worst-case error Γ_n and on the membership-query complexity of the class. If the class has low query complexity, an attacker who can estimate membership can 'triangulate' the boundary with few estimates, and the total cost of an attack is the discovery cost plus the amplification cost $\Theta(1/\Gamma_n)$. This gives a formal sense in which hiding a structured safety boundary is not the same as removing it: as long as the generator's support is broad and the boundary has low query complexity, sampling access makes the boundary learnable. The converse caveat is equally important: for classes with high query complexity, or

where each estimate is computationally expensive, obscurity may act as a practical defense, which our information-theoretic analysis does not rule out.

B.13 Guardrails, limitations, and future work

Architectural guardrails. The robustness–breadth trade-off implies that as generators become visibly broader, making them robust from positive data alone becomes harder. This supports modular safety architectures, in which an external, possibly less expressive, filter or guardrail acting as the membership oracle of Section 5 monitors the outputs of a high-expressivity generator. Theorem 2(a) shows that an exact verifier removes the barrier trivially, and Theorem 3 that noisy verifiers suffice if they satisfy (3). Our results specify what such a component must provide, namely boundary information of membership type, which gives a target for designing and evaluating verifiers; the worst-case behavior of a verifier under probing can itself be analyzed within the same framework.

Scope and future directions. Several features of the framework point to natural extensions. First, prompts are modelled as a countable set, which matches the token-sequence interface through which models are queried; extending the analysis to continuous prompt representations is a direction for future work. Second, the estimator of Lemma 1 is information-theoretic, which makes the impossibility independent of the attacker's computational resources; a complementary computational theory, asking when locating a failure-inducing prompt is tractable, would refine the discovery costs of Section 6 and give a precise account of 'safety through obscurity'. Third, the class-uniform rates that allow a single identification procedure suggest a natural generalization to (K, P)-dependent rates (Remark 4). Fourth, the Gap Condition is sufficient for the reduction, and characterizing the sharp threshold between γ_n and $\alpha_n/m(n)$ is an appealing open question. Finally, the ratio $\gamma_n m(n)/\alpha_n$ is directly measurable: estimating the prefix masses μ_n on structured-output tasks, where prefix-closedness holds, would connect the theory to model scaling and to the evaluation of practical verifiers.

C Additional Illustrations

C.1 Why support-based membership tests fail

Let $\widetilde{M}_n(x) := \mathbb{1}\{\exists t \le T(n) : x \in \operatorname{supp} G_n(\cdot \mid p^{(t)}; S_n)\}$ be the support-based test of Appendix B.4. If each $G_n(\cdot \mid p; S_n)$ has full support on X, as for a softmax-based iterative generator with a positive end-of-sequence probability at every step, then $\widetilde{M}_n(x) = 1$ for every x and every n, so the test carries no information about K. The same example shows that completeness stated through support is uninformative: every full-support generator is complete, and none is sound unless $K = X$. What is informative is *how much* mass the generator places on prefixes, which the estimator of Lemma 1 measures by sampling; it requires no access to the generator's internals and is implementable against a black-box interface.

C.2 A worked example of the estimator

On $\mathcal{L}_{\text{chain}}$ (Proposition 1) with $K = L_k$ and $x = a^j$: if $j > k$, every output beginning with a^j is itself a violation, so each probe sample matches with probability at most Γ_n; if $j \le k$, breadth supplies a budgeted prompt whose outputs begin with a^j with probability at least α_n. Counting matches over the batch separates the two cases whenever $1/\alpha_n \ll k(n) \ll 1/(m(n)\gamma_n)$.

C.3 A numerical illustration of the estimator window

To illustrate Lemma 1 and the role of the Gap Condition, we simulate the estimator on $\mathcal{L}_{\text{chain}}$ with target $K = L_{20}$. We use a synthetic generator with $m = 25$ budgeted prompts. Under prompt $p^{(t)}$ it outputs the valid string $a^{t \bmod 21}$ with probability $\alpha = 0.05$, the invalid string $a^{21+(t \bmod 5)}$ with probability $\gamma = 10^{-4}$, and the empty string otherwise. Every $x \in K$ therefore has prefix mass at least α at some budgeted prompt, and every prompt has error γ, so the window of the proof of Theorem 1 is $1/\alpha = 20 \ll k \ll 1/(m\gamma) = 400$. For each sampling depth k we run the estimator on the 21 valid strings $a^0, \ldots, a^{20}$ and the 5 invalid strings $a^{21}, \ldots, a^{25}$, averaging over 500 independent repetitions (random seed 0).

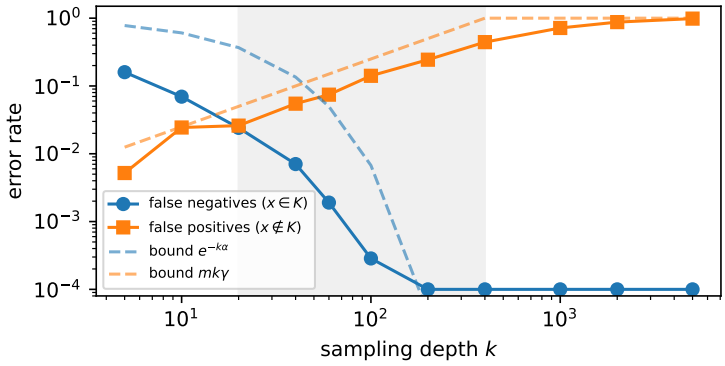


Figure 2: Error rates of the membership estimator $\widehat{M}_n$ as a function of the sampling depth k on the chain class ($K = L_{20}$, $m = 25$, $\alpha = 0.05$, $\gamma = 10^{-4}$; 500 repetitions). Dashed lines are the bounds of Lemma 1 (with $\beta = \rho = 0$), and the shaded band is the window $[1/\alpha, 1/(m\gamma)]$ of the proof of Theorem 1. False negatives fall and false positives rise with k, and both are small only inside the window, which is non-empty because $m\gamma/\alpha = 0.05$. The simulation illustrates the mechanism of the lemma.

k	false-negative rate	bound $e^{-k\alpha}$	false-positive rate	bound $mk\gamma$
5	0.160	0.779	0.005	0.013
20	0.024	0.368	0.026	0.050
60	0.002	0.050	0.074	0.150
100	0.0003	0.007	0.141	0.250
400	0.000	0.000	0.443	1.000
2000	0.000	0.000	0.875	1.000

Table 3: Selected values from Figure 2. Rates are averaged over strings and repetitions.

Figure 2 and Table 3 show the two failure modes that the window balances. For small k the estimator misses valid strings (false negatives decay like $e^{-k\alpha}$); for large k it accumulates matches from rare violations (false positives grow like $mk\gamma$). Both empirical rates respect the bounds of Lemma 1, and both are small only inside the window, which is non-empty here because $m\gamma/\alpha = 0.05$ is small. As the ratio $m\gamma/\alpha$ approaches one the two curves meet and no depth yields a reliable estimator; this is the regime in which Theorem 1 makes no claim. The illustration uses a fixed n and a hand-designed generator; it makes the mechanism concrete and is not evidence about trained models.

D Proofs

Throughout, $K = L_{i^\star} \in \mathcal{L}$ and P with $\mathrm{supp}(P) = K$ are fixed but unknown, and every probability is taken over S_n and, where indicated, over probe samples, oracle answers, or a learner's internal randomness.

D.1 Proofs for Section 4

Proof of Lemma 1. Fix n, K, P and $x \in X$. Conditional on S_n, the generator $G_n(\cdot \mid \cdot\,; S_n)$ is fixed and the $m(n)k(n)$ probe samples $y^{(t,j)}$ are independent, with $y^{(t,j)} \sim G_n(\cdot \mid p^{(t)}; S_n)$.

Case $x \notin K$. By Remark 2, K is prefix-closed, so every string that begins with x lies outside K: if some $y \in K$ began with x, then x would be a prefix of a member and hence a member. Equivalently, by Assumption 1 with $\ell(x) \le |x|$, any y with $y_{1:|x|} = x$ also satisfies $y_{1:\ell(x)} = x_{1:\ell(x)}$ and is therefore invalid. Consequently, for every prompt,

$$\mu_n\big(x; p^{(t)}, S_n\big) \;=\; \Pr_{y \sim G_n(\cdot \mid p^{(t)}; S_n)}[y \text{ begins with } x] \;\le\; e_n\big(p^{(t)}; S_n\big) \;\le\; \Gamma_n(S_n).$$

Let $E := \{\Gamma_n(S_n) \le \gamma_n\}$. On E, a union bound over the $m(n)k(n)$ samples gives

$$\Pr\left[\widehat{M}_n(x) = 1 \,\middle|\, S_n\right] \;\le\; \sum_{t=1}^{m(n)} \sum_{j=1}^{k(n)} \mu_n\big(x; p^{(t)}, S_n\big) \;\le\; m(n)k(n)\gamma_n.$$

By the uniform envelope, $\Pr_{S_n}[E^c] \le \rho_n$. Taking expectations over S_n,

$$\Pr[\widehat{M}_n(x) = 1] \;\le\; \Pr_{S_n}[E^c] + \mathbb{E}_{S_n}\left[\mathbb{1}_E \Pr[\widehat{M}_n(x) = 1 \mid S_n]\right] \;\le\; \rho_n + m(n)k(n)\gamma_n.$$

Case $x \in K$ and $\mathrm{idx}(x) \le s(n)$. Let $B := \{\exists\, t \le m(n) : \mu_n(x; p^{(t)}, S_n) \ge \alpha_n\}$, an event depending only on S_n. By Definition 5, $\Pr_{S_n}[B^c] \le \beta_n$. On B, let $t^\star = t^\star(S_n) \le m(n)$ be the smallest such index. The $k(n)$ samples drawn from $p^{(t^\star)}$ are independent given S_n, and each begins with x with probability at least α_n. Since $\widehat{M}_n(x) = 0$ requires that none of them does,

$$\Pr\left[\widehat{M}_n(x) = 0 \,\middle|\, S_n\right] \;\le\; \big(1 - \alpha_n\big)^{k(n)} \;\le\; e^{-k(n)\alpha_n} \qquad \text{on } B,$$

using $1 - u \le e^{-u}$. Hence $\Pr[\widehat{M}_n(x) = 0] \le \Pr_{S_n}[B^c] + e^{-k(n)\alpha_n} \le \beta_n + e^{-k(n)\alpha_n}$.

Convergence. Fix x. If $x \notin K$, the bound in the first case tends to 0 because $\rho_n \to 0$ and $m(n)k(n)\gamma_n \to 0$. If $x \in K$, then $\mathrm{idx}(x) \le s(n)$ for all large n since $s(n) \to \infty$, and the bound in the second case tends to 0 because $\beta_n \to 0$ and $k(n)\alpha_n \to \infty$. Hence $\widehat{M}_n(x) \xrightarrow{p} \mathbb{1}\{x \in K\}$. □

Proof of Theorem 1. Assume for contradiction that some (G_n) has a uniform robustness envelope (γ_n, ρ_n) and uniform detectable breadth (m, s, α, β) with $\gamma_n = o(\alpha_n / m(n))$, possibly only for $n \ge N_0$. We construct a learner that identifies $\mathcal{L}$ statistically.

Step 1: Choice of the sampling depth. With the convention $1/0 = \infty$, set

$$k(n) := \left\lceil \alpha_n^{-1} \min\left\{n,\; \big(\alpha_n / (m(n)\gamma_n)\big)^{1/2}\right\}\right\rceil.$$

Then $k(n)\alpha_n \ge \min\{n, (\alpha_n/(m(n)\gamma_n))^{1/2}\} \to \infty$ by (2), and, since $\lceil u \rceil \le u + 1$,

$$m(n)k(n)\gamma_n \;\le\; \left(\frac{m(n)\gamma_n}{\alpha_n}\right)^{1/2} + m(n)\gamma_n \;\longrightarrow\; 0,$$

where the second term vanishes because $m(n)\gamma_n = o(\alpha_n)$ and $\alpha_n \le 1$. The window $1/\alpha_n \ll k(n) \ll 1/(m(n)\gamma_n)$ is thus non-empty exactly because of the Gap Condition, and the choice of $k(n)$ depends only on the class-uniform $\alpha_n, m(n), \gamma_n$, not on (K, P).

Step 2: Per-string error. Let

$$r_n := \max\left\{\rho_n,\; m(n)k(n)\gamma_n,\; \beta_n,\; e^{-k(n)\alpha_n}\right\} \longrightarrow 0.$$

By Lemma 1, for every $n \ge N_0$, every (K, P), and every x with $\mathrm{idx}(x) \le s(n)$,

$$\Pr\left[\widehat{M}_n(x) \ne \mathbb{1}\{x \in K\}\right] \;\le\; \max\left\{\rho_n + m(n)k(n)\gamma_n,\; \beta_n + e^{-k(n)\alpha_n}\right\} \;\le\; 2r_n.$$

Step 3: Test set and union bound. Let $t(n) := \min\{s(n), n, \lfloor r_n^{-1/2} \rfloor\}$. Since $s(n) \to \infty$ and $r_n \to 0$, $t(n) \to \infty$, and $t(n)$ is class-uniform. Let $T_n := \{x^{(1)}, \ldots, x^{(t(n))}\}$. The learner draws a *single* batch of $m(n)k(n)$ probe samples $y^{(t,j)} \sim G_n(\cdot \mid p^{(t)}; S_n)$, $t \le m(n)$, $j \le k(n)$, and evaluates $\widehat{M}_n(x)$ in (1) for every $x \in T_n$ from this same batch. For each fixed x, the batch has exactly the law assumed in Lemma 1, so the per-string bound of Step 2 applies; the estimates for different strings are dependent, but a union bound requires no independence. Every $x \in T_n$ satisfies $\mathrm{idx}(x) \le t(n) \le s(n)$, and each x is either in K (bound (b)) or not (bound (a)), so

$$\begin{aligned}\Pr\left[\exists x \in T_n : \widehat{M}_n(x) \ne \mathbb{1}\{x \in K\}\right] \;&\le\; t(n)\Big[\underbrace{\rho_n + m(n)k(n)\gamma_n}_{\text{false positives}} + \underbrace{\beta_n}_{\text{breadth fails}} + \underbrace{e^{-k(n)\alpha_n}}_{\text{false negatives}}\Big] \\ &\le\; 4\,t(n)\,r_n \;\le\; 4r_n^{1/2},\end{aligned}$$

using $t(n) \le r_n^{-1/2}$. Let F_n be the complement of this event, i.e. the event that $\widehat{M}_n(x) = \mathbb{1}\{x \in K\}$ for all $x \in T_n$; then $\Pr[F_n^c] \le 4r_n^{1/2} \to 0$.

Step 4: The elimination identifier. Given S_n, the learner (i) computes the generator $G_n(\cdot \mid \cdot\,; S_n)$, which it can do because the generating algorithm is fixed and known; (ii) draws the probe batch of Step 3 with its own internal randomness; (iii) evaluates $\widehat{M}_n$ on T_n; and (iv) outputs

$$A_n := \min\big\{i \le n : S_n \subseteq L_i \text{ and } \mathbb{1}\{x \in L_i\} = \widehat{M}_n(x) \ \ \forall x \in T_n\big\},$$

with $A_n := 1$ if this finite set is empty. The learner uses only S_n and its own randomness, and $k(n)$, $t(n)$ and T_n do not depend on (K, P). Restricting to $i \le n$ makes each step a finite computation whenever membership in the L_i is uniformly decidable and G_n can be sampled; for the information-theoretic notion of SIL used here, this is not needed, but it shows that the identifier is an explicit procedure.

Step 5: Analysis of the identifier. Fix (K, P) and let $i^\star$ be the *least* index with $L_{i^\star} = K$. For each $i < i^\star$ we have $L_i \neq K$; let w_i be the string of least index in the symmetric difference $L_i \triangle K$, and let $W := \max_{i<i^\star} \mathrm{idx}(w_i)$ ($W := 0$ if $i^\star = 1$). Both $i^\star$ and W depend only on K. Let n be large enough that $n \ge \max\{N_0, i^\star\}$ and $t(n) \ge W$. On the event F_n:

- index $i^\star$ belongs to the set in Step 4, because $i^\star \le n$, $S_n \subseteq K = L_{i^\star}$ (every sample lies in $\mathrm{supp}(P) = K$), and $\mathbb{1}\{x \in L_{i^\star}\} = \mathbb{1}\{x \in K\} = \widehat{M}_n(x)$ for all $x \in T_n$;
- every $i < i^\star$ is excluded, because $w_i \in T_n$ (as $\mathrm{idx}(w_i) \le W \le t(n)$) and $\mathbb{1}\{w_i \in L_i\} \neq \mathbb{1}\{w_i \in K\} = \widehat{M}_n(w_i)$.

Hence $A_n = i^\star$ on F_n, and therefore

$$\Pr\left[L_{A_n} = K\right] \ \ge\ \Pr[A_n = i^\star] \ \ge\ 1 - \Pr[F_n^c] \ \ge\ 1 - 4r_n^{1/2} \ \longrightarrow\ 1.$$

The probability is over $S_n \sim P^n$ and the learner's internal randomness, as allowed in the definition of SIL.

Step 6: Conclusion. Step 5 holds for every $K \in \mathcal{L}$ and every P with $\mathrm{supp}(P) = K$, with a learner that does not depend on (K, P), so $\mathcal{L}$ is SIL, contradicting the hypothesis. Only indices $n \ge N_0$ were used, which proves the final sentence of the theorem. Each hypothesis enters at one point: prefix-closedness and the envelope control false positives (Lemma 1(a)); breadth, including the budget and the horizon, controls false negatives (Lemma 1(b)); the Gap Condition makes the choice of $k(n)$ in Step 1 possible; class-uniformity makes $k(n)$, $t(n)$ and T_n independent of (K, P); and non-SIL supplies the contradiction. □

Proof of Proposition 1. **Prefix-closedness.** Every prefix of a^i is $a^{i'}$ with $i' \le i$, so each L_k and L_∞ is prefix-closed, and Assumption 1 holds by Remark 2.

Non-identifiability. Suppose a possibly randomized learner A is SIL for $\mathcal{L}_{\mathrm{chain}}$. We construct P with $\mathrm{supp}(P) = a^*$ on which A fails. Let U_k be the uniform distribution on L_k. Set $k_1 := 0$ and $P_1 := U_0$, and proceed recursively for $j \ge 1$: since $\mathrm{supp}(P_j) = L_{k_j}$ and A is SIL, choose $n_j > n_{j-1}$ with

$$\Pr_{S \sim P_j^{n_j}}\left[L_{A_{n_j}(S)} = L_{k_j}\right] \ge 3/4;$$

then set $\varepsilon_j := 2^{-j}/(8n_j)$, $k_{j+1} := k_j + 1$, and $P_{j+1} := (1 - \varepsilon_j)P_j + \varepsilon_j U_{k_{j+1}}$. By induction $\mathrm{supp}(P_{j+1}) = L_{k_{j+1}}$, and $\|P_{j+1} - P_j\|_{\mathrm{TV}} = \varepsilon_j \|U_{k_{j+1}} - P_j\|_{\mathrm{TV}} \le \varepsilon_j$.

Since $\sum_j \varepsilon_j < \infty$, (P_j) is Cauchy in total variation and converges to a distribution P. Each a^i receives positive mass $\varepsilon_i/(i+1)$ at stage $i+1$ (for $i \ge 1$; a^0 at stage 1), and later stages multiply this mass by factors $(1 - \varepsilon_{j'})$ and only add to it, so $P(a^i) \ge \frac{\varepsilon_i}{i+1} \prod_{j'}(1 - \varepsilon_{j'}) > 0$; hence $\mathrm{supp}(P) = a^* = L_\infty$. Moreover, for each j, using $n_i \ge n_j$ for $i \ge j$,

$$\|P - P_j\|_{\mathrm{TV}} \ \le\ \sum_{i \ge j} \varepsilon_i \ =\ \sum_{i \ge j} \frac{2^{-i}}{8n_i} \ \le\ \frac{1}{8n_j} \sum_{i \ge j} 2^{-i} \ \le\ \frac{1}{4n_j}.$$

Total variation between product measures is subadditive, so $\|P^{n_j} - P_j^{n_j}\|_{\mathrm{TV}} \le n_j \|P - P_j\|_{\mathrm{TV}} \le 1/4$. The learner's output is a (randomized) function of its sample, so the probability of any event

concerning the output changes by at most the total variation distance between the sample laws. Therefore

$$\Pr_{S\sim P^{n_j}}\left[L_{A_{n_j}(S)} = L_{k_j}\right] \;\geq\; \frac{3}{4} - \frac{1}{4} \;=\; \frac{1}{2}, \qquad \text{hence} \qquad \Pr_{S\sim P^{n_j}}\left[L_{A_{n_j}(S)} = L_{\infty}\right] \;\leq\; \frac{1}{2}$$

for every j. As $n_j \to \infty$, this contradicts SIL for the pair (L_∞, P). □

Proof of Corollary 1. A generator with all three properties would in particular have a uniform robustness envelope and uniform detectable breadth with $\gamma_n = o(\alpha_n/m(n))$, contradicting Theorem 1. Consistency is not used; it is implied by robustness (Remark 5). □

Proof of Proposition 2 (Refusal). Fix m, s, β, α as in the statement and suppose, for contradiction, that the conclusion fails. Then there is N_0 such that for every $n \geq N_0$, every $K \in \mathcal{L}$, every P with $\mathrm{supp}(P) = K$, and every $x \in K$ with $\mathrm{idx}(x) \leq s(n)$,

$$\Pr_{S_n}\left[\mu_n(x; p^{(t)}, S_n) < \alpha_n \;\; \forall t \leq m(n)\right] \leq \beta_n, \quad \text{i.e.}$$

$$\Pr_{S_n}\left[\exists t \leq m(n) : \mu_n(x; p^{(t)}, S_n) \geq \alpha_n\right] \geq 1 - \beta_n.$$

Thus (G_n) has uniform detectable breadth (m, s, α, β) for all $n \geq N_0$. Together with the uniform envelope and $m(n)\gamma_n = o(\alpha_n)$, which is (2), this contradicts the final sentence of Theorem 1. □

Proof of Proposition 3 (Vulnerability). Fix $\gamma_n = o(\alpha_n/m(n))$ and $\rho_n \to 0$, and suppose, for contradiction, that there is N_0 such that for every $n \geq N_0$ and every (K, P), $\Pr_{S_n}[\Gamma_n(S_n) > \gamma_n] \leq \rho_n$. Then (γ_n, ρ_n) is a uniform robustness envelope for $n \geq N_0$, and together with uniform detectable breadth and the Gap Condition this contradicts Theorem 1. For the last claim, on the event $\Gamma_n(S_n) > \gamma_n$, Theorem 5 gives $Q_n(S_n) \leq \lceil 2\ln 2/\Gamma_n(S_n)\rceil \leq \lceil 2\ln 2/\gamma_n\rceil$. □

D.2 Proofs for Section 5

Proof of Theorem 2. **Part (a).** For each p, the normalizer $Z(p) := \sum_{x'\in K}\pi(x' \mid p)$ is positive because $K \neq \emptyset$ and π has full support. Hence $G_n(x \mid p) = \pi(x \mid p)\mathbb{1}\{x \in K\}/Z(p)$ is a well-defined distribution, supported on K, so $e_n(p) = 0$ for every p and $\Gamma_n \equiv 0$, and it is positive on every $x \in K$.

Part (b), Step 1: Identification via elimination. Let $\{L_i\}_{i\geq 1}$ enumerate $\mathcal{L}$ and let $T_1 \subseteq T_2 \subseteq \cdots$ be finite sets with $\bigcup_n T_n = X$. Define the hypothesis set of languages consistent with the data and with the oracle on the test set,

$$H_n := \left\{L_i : \; S_n \subseteq L_i \text{ and } \mathcal{O}(x) = \mathbb{1}\{x \in L_i\} \;\; \forall x \in T_n\right\}, \qquad A_n := \min\{i : L_i \in H_n\}.$$

The index $i^\star$ of K always belongs to H_n, since $S_n \subseteq K$ and $\mathcal{O} = \mathbb{1}_K$, so $A_n \leq i^\star$. For each $i < i^\star$ there is a witness $w_i \in L_i \triangle K$, where $\mathbb{1}\{w_i \in L_i\} \neq \mathcal{O}(w_i)$. Because the T_n increase to X, there is N with $\{w_i : i < i^\star\} \subseteq T_n$ for all $n \geq N$, and then every $i < i^\star$ is excluded from H_n. Hence $A_n = i^\star$ for all $n \geq N$, for every realization of S_n.

Step 2: Filtered prompt-conditional generation. Define

$$G_n(x \mid p; S_n) := \frac{\pi(x \mid p)\,\mathbb{1}\{x \in L_{A_n}\}}{\sum_{x'\in L_{A_n}}\pi(x' \mid p)}, \tag{4}$$

which is well defined whenever $L_{A_n} \neq \emptyset$ (in particular when $A_n = i^\star$); otherwise set $G_n(\cdot \mid p; S_n) := \pi(\cdot \mid p)$. This generator is conditioned on the prompt while being gated by the hypothesized boundary.

Step 3: Robustness and coverage. On $\{A_n = i^\star\}$ we have $L_{A_n} = K$, so for every prompt

$$\Gamma_n(S_n) \;=\; \sup_{p\in\mathcal{P}}\sum_{x\notin K} G_n(x \mid p; S_n) \;=\; 0,$$

$$G_n(x \mid p; S_n) \;=\; \frac{\pi(x \mid p)}{\sum_{x'\in K}\pi(x' \mid p)} \;>\; 0 \quad \forall x \in K.$$

Since $A_n = i^\star$ for all $n \geq N$, $\Pr_{S_n}[\Gamma_n(S_n) = 0] = 1$ and $\Pr_{S_n}[\forall p : G_n(x \mid p; S_n) > 0] = 1$ for every $x \in K$ and $n \geq N$, which proves both claims. □

Proof of Proposition 4. **Lower bound.** Suppose the learner, on sample S_n, queries a set T (chosen adaptively or not) and then determines K with certainty. If some $L \in \mathcal{H}_n$ with $L \neq K$ satisfied $L \cap T = K \cap T$, then L would be consistent with S_n (as $S_n \subseteq L$) and with every oracle answer, so the learner's information would be identical under targets K and L and it could not determine K with certainty. Hence T is a teaching set for K within $\mathcal{H}_n$, and $|T| \geq \mathrm{TD}(K; \mathcal{H}_n)$.

Comparison and monotonicity. Since $\mathcal{H}_n \subseteq \mathcal{L}$, any set that distinguishes K from all other members of $\mathcal{L}$ also distinguishes it from all other members of $\mathcal{H}_n$, so $\mathrm{TD}(K; \mathcal{H}_n) \leq \mathrm{TD}(K; \mathcal{L})$. Because samples are nested, $S_n \subseteq S_{n+1}$ and therefore $\mathcal{H}_{n+1} \subseteq \mathcal{H}_n$; the same argument gives $\mathrm{TD}(K; \mathcal{H}_{n+1}) \leq \mathrm{TD}(K; \mathcal{H}_n)$ along every sample path.

Upper bound. For a finite class $\mathcal{H}$, Hegedüs (1995) shows that exact identification with membership queries is possible with $O(\mathrm{XTD}(\mathcal{H}) \log |\mathcal{H}|)$ adaptive queries. Applying this to $\mathcal{H}_n$, which contains K, gives the claim. □

Proof of Theorem 3. **Step 1: Majority-vote identification.** Let T_n be the first $\lfloor n^c \rfloor$ strings of the enumeration of X, and for each $x \in T_n$ let $\hat{\mathcal{O}}_n(x)$ be the majority vote of $q(n) := \lceil C \log n \rceil$ independent answers of $\widetilde{\mathcal{O}}(x)$; the total number of queries is $O(n^c \log n)$. Each answer is correct with probability at least $1 - \eta > 1/2$, so by Hoeffding's inequality

$$\Pr\left[\hat{\mathcal{O}}_n(x) \neq \mathbb{1}\{x \in K\}\right] \;\leq\; \exp\left(-2(1/2 - \eta)^2 q(n)\right) \;\leq\; n^{-2(1/2-\eta)^2 C}.$$

Choosing $C > (c+1)/\left(2(1/2-\eta)^2\right)$ and taking a union bound over T_n,

$$\Pr\left[\exists x \in T_n : \hat{\mathcal{O}}_n(x) \neq \mathbb{1}\{x \in K\}\right] \;\leq\; n^c \cdot n^{-2(1/2-\eta)^2 C} \;\leq\; n^{-1}.$$

Define H_n and A_n as in Step 1 of the proof of Theorem 2, with $\hat{\mathcal{O}}_n$ in place of $\mathcal{O}$. Let N be such that T_n contains the finitely many witnesses w_i, $i < i^\star$, for all $n \geq N$. For $n \geq N$, on the event that all majority votes on T_n are correct, the argument of that proof gives $A_n = i^\star$. Hence

$$\varepsilon_n := \Pr\left[A_n \neq i^\star\right] \;\leq\; n^{-1} \quad \text{for } n \geq N, \qquad \text{so } \varepsilon_n \to 0.$$

Step 2: Prompt-conditional filtering. Define $G_n(\cdot \mid p)$ by (4), with A_n from Step 1.

Step 3: Error analysis. For any prompt p, a violation can occur only if L_{A_n} contains strings outside K. On $\{A_n = i^\star\}$, $L_{A_n} = K$ and the error is zero; in all cases it is at most one. Thus

$$\Gamma_n \;=\; \sup_p e_n(p) \;\leq\; \mathbb{1}\{A_n \neq i^\star\}, \qquad \mathbb{E}_{S_n, \widetilde{\mathcal{O}}}[\Gamma_n] \;\leq\; \Pr[A_n \neq i^\star] \;=\; \varepsilon_n \;\to\; 0,$$

establishing (ε_n)-robustness. On $\{A_n = i^\star\}$, $G_n(x \mid p) > 0$ for every $x \in K$ and every p, so for every $x \in K$, $\Pr_{S_n, \widetilde{\mathcal{O}}}[\forall p : G_n(x \mid p) > 0] \geq 1 - \varepsilon_n$, establishing (δ_n)-coverage with $\delta_n := \varepsilon_n$, uniformly in x. □

Lemma 2 (Boundary-uninformative preferences carry no information)**.** *Let r be boundary-uninformative for $\mathcal{L}$. For every algorithm B that receives S_n and may query $\mathcal{C}_r$ adaptively, there is an algorithm B' that receives only S_n and produces the same output as B on every realization of S_n and of the internal randomness, under every (K, P).*

Proof. Because r is the same function for every $K \in \mathcal{L}$, so is $\mathcal{C}_r$: $\mathcal{C}_r(x, y) = x$ iff $r(x) \geq r(y)$. Let B' run B with the same S_n and internal randomness and answer each query (x, y) of B by evaluating this fixed function. By induction on the number of queries, the transcript of B' coincides with that of B under every (K, P), and so do the outputs. □

Proposition 6 (Preference Indistinguishability)**.** *If $\mathcal{L}$ is not SIL and r is boundary-uninformative for $\mathcal{L}$, then no learner using $(S_n, \mathcal{C}_r)$ is SIL for $\mathcal{L}$. In particular, on $\mathcal{L}_{\mathrm{chain}}$ with the safety-consistent reward r_0, preference answers do not help to decide whether the target is L_∞ or some L_k: every learner using $(S_n, \mathcal{C}_{r_0})$ fails, for some P with $\mathrm{supp}(P) = L_\infty$, to output L_∞ with probability tending to one.*

Proof. If a learner B using $(S_n, \mathcal{C}_r)$ were SIL, the learner B' of Lemma 2 would have the same output on every realization, hence the same success probability under every (K, P), and would be SIL using S_n alone, a contradiction. For $\mathcal{L}_{\text{chain}}$, $r_0 = \mathbb{1}\{x = \epsilon\}$ is boundary-uninformative and safety-consistent, and the construction in the proof of Proposition 1, applied to B', yields P with $\text{supp}(P) = L_\infty$ and $\Pr[L_{A_{n_j}} = L_\infty] \le 1/2$ for infinitely many n_j. □

Proof of Theorem 4. Suppose a generator (G_n) trained on $(S_n, \mathcal{C}_r)$ had a uniform robustness envelope and uniform detectable breadth with $\gamma_n = o(\alpha_n/m(n))$. The training procedure is an algorithm that receives S_n and queries $\mathcal{C}_r$, and its output is the family $\{G_n(\cdot \mid p; S_n)\}_p$. By Lemma 2 there is a training procedure using only S_n that outputs the same family on every realization of S_n and of the training randomness. The resulting generator (G'_n) satisfies $G'_n = G_n$ pathwise, so it has the same envelope and the same uniform detectable breadth, and it satisfies the Gap Condition. Since $\mathcal{L}$ is non-SIL and prefix-closed, this contradicts Theorem 1. The reward r_0 is boundary-uninformative because $\bigcap_{L \in \mathcal{L}} L$ does not depend on K, and safety-consistent because $r_0(x) = 1$ implies $x \in \bigcap_L L \subseteq K$. □

Proof of Proposition 5. Let $y \in S_n \subseteq K$, so $r(y) = 1$. If $x \in K$, then $r(x) = 1 \ge r(y)$ and $\mathcal{C}_r(x, y) = x$; if $x \notin K$, then $r(x) = 0 < 1 = r(y)$ and $\mathcal{C}_r(x, y) = y$. Hence $\mathbb{1}\{\mathcal{C}_r(x, y) = x\} = \mathbb{1}\{x \in K\} = \mathcal{O}(x)$. If ties are instead broken uniformly at random, then for $x \in K$ each comparison returns x with probability $1/2$, and for $x \notin K$ it never does. Declaring $x \in K$ iff some of $\lceil \log_2 n \rceil + 1$ independent comparisons returns x errs only when $x \in K$ and all comparisons return y, which has probability at most $2^{-\lceil \log_2 n \rceil - 1} \le 1/(2n)$. The simulated oracle can therefore replace $\mathcal{O}$ in Theorem 2(b); with random tie-breaking, a union bound over T_n applies as in the proof of Theorem 3 after choosing T_n of size $o(n)$. □

D.3 Proofs for Section 6

Proof of Theorem 5. Fix a realization of S_n with $\Gamma_n := \Gamma_n(S_n) > 0$.

Upper bound. By the definition of the supremum there is p^* with $e_n(p^*; S_n) \ge \Gamma_n/2$. A prober that queries p^* in every round receives independent outputs, each invalid with probability at least $\Gamma_n/2$, so

$$\Pr\left[\forall t \le T : \phi_t \in K \,\middle|\, S_n\right] \;\le\; (1 - \Gamma_n/2)^T \;\le\; e^{-T\Gamma_n/2} \;\le\; \tfrac{1}{2} \qquad \text{whenever } T \ge 2\ln 2/\Gamma_n.$$

Hence $Q_n(S_n) \le \lceil 2 \ln 2/\Gamma_n \rceil$.

Lower bound. For any prober, possibly adaptive, let $E_t := \{\phi_t \notin K\}$. Conditional on S_n and on the history $\phi_1, \dots, \phi_{t-1}$, the prompt p_t is determined and $\Pr[E_t \mid S_n, \phi_{1:t-1}] = e_n(p_t; S_n) \le \Gamma_n$; hence $\Pr[E_t \mid S_n] \le \Gamma_n$. By a union bound,

$$\Pr\left[\exists t \le T : E_t \,\middle|\, S_n\right] \;\le\; \sum_{t=1}^{T} \Pr[E_t \mid S_n] \;\le\; T\,\Gamma_n,$$

which is at least $1/2$ only if $T \ge 1/(2\Gamma_n)$. Thus $Q_n(S_n) \ge 1/(2\Gamma_n)$.

Envelope. On $\{\Gamma_n(S_n) \le \gamma_n\}$, which has probability at least $1 - \rho_n$, the lower bound gives $Q_n(S_n) \ge 1/(2\Gamma_n) \ge 1/(2\gamma_n)$. □

Proof of Corollary 2. **(a) Simulation of a query algorithm.** Let $\mathcal{M}$ be a deterministic membership-query algorithm. Run with exact answers $\mathcal{O} = \mathbb{1}_K$, it issues a fixed sequence of queries $z_1, \dots, z_q$ determined by K (and by S_n if $\mathcal{M}$ uses it). Now run $\mathcal{M}$ with each answer replaced by $\widehat{M}_n$, computed from fresh probe samples. By induction on j: as long as the answers to $z_1, \dots, z_{j-1}$ were correct, the next query is z_j. Hence the two runs coincide on the event $\bigcap_{j \le q} \{\widehat{M}_n(z_j) = \mathbb{1}\{z_j \in K\}\}$. Each z_j is a fixed string within the horizon, so by Step 2 of the proof of Theorem 1 its estimate errs with probability at most $2r_n$, and a union bound gives failure probability at most $2qr_n$. Each estimate uses $m(n)k(n)$ probe samples, for $q\,m(n)k(n)$ in total.

(b) Lower bound. This is the lower bound of Proposition 4; for $S_n = \emptyset$ it is the teaching dimension of Goldman & Kearns (1995).

(c) Oblivious versus adaptive on the finite chain. For $1 \le j \le N$, the languages L_{j-1} and L_j differ only at a^j, so any query set that distinguishes every pair of members must contain $a^1, \dots, a^N$ and has size at least N. Conversely, since $a^j \in L_k$ iff $j \le k$, the answers are monotone in j, and binary search over $j \in \{1, \dots, N\}$ determines $k \in \{0, \dots, N\}$ with $\lceil \log_2(N+1) \rceil$ adaptive queries. $\square$