

On the Limits of Language Generation: Trade-Offs Between Hallucination and Mode Collapse

Alkis Kalavasis

Yale University

alkis.kalavasis@yale.edu

Anay Mehrotra

Yale University

anaymehrotra1@gmail.com

Grigoris Velegkas

Yale University

grigoris.velegkas@yale.edu

Abstract

Specifying all desirable properties of a language model is challenging, but certain requirements seem essential for any good model. Given samples drawn from an unknown language, the trained model should (1) produce valid strings that have not been seen in the training data, and (2) be expressive enough to capture the full richness of the language. Otherwise, if the language model outputs invalid strings, it “hallucinates,” and if it fails to capture the full range of the language, it suffers from “mode collapse.” In this paper, we ask whether it is possible for a language model to meet both of these requirements.

We investigate this question within a statistical setting of language generation, building on the seminal works of Gold [Gol67, Inf. Control], Angluin [Ang79, STOC], and Angluin [Ang88, Tech. Report]. In this setting, the language model is presented with randomly sampled strings from a distribution supported on an unknown language K , which is only known to belong to a possibly infinite collection of candidate languages. The goal of the model is to generate unseen strings from this target language. We say that the language model generates from K with consistency and breadth if, as the size of the training set increases, the set of strings it can output converges to the set of all unseen strings in K .

Kleinberg and Mullainathan [KM24, NeurIPS] posed an open question of whether consistency and breadth in language generation are both possible. We answer this question negatively: for a large class of language models – including next-token-prediction-based models – this is impossible for most collections of candidate languages. This contrasts with the recent positive result of Kleinberg and Mullainathan [KM24, NeurIPS], which demonstrated that consistent generation, without requiring breadth, is possible for any countable collection of candidate languages. Our finding highlights that generation with breadth is fundamentally different from generation without breadth.

As a byproduct of our result, we also examine how many samples are required for generation with or without breadth, establishing near-tight bounds on the “learning curves” for generation in the statistical framework of Bousquet, Hanneke, Moran, van Handel, and Yehudayoff [BHM+21, STOC].

Finally, our results also give some hope for consistent generation with breadth: it is achievable for any countable collection of languages when negative examples – in the form of strings outside of K – are available in addition to strings inside of K . This suggests that feedback in post-training, which encodes negative examples, can be crucial in reducing hallucinations while also limiting mode collapse.

Contents

1	Introduction	1
1.1	Informal Results	3
1.1.1	Setup and Definitions	4
1.1.2	Main Results	5
1.2	Technical Overview	7
1.3	Additional Results With Relaxation of Consistency and Breadth	12
1.4	Takeaways, Discussion, and Open Problems	14
1.5	Further Related Works	16
2	Model and Preliminaries	19
2.1	Language Identification and Generation in the Limit	20
3	Overview of Results	23
3.1	Results for Identification and Generation Without Breadth	24
3.1.1	Universal Rates: Model and Preliminaries	24
3.1.2	Universal Rates for Identification	25
3.1.3	Universal Rates for Consistent Generation	26
3.2	Results for Generation With Breadth	27
3.2.1	Membership Oracle Problem	27
3.2.2	Results for Generators for Which $\text{MOP}(\cdot)$ Is Decidable	28
3.2.3	A Family of Generators for Which $\text{MOP}(\cdot)$ Is Decidable	29
3.2.4	Results for Generation With Breadth in the Limit	30
3.3	Results for Generation With Approximate Consistency and Breadth	30
3.4	Further Results for Identification	32
3.4.1	Exponential Rates for Identification Using Subset Oracle	32
3.4.2	Exponential Rates for Identification of Finite Collections	33
3.4.3	Exponential Rates for Identification of Collections of Finite Languages	33
3.4.4	Exponential Rates for Identification from Positive and Negative Examples	33
4	Organization of the Rest of the Paper	34
5	Proofs from Section 3.1 (Rates for Identification and Generation)	35
5.1	Proof of Theorem 3.1 (Rates for Identification)	35
5.2	Proof of Theorem 3.2 (Rates for Generation)	46
5.2.1	Optimal Rate for Non-Trivial Collections for Generation	46
5.2.2	A Sufficient Condition To Achieve Exponential Rate	51
5.2.3	Algorithm With Access To Subset Oracle	52
5.2.4	Algorithm With Access To Membership Oracle	53
6	Proofs from Section 3.2 (Generation With Breadth)	55
6.1	Proof of Theorem 3.4 ($\text{MOP}(\cdot)$ Is Decidable For Iterative Generators)	55
6.2	Proof of Theorem 3.3 (Impossibility for Generation With Breadth)	56
6.3	Proof of Theorem 3.5 (Impossibility for Generation With Breadth in the Limit)	59

7	Proofs from Section 3.3 (Generation With Approximate Consistency and Breadth)	60
7.1	Proof of Theorem 3.7 (Impossibility in the Limit)	60
7.2	Proof of Theorem 3.6 (Impossibility in the Statistical Setting)	63
8	Proofs from Section 3.4 (Further Results for Identification)	66
8.1	Proof of Proposition 3.8 (Identification Using Subset Oracle)	66
8.2	Proof of Proposition 3.9 (Identification of Finite Collections)	67
8.3	Proof of Proposition 3.10 (Identification of Collections of Finite Languages)	68
8.4	Proof of Theorem 3.11 (Identification from Positive and Negative Examples)	69
A	Further Discussion on Decidability of $\text{MOP}(\cdot)$	90
B	Results With Subset Oracles	90
B.1	Identification in the Limit Without Tell-Tale Oracle via Subset Oracles	90
B.2	Best-Of-Both Worlds: Generating With Breadth When Possible	92
C	Further Results for Consistent Generation With Approximate Breadth	92
D	Further Comparison With Online Learning	96
E	Borel-Cantelli Lemmas	98

1 Introduction

Language acquisition is a fundamental mystery across multiple scientific fields, ranging from Biology and Neuroscience to Sociology [SAN96; Bre07; Cla14; MIB+24]. Theoretical Computer Scientists have been fascinated by language since the early days of the field: in the 1950s, Turing [Tur50] introduced his famous test using language as an interface to cognition, Shannon [Sha51a] studied statistics of printed English aiming at understanding its entropy and the extent to which it could be compressed, and Mandelbrot [Man53] designed a statistical model to capture connections between language and the brain.

Over the years, language modeling has advanced through simple models, such as the word n -gram model introduced by Shannon [Sha51b] and widely used in natural language processing [BDd+92]. In the early 2000s, neural networks achieved a significant breakthrough in the field [BDV00], leading to fascinating deep learning systems [MKB+10; LBH15; Gol16] built using traditional architectures like Recurrent Neural Networks [RHW86] and Long Short-Term Memory [HS97]. In 2017, the field of language modeling was revolutionized by the introduction of the Transformer architecture [Bah14; SVL14; VSP+17], which led to the development of Large Language Models (LLMs). The achievements of LLMs have been groundbreaking; recent models can perform well on tasks far beyond natural language processing [BCE+23; TLI+23]. Despite their impressive performance, their extensive use has revealed that LLMs exhibit various bizarre behaviors even in seemingly mundane tasks [Bor23].

Perhaps the most well-known issue with current LLMs is *hallucinations*: the models generate false but plausible-sounding text with surprising frequency [JLF+23; ZLC+23].¹ Such hallucinations, highlighted by popular media [WM23], could significantly impact safety, reliability, and user trust as the adoption of these systems extends to new tasks [AOS+16; HCS+22]. The importance of this problem, among other concerns, led both the US [Bid23] and the EU [Sat23] to issue calls for safeguards against misleading outputs generated by LLMs. In this direction, designing LLMs that generate responses *consistent* with the ground truth is an effort that has gained a lot of attention from Machine Learning (ML) practitioners [WWS+22; AP23; GZA+23; HYM+23; JLF+23; FSW+24; KWT+24], policymakers [Bid23; Sat23; SK23], and theorists [HKK+18; KV24; KM24].

If the sole goal is to avoid hallucinations, then, of course, one could simply limit the range of outputs generated by the language model. As an extreme example, consider a language model that only outputs “I am a language model” and, therefore, never hallucinates. However, modern LLMs do not just aim to generate *a few* valid outputs; their goal is to obtain the ability to express *a wide range* of plausible outputs, thus capturing the richness of human language. The key challenge lies in avoiding hallucinations while achieving *breadth*. The problem of achieving consistent generation with breadth is not new in the ML community, dating back at least to the era of Generative Adversarial Networks (GANs) [GPM+20]. In this line of work, *mode collapse* [GPM+20] is the analog of lack of breadth; it refers to the phenomenon where the GAN assigns non-zero mass only to a few modes of the true data distribution, thus producing a limited variety of samples and becoming repetitive [AB17; SSA18; BZW+19]. *The starting point of our work is exactly this puzzling tension between consistent generation and breadth in language generation.*

¹We stress that LLMs outputting wrong facts based on errors in training data (e.g., “The Earth is flat”) or miscalculations (e.g., “1+1 = 3”) do not constitute hallucinations. A hallucination is a plausible but false text with unclear origin (e.g., “Barack Obama was the president of the US and was born on January 1, 1958”).

We start with a mathematical specification inspired by classical work on learning theory, tracing back to the seminal work of Angluin [Ang88], and the recent formulation of Kleinberg and Mullainathan [KM24]: the domain \mathcal{X} is a countable collection of strings, and there is an unknown target language K which is a subset of this domain. We know that the true language lies within a collection of possibly infinite but countably many languages $\mathcal{L} = \{L_1, L_2, \dots\}$. There exists an unknown distribution \mathcal{P} over strings in $K \in \mathcal{L}$ that satisfies $\text{supp}(\mathcal{P}) = K$; any distribution with this property is said to be *valid* for K . The algorithm observes i.i.d. samples from \mathcal{P} and aims to learn how to generate unseen strings from the target language K – this, at a high level, is the language generation problem. Intuitively, the target language K is capturing “facts” of the world; everything that belongs to K is correct, whereas everything outside of K is unambiguously incorrect and can be thought of as a “hallucination.” Observe that K has to be infinite for the problem to be well-defined as, otherwise, at some point, the algorithm will see all possible strings of K and, from then on, would have no unseen strings to generate from.

Let us explore language generation further, with the immediate aim of quantifying an algorithm’s progress toward becoming a useful generator. Consider a generating algorithm \mathcal{G}_n ² that is trained on a set S of n i.i.d. examples from \mathcal{P} . To quantify the inconsistency of \mathcal{G}_n , we need an objective. As discussed above, this objective should penalize \mathcal{G}_n for outputting strings outside of K and for repeating examples already seen in the training data S .³ For a target language K and a model \mathcal{G}_n trained on S , we consider the following generation error

$$\text{gen_er}(\mathcal{G}_n) := \Pr_{S \sim \mathcal{P}^n} [\text{supp}(\mathcal{G}_n) \supset K \setminus S]. \quad (1)$$

In words, a model errs according to $\text{gen_er}(\cdot)$ if it either hallucinates by outputting strings from $\mathcal{X} \setminus K$ or if it outputs something already contained in the training set S . This is inspired by the notion of generation considered by Kleinberg and Mullainathan [KM24]; they call an algorithm a *consistent* generator if its support becomes a subset of $K \setminus S$ after seeing finitely many training examples S . We relax this definition and call an algorithm a *consistent generator for the collection \mathcal{L}* if its error, as defined in Equation (1), asymptotically goes to zero for any valid distribution \mathcal{P} .

Let us now review how prior work has approached issues with language generation algorithms – foremost, hallucination. Under the above statistical setting, Kalai and Vempala [KV24] made important progress showing that calibrated models must hallucinate by lower bounding the hallucination rate by the model’s calibration. For a detailed comparison with our work, we refer to Section 1.5. Closer to our paper, the work of Kleinberg and Mullainathan [KM24] explored language generators that must not hallucinate, *i.e.*, they must be consistent. They studied language generation in an online setting where the data are not drawn from \mathcal{P} but are given as a stream to the learner, *i.e.*, as an *adversarial enumeration* of the strings of the true language K . In their setting, \mathcal{G}_n is said to *generate in the limit* from K if, after some finite time n_0 in the enumeration of K , \mathcal{G}_n is able to generate new unseen strings from K for all subsequent times $n \geq n_0$. They showed

²Formally, a generating algorithm is a sequence of mappings $(\mathcal{G}_n)_{n \in \mathbb{N}}$: for each n , it is a computable mapping from a training dataset of size n to a (computable) distribution (*i.e.*, a sampling algorithm) over \mathcal{X} . We will use the notation $(\mathcal{G}_n)_n$ to refer to the generating algorithm and the notation \mathcal{G}_n or simply \mathcal{G} for the induced distribution (*generator*) after training; hence when we write $x \sim \mathcal{G}_n$ or $\text{supp}(\mathcal{G}_n)$, we refer to the distribution obtained after training.

³When we require generating algorithm to achieve breadth, it is not important to enforce that the support does not contain S . We will elaborate after the formal statement of Definition 4.

that there exists an algorithm that can generate in the limit from *every* countable list of candidate languages.

This result is surprising because it contrasts with strong negative results for the well-studied problem of *language identification in the limit* (where one wants to identify K in the limit and not simply generate from it;⁴ see also Definition 9). The family of languages identifiable in the limit is very limited: the results of Gold [Gol67] and Angluin [Ang79] showed that language identification is a very difficult problem and most collections of languages are non-identifiable (in fact, there is a tight characterization due to Angluin [Ang80] which we state in Definition 10). Hence, the algorithm of Kleinberg and Mullainathan [KM24] shows that language generation in the limit is much more tractable than identification. We note that while their algorithm operates in a non-statistical setting, it will be an important building block for our results.

Kleinberg and Mullainathan [KM24] observed that their algorithm eventually becomes a consistent generator but *suffers from mode collapse*: initially, it generates with breadth while being inconsistent with the target language; later on, as a larger part of the stream is seen, it starts sacrificing breadth in order to generate valid outputs. This behavior led them to leave the existence of a consistent generator that achieves breadth as an interesting open question. In this work, we will formally introduce a notion of breadth for language generation in our statistical setting (Section 1.1.1). For now, we mention that our definition roots in the notion of mode collapse from Generative Adversarial Networks (GANs) [AB17; GPM+20] and, roughly speaking, states that an algorithm (\mathcal{G}_n) generates with breadth from K if the probability that its support contains all the unseen examples from the target language goes to 1, as the training samples from a valid distribution go to infinity. Now it is a good point to contrast breadth with consistency: consistent generators aim at avoiding any elements outside of K while generators achieving breadth try to cover all unseen elements of K . The question of Kleinberg and Mullainathan [KM24] is asking whether the equilibrium condition that the support of the generator *exactly* matches the unseen elements of K can eventually be achieved by some algorithm. This is the main question we aim to address in this paper.

*Is it possible to achieve consistent language generation with breadth or
is there some inherent trade-off between consistency and breadth?*

1.1 Informal Results

Our main results confirm the tension between consistent generation and breadth for language models, conjectured by Kleinberg and Mullainathan [KM24], in a strong way: *informally*, we show that

A language model that generates with breadth must be inconsistent, i.e., it must hallucinate.

We focus on the probabilistic setting of Angluin [Ang88] which we have already introduced informally. En route to our results in the probabilistic setting, we also obtain results in the online setting of Gold [Gol67], Angluin [Ang79], and Kleinberg and Mullainathan [KM24], as we will

⁴Very briefly, a language collection $\mathcal{L} = \{L_1, L_2, \dots\}$ is called identifiable in the limit if there exists an algorithm $(\mathcal{A}_n: \mathcal{X}^n \rightarrow \mathbb{N})_n$ such that for any $K \in \mathcal{L}$ and any enumeration x_1, x_2, \dots of the strings of K appearing as a stream to (\mathcal{A}_n) , there is a finite time $n_0 \in \mathbb{N}$ after which the algorithm predicts the correct index of the true language, i.e., $L_{\mathcal{A}_n(x_1, \dots, x_n)} = K$ for any $n \geq n_0$.

see later. To facilitate a formal discussion of our contributions, we need to introduce some further definitions.

1.1.1 Setup and Definitions

A generating (or learning) algorithm is a sequence of computable mappings $(\mathcal{G}_n) = (\mathcal{G}_n)_{n \in \mathbb{N}}$ from samples $S \subseteq \mathcal{X}^n$ to generators, which are simply distributions over the domain \mathcal{X} . More formally, a generating algorithm is a sequence of mappings from samples to Turing machines that generate samples from an (explicitly or implicitly) defined distribution over strings.

In the statistical setting we consider, the learner observes samples from an unknown distribution which is valid for some unknown language K in the collection $\mathcal{L} = \{L_1, L_2, \dots\}$.

Definition 1 (Valid Distribution [Ang88]). *A distribution \mathcal{P} over a countable domain \mathcal{X} is valid with respect to a countable language collection \mathcal{L} if its support is the same as some language $K \in \mathcal{L}$. In this case, when we want to be specific about the language that \mathcal{P} draws samples from, we say \mathcal{P} is valid for K .*

If the collection \mathcal{L} is clear from context, we will simply say that \mathcal{P} is valid. Based on this definition and building on the model studied by Kleinberg and Mullainathan [KM24], we give the following adaptation for consistent generation from a collection \mathcal{L} in the statistical setting.

Definition 2 (Consistency). *A generating algorithm (\mathcal{G}_n) for a language collection \mathcal{L} is consistent if for any valid distribution \mathcal{P} , it holds that $\lim_{n \rightarrow \infty} \text{gen_er}(\mathcal{G}_n) = 0$. Otherwise, the algorithm is said to be inconsistent.*

Hence, an algorithm is said to be consistent if the generators it produces by training on any valid distribution \mathcal{P} converge to generating examples from the unseen part of \mathcal{P} . Some of our results explore when asymptotic consistency is achievable. However, the main focus of our work is on understanding the *rates* at which consistency (and other desirable properties) can be attained – if possible at all. In particular, we want to study the rate at which the generation error $\text{gen_er}(\mathcal{G}_n)$ decreases as the number of samples n goes to infinity – that is, we want to study the *learning curve* of consistent generation (and other tasks that we introduce later in this section). Bousquet, Hanneke, Moran, van Handel, and Yehudayoff [BHM+21] characterized learning curves for binary classification, formalizing the *universal rates* framework, earlier explored by Schuurmans [Sch97] and Antos and Lugosi [AL98]. To this end, we borrow their definition of universal rates.

Definition 3 (Informal, Universal Rates; [BHM+21], see Definition 12). *A generating algorithm (\mathcal{G}_n) has rate $R(\cdot)$, where $\lim_{n \rightarrow \infty} R(n) = 0$, for a language collection \mathcal{L} if*

$$\forall \mathcal{P} \in \text{Val}(\mathcal{L}) \quad \exists C, c > 0 \quad \text{such that} \quad \text{gen_er}(\mathcal{G}_n) \leq C \cdot R(c \cdot n) \quad \forall n \in \mathbb{N},$$

where $\text{Val}(\mathcal{L})$ is the class of valid (realizable) distributions for \mathcal{L} .

Observe that these learning curves are distribution-dependent since the constants c and C are allowed to depend on \mathcal{P} . This difference turns out to be crucial and can, sometimes, lead to significant differences between universal rates and the corresponding distribution-independent rates [BHM+21]. Among different universal rates, exponential universal rates are of specific interest as they are often the best possible rate, as we will see later. We say that the algorithm (\mathcal{G}_n) generates with an exponential universal rate if $R(n) = \exp(-n)$ in the above definition. Next, we turn to language generation with breadth.

Definition 4 (Breadth). *A generating algorithm (\mathcal{G}_n) for a language collection \mathcal{L} is said to achieve breadth if, for any valid distribution \mathcal{P} , it holds that $\lim_{n \rightarrow \infty} \Pr[\text{supp}(\mathcal{G}_n) \supseteq K \setminus S_n] = 1$, where S_n is the dataset used to train \mathcal{G}_n , i.i.d. from \mathcal{P} . Otherwise, the algorithm suffers from mode collapse.*

Definition 4 is inspired by the literature on GANs (see e.g., [AB17; GPM+20]). For instance, consider the work of Arjovsky and Bottou [AB17], which studies distributions \mathcal{G} and \mathcal{P} induced by the generator and nature, respectively, and says that mode collapse occurs when the KL divergence $\text{KL}(\mathcal{P} \parallel \mathcal{G}) := \int \log(\mathcal{P}(x)/\mathcal{G}(x)) \, d\mathcal{P}(x) \rightarrow \infty$. In particular, mode collapse happens when there is some string $x \in \text{supp}(\mathcal{P})$ for which $\mathcal{G}(x) = 0$. In other words, the generator has breadth when $\text{supp}(\mathcal{G}) \cup S_n \supseteq \text{supp}(\mathcal{P})$, which recovers our definition for breadth by noting that $\text{supp}(\mathcal{P}) = K$ since \mathcal{P} is valid for K and that, to be compatible with the definition of consistency (Definition 2), we bar a generator from repeating strings it has already seen. (It is worth mentioning that one can modify the definition of breadth to require $\text{supp}(\mathcal{G}_n) \supseteq K$ without changing any of our results; see Remark 2.) We also note that the definition of consistency we use can also be derived in an analogous fashion by requiring the reverse KL divergence (i.e., $\text{KL}(\mathcal{G} \parallel \mathcal{P})$) to be finite.

Putting the definitions of consistency and breadth together implies that an algorithm generates with consistency and breadth if, eventually, its support *matches* the set of unseen strings in K , i.e., $K \setminus S_n$ at the n -th step. After presenting our main results, in Section 1.3, we discuss relaxations of this notion of consistent generation with breadth.

A last ingredient for our results concerns the decidability of a folklore Theoretical Computer Science problem, which we call the *membership oracle problem*, that has motivated extensive work in formal languages and complexity theory [Soa99; Sip12]. A generator \mathcal{G} , which is the output of some generating algorithm, corresponds to some Turing machine, as is standard in the language inference literature, that samples according to a distribution over \mathcal{X} [BB75; Ang79; AS83; AB91].

Definition 5 (Membership Oracle Problem). *Given a generator \mathcal{G} , the membership oracle problem for \mathcal{G} , denoted as $\text{MOP}(\mathcal{G})$, is defined as follows: given the description of \mathcal{G} and a string x , output Yes if $x \in \text{supp}(\mathcal{G})$ and output No otherwise.*

This problem is, in general, undecidable due to a reduction to the halting problem (Section A); nevertheless, its decidability depends on the structure of the Turing machine as we will see shortly. The above definition naturally extends to generating algorithms.

Definition 6 (MOP for Generating Algorithms). *The membership oracle problem is decidable for a generating algorithm (\mathcal{G}_n) if, for any $n \in \mathbb{N}$ and any $S \subseteq \mathcal{X}^n$, $\text{MOP}(\cdot)$ is decidable for the induced generator $\mathcal{G} = \mathcal{G}_n(S)$.*

We note that the above definitions implicitly assume that the generator $\mathcal{G}_n(S)$ depends only on the randomness of S ; we could extend this by allowing $\mathcal{G}_n(S)$ to be a distribution over generators.

1.1.2 Main Results

We now have all the ingredients to state our first result, which establishes that, for all generating algorithms for which $\text{MOP}(\cdot)$ is decidable, (consistent) generation with breadth is as hard as language identification in the statistical setting.

As in [Definition 3](#), we will say that the generating algorithm (\mathcal{G}_n) generates with breadth from \mathcal{L} at some rate $R(\cdot)$ if, for any $K \in \mathcal{L}$, valid distribution \mathcal{P} , and $n \in \mathbb{N}$,

$$\mathbb{E}_{S \sim \mathcal{P}^n} \mathbb{1} \{ \text{supp}(\mathcal{G}_n) \neq K \setminus S \} \leq C \cdot R(c \cdot n),$$

for some distribution-dependent constants $C, c > 0$. If no rate $R(\cdot)$ satisfying $\lim_{n \rightarrow \infty} R(n) = 0$ exists, we will say that (\mathcal{G}_n) does not generate with breadth *at any rate*.

Informal Theorem 1 (see [Theorem 3.3](#)). *For every language collection \mathcal{L} that is not identifiable in the limit, no generating algorithm (\mathcal{G}_n) , for which $\text{MOP}(\cdot)$ is decidable, can generate from \mathcal{L} with breadth at any rate.*

Recall that the family of languages non-identifiable in the limit is quite broad. Based on the results of Gold [[Gol67](#)] and Angluin [[Ang79](#); [Ang80](#)] on the problem of language identification in the limit, our impossibility result holds for most interesting collections of languages. For [Informal Theorem 1](#) to be valuable and meaningful though, we further need to show that there exists an algorithm that generates without breadth for the collections of languages for which our impossibility result is true. Our next result states that this is indeed possible: there exists an algorithm that generates with (almost) exponential universal rates for *any* countable language collection \mathcal{L} .

Informal Theorem 2 (see [Theorem 3.3](#)). *For every language collection \mathcal{L} that is not identifiable in the limit, there exists a generating algorithm (\mathcal{G}_n) , for which $\text{MOP}(\cdot)$ is decidable, that generates (possibly) without breadth from \mathcal{L} at exponential rates. Further, if \mathcal{L} is identifiable in the limit, then there exists a generating algorithm (\mathcal{G}_n) , for which $\text{MOP}(\cdot)$ is decidable, that generates with breadth from \mathcal{L} at (almost) exponential rates.*

[Informal Theorem 2](#) shows that *any* countable collection of languages not only admits a consistent generator in the limit under an adversarial enumeration of the target language (as shown by Kleinberg and Mullainathan [[KM24](#)]), but the statistical rate at which consistency (as per [Definition 2](#)) is achieved is exponential in the number of samples. Further, for identifiable collections of languages, we give an algorithm that generates with breadth at an (almost) exponential rate.

The combination of [Informal Theorem 1](#) and [2](#), reveals a strong separation between generation with and without breadth for any generating algorithm for which $\text{MOP}(\cdot)$ is decidable. What is missing is an answer to: how large is the class of generators for which the membership oracle problem $\text{MOP}(\cdot)$ is decidable? It turns out there is a very broad class of language generators for which this is the case and which also captures modern LLMs, as we show next.

A Family of Generators for Which $\text{MOP}(\cdot)$ Is Decidable. Motivated by the structure of modern language models [[BJM83](#); [BCP+90](#); [BCE+23](#); [TLI+23](#); [OAA+24](#)], we consider a family of *iterative* generators. A generator is said to be iterative if it generates text one alphabet or “token” at a time (see [Definition 14](#)). To generate each token, the generator can perform an arbitrary (but finite) amount of computation and, possibly, use randomness. For this to make sense, one has to imagine strings of \mathcal{X} as strings over some finite alphabet Σ . This holds without loss of generality as \mathcal{X} is countably infinite and, hence, there is a one-to-one mapping from \mathcal{X} to strings over Σ (due

to which \mathcal{X} can be thought of as a set of strings over Σ).⁵ We show that for any iterative generator, the membership oracle problem is decidable and our [Informal Theorem 1](#) is applicable.

Informal Theorem 3 (see [Theorem 3.4](#)). *For any iterative generator G , $\text{MOP}(G)$ is decidable.*

Observe that this family of next-token generators is very general. First, it captures existing large language models: for instance, to simulate an LLM L , we define the next-token predictor as a Turing machine that simulates L on the provided string until L generates one new token. Next, it also captures systems where an LLM can interact with another Generative AI model or algorithmic system (such as a diffusion model or a code interpreter) – as these auxiliary systems can also be simulated by the generator. Given this, it becomes evident that this class of generators for which $\text{MOP}(\cdot)$ is decidable is fairly large and interesting.

Implications for the Gold-Angluin Model. We repeat that all the aforementioned results hold in the statistical setting. En route to obtaining our results in this setting ([Informal Theorems 1 and 2](#)), we show several connections to the online setting of Gold [[Gol67](#)], Angluin [[Ang79](#); [Ang80](#)], and Kleinberg and Mullainathan [[KM24](#)], which lead to the following result.

Informal Theorem 4 (see [Theorem 3.5](#)). *For any language collection \mathcal{L} that is not identifiable in the limit, no generating algorithm (G_n) , for which $\text{MOP}(\cdot)$ is decidable, can generate from \mathcal{L} with breadth in the limit.*

To be more concrete, a generating algorithm generates with breadth in the limit if its support is eventually $K \setminus S_n$, where S_n is the set of the first n positive examples (*i.e.*, examples that belong to K). We emphasize that [Informal Theorem 4](#) is in a similar spirit as our [Informal Theorem 1](#), but holds in the online model instead of the statistical model discussed earlier. In particular, [Informal Theorem 4](#) combined with the algorithm of Kleinberg and Mullainathan [[KM24](#)] give a separation between consistent generation with and without breadth in the Gold-Angluin model. Further, as explained before, this result applies to any iterative generator due to [Informal Theorem 3](#). Moreover, as $\text{MOP}(\cdot)$ is decidable for the generating algorithm of Kleinberg and Mullainathan [[KM24](#)] (since its support contains a singleton element x which can be computed by running their algorithm), the above result, in particular, shows that the algorithm of Kleinberg and Mullainathan [[KM24](#)] cannot generate with breadth in the limit from any non-identifiable collection.

Organization of Rest of the Introduction. We proceed with an exposition of our techniques in order to obtain our main results presented above. In [Section 1.3](#), we relax the definitions of consistency and breadth and give more “robust” trade-offs between hallucination and breadth. Next, in [Section 1.4](#), we give a list of open problems for future work. Finally, [Section 1.5](#) contains an extensive overview of related works.

1.2 Technical Overview

In this section, we present the technical tools we develop to obtain our main results.

⁵In a bit more detail, since \mathcal{X} and Σ^* are countably infinite, they have enumerations x_1, x_2, \dots and s_1, s_2, \dots . Therefore, given any string $s_i \in \Sigma^*$ generated by an iterative generator, one can map it to a string $x_i \in \mathcal{X}$, thereby getting a generator for \mathcal{X} .

A Natural Strategy to Prove Informal Theorem 1. At first glance, there seems to be a natural strategy to prove Informal Theorem 1: assume that there exists a consistent generating algorithm with breadth $\mathcal{G} = (\mathcal{G}_n)$ for some non-identifiable collection \mathcal{L} in the statistical setting and then show that this implies identification in the statistical setting, which would contradict the fact that \mathcal{L} is non-identifiable. To implement this strategy one needs a method to utilize \mathcal{G} , along with the positive samples from the target language K , for identification. This raises the question: what additional power can \mathcal{G} give that the positive samples do not *already* provide?

Initial Attempts to Implement the Strategy. Indeed, if one uses *no* additional properties of \mathcal{G} , then its outputs provide no more information than an adversarial enumeration of K . To develop some intuition, we begin by considering some properties of the generator and explaining why they are insufficient to enable identification.

1. *\mathcal{G} is non-adaptive.* First, one may want to utilize the fact that the generator \mathcal{G} is fixed and, hence, the samples it outputs cannot adapt to the specific algorithm being used based on the outputs of the algorithm. Hence, it will probably provide an algorithm-independent enumeration of the true language. However, this is not helpful in general since there exist simple non-identifiable language collections that remain non-identifiable for many enumerations of the target language.

2. *\mathcal{G} samples from a fixed distribution.* Another property one may want to leverage is the stochasticity of the generator: \mathcal{G} samples its outputs from a fixed distribution (which is valid for K). However, even this does not enable the identification of non-identifiable collections due to a result by Angluin [Ang88]. Angluin shows that even if the positive examples are i.i.d. from a valid distribution and do not appear as an adversarial enumeration (as in Gold [Gol67]), this does not enable identification of any collection \mathcal{L} that is non-identifiable in the limit. (We prove a stronger version of this result in Lemma 5.5.)

3. *\mathcal{G} samples from a simple distribution.* Moreover, the difficulty in the above negative result is not the complexity of the encoded distribution: it holds even when \mathcal{G} samples from a distribution that is computable by a Turing machine.

At this point, it is not clear how to utilize access to a generator \mathcal{G} which generates with breadth from K . Next, we present a strong form of access to the generator \mathcal{G} that is useful for identification.

4. *Access to Subset Queries “ $\text{supp}(\mathcal{G}) \subseteq L_i$ ” and “ $L_i \subseteq L_j$ ”.* For one of their algorithms, Kleinberg and Mullainathan [KM24] utilize a subset oracle that answers queries of the form “is $L_i \subseteq L_j$?”. (In general, this oracle is not guaranteed to be computable). One can imagine an extension of this oracle that, given an index i and description of the generator \mathcal{G} , outputs whether $\text{supp}(\mathcal{G}) \subseteq L_i$. The existence of this oracle turns out to be sufficient to identify K , as we explain next: After a finite amount of time, Kleinberg and Mullainathan [KM24]’s algorithm creates a list of “critical” languages C_1, C_2, \dots , of the following form (see Theorem 4.1 in Kleinberg and Mullainathan [KM24])

$$C_1 \supseteq C_2 \supseteq \dots \supseteq (C_i := K) \supseteq C_{i+1} \supseteq \dots$$

In words, this list has two properties (1) K appears in this list, say, at $C_i = K$ for some $i < \infty$ and (2) each language C_j in the list is a subset of the preceding language C_{j-1} . Given this list and the aforementioned subset oracle, one can easily identify the index of K as the largest j for which $\text{supp}(\mathcal{G}) = K \subseteq C_j$. This assumption allows to identify *any* collection in the limit given access to a

consistent generation \mathcal{G} with breadth. However, this type of access is not very practical since it is not clear when such an oracle is implementable.

Our Approach. Our first idea is that a much weaker form of access to \mathcal{G} – *membership oracle to $\text{supp}(\mathcal{G})$* – is sufficient for identification. This is where the membership oracle problem $\text{MOP}(\cdot)$ (Definition 5) appears in the proof. In fact, given this idea, it is not difficult to show that with that type of access, we can go from a generator with breadth in the online setting to an identification algorithm in the *online setting*; and, hence, get Informal Theorem 4. However, our focus is the statistical setting where there are several additional challenges in using the membership oracle to $\text{supp}(\mathcal{G})$.

A. Need for Universal Rates for Generation and Identification. The key issue is the following: In the statistical setting, if we assume that we have a generator with breadth at rate $R(\cdot)$, then we can hope to show an implication that we can get an identification algorithm at rate $R(\cdot)$. However, this need not imply a contradiction to the identifiability of \mathcal{L} as in the online setting. This is because, even though \mathcal{L} is non-identifiable in the online setting, it may become identifiable at *some* rate $R'(\cdot)$ in the statistical setting. Indeed, this is the case in binary classification, where there are simple hypothesis classes (such as thresholds over reals) that are not learnable in Littlestone’s online setting [Lit88] but become learnable (at a universal – and uniform – linear rate) in the statistical setting; in fact, *any* hypothesis class is learnable in the statistical setting under universal rates, since there is a Bayes consistent algorithm, under benign assumptions [BHM+21].

Hence, to get a contradiction, we first need to understand the taxonomy of universal rates for generation and identification. We remark here that both the learning task (e.g., classification, regression, identification, and generation) and loss function used in the problem are pivotal for the landscape of rates that one gets; for instance, with the zero-one loss for binary classification one gets a trichotomy of rates [BHM+21], but with the L_1 -loss for regression, one gets infinitely many rates [AHK+24].

To overcome the above challenge, we provide statistical rates for identification and generation. We start with identification. We show that if \mathcal{L} is identifiable in the limit in the adversarial Gold-Angluin setting with positive examples [Gol67; Ang80], then it is identifiable under Definition 12 with (almost) exponential (universal) rates. This is the less technical part of the proof so we will give a high-level approach.

B. Identification in the Limit \implies Identification at (Almost) Exponential Rates. Our idea is reminiscent of Bousquet, Hanneke, Moran, van Handel, and Yehudayoff [BHM+21] and requires splitting the input dataset into multiple batches whose size is carefully selected, running the online algorithm on each batch, and then taking a majority vote over the outputs of the algorithm. We remark that there are some technical issues which require further care, compared to Bousquet et al. [BHM+21]. First, unlike the setting of Bousquet et al. [BHM+21], we only see positive examples and we get no feedback about our guesses. Thus, we cannot use their approach to “estimate” a time after which the learner will stop making mistakes. Moreover, when we run the learners on multiple batches, it can be the case that different batches output different indices of languages that correspond to K (since the target language can appear at multiple positions in the countable collection \mathcal{L}).

Thus, taking a majority vote over these indices might not work. Nevertheless, we manage to handle these issues and get almost exponential rates for collections that satisfy Angluin’s criterion

for identification in the limit [Ang79]. A bit more concretely, to circumvent the first issue, our approach is to “guess” the right batch size, and this guess needs to be an increasing function of n – this is why we get almost exponential rates instead of exactly exponential rates (Lemma 5.5). The second issue is more subtle. At a high level, we use a voting scheme where the output of every batch \hat{L}_i gives a “vote” to every language $L \in \mathcal{L}$ such that $\hat{L}_i = L$, and we predict the lowest-indexed language that is voted by at least half of the batches. In its current form, this scheme is not computable, nevertheless, we show that it can be modified so that it becomes computable (Lemma 5.4).

The more interesting half of establishing universal rates for identification is the lower bound showing that if a collection is not identifiable in the limit, it is also not identifiable in the statistical setting at any rate $R(\cdot)$ such that $\lim_{n \rightarrow \infty} R(n) = 0$.

C. Impossible to Identify in the Limit \implies Impossible to Identify at Any Rate. Recall that the statistical setting was studied by Angluin [Ang88]. Angluin [Ang88] showed that every learner, with probability at least $1/3$, does not converge to outputting a (stable) index of the target language in an infinite stream of examples drawn from a valid distribution. In other words, with probability at least $1/3$, the algorithm will either stabilize to an index that does not correspond to the target language or it will not stabilize to any index. Notice that this does not rule out algorithms that output different indices of the target language, for all but finitely many $n \in \mathbb{N}$. The first step towards establishing our desired lower bound is to strengthen Angluin’s result: we show that any learning algorithm, with probability at least $1/3$, outputs indices that do not correspond to the target language infinitely often. More formally, let us consider an identification algorithm h_n , which maps a training set of n examples x_1, \dots, x_n to an index $h_n(x_1, \dots, x_n)$ so that $L_{h_n(x_1, \dots, x_n)}$ is the n -th prediction for the true language. The aforementioned lower bound means that⁶

$$\Pr_{\{x_i: i \in \mathbb{N}\} \sim \mathcal{P}^\infty} \left[\limsup_{n \rightarrow \infty} \left\{ L_{h_n(x_1, \dots, x_n)} \neq K \right\} \right] \geq \frac{1}{3}, \quad (2)$$

where $X \sim \mathcal{P}^\infty$ corresponds to an infinite i.i.d. draw from \mathcal{P} . One may be tempted to conclude that this implies that with probability $1/3$ we cannot identify the target language (in the statistical setting). However, the quantity we wish to bound away from 0 to derive the desired lower bound is

$$\limsup_{n \rightarrow \infty} \Pr_{x_1, \dots, x_n \sim \mathcal{P}^n} \left[\left\{ L_{h_n(x_1, \dots, x_n)} \neq K \right\} \right].$$

It is well-known that for any sequence of events $\{\mathcal{E}_n\}_{n \in \mathbb{N}}$,

$$\Pr \left[\limsup_{n \rightarrow \infty} \mathcal{E}_n \right] \geq \limsup_{n \rightarrow \infty} \Pr[\mathcal{E}_n]. \quad (3)$$

This, however, is not sufficient to deduce the result we need; we need the opposite inequality. Hence, Angluin’s guarantee does not suffice to get our lower bound. In order to show our result, we use a boosting argument (Lemma 5.8): if there exists a learner h_n whose probability of

⁶Informally, \limsup of a sequence of events captures the events that occur infinitely often. For instance, $\Pr[\limsup_{n \rightarrow \infty} \mathcal{E}_n]$ represents the probability that infinitely many of the events $\mathcal{E}_1, \mathcal{E}_2, \dots$ occur. On the other hand, $\limsup_{n \rightarrow \infty} \Pr[\mathcal{E}_n]$, roughly speaking, denotes the largest value that the probabilities $\Pr[\mathcal{E}_1], \Pr[\mathcal{E}_2], \dots$ approach infinitely often as $n \rightarrow \infty$.

misidentification

$$\Pr_{x_1, \dots, x_n \sim \mathcal{P}^n} [L_{h_n}(x_1, \dots, x_n) \neq K]$$

converges to a number strictly less than $1/2$, then we can convert it to a learner whose error rate decreases (almost) exponentially quickly. This (almost) exponential rate, in particular, implies that

$$\sum_{n=1}^{\infty} \Pr_{x_1, \dots, x_n \sim \mathcal{P}^n} [L_{h_n}(x_1, \dots, x_n) \neq K] < \infty.$$

This, crucially, enables us to use the Borel-Cantelli lemma (see [Lemma E.1](#)) which gives us that $\Pr \left[\limsup_{n \rightarrow \infty} \{L_{h_n}(x_1, \dots, x_n) \neq K\} \right] = 0$, and, thus, a contradiction to [Equation \(2\)](#). This implies the desired impossibility result.

As consequence of the above results, we get a dichotomy for universal identification rates:

Informal Theorem 5 (see [Theorem 3.1](#)). *For any language collection \mathcal{L} that is identifiable in the limit and for any $g(n) = o(n)$, there exists a learner that identifies \mathcal{L} at rate $\exp(-g(n))$. Otherwise, \mathcal{L} is not identifiable at any rate.*

We remark that if we have access to subset queries for \mathcal{L} , we can show that there exists an algorithm that achieves exactly exponential rates, for all identifiable collections (see [Proposition 3.8](#)).

Next, we move to understanding universal rates for language generation.

D. Universal Rates for Generation (Possibly Lacking Breadth) Without Boosting. One might suspect that a similar batching argument would give us exponential rates for generation: just run the on-line algorithm of Kleinberg and Mullainathan [[KM24](#)] multiple times and aggregate. The issue is that aggregation for generation is different than prediction: for prediction, it is clear how to implement majority vote as a boosting technique; for generation, it is unclear how to aggregate different generated strings which is, typically, necessary to obtain a boosting algorithm. One immediate attempt is to take majority votes over the strings that each batch outputs; unfortunately, even if the majority of them are generating from the target language, they might be outputting different strings, thus, even a few batches outputting the same invalid strings are enough to fool our aggregation rule.

Another tempting approach is to mimic the strategy we used to aggregate different indices of the target language in the identification setting: we go over every output of the batches and we let them give a vote to each of the languages in \mathcal{L} they belong to.⁷ It is not hard to see that every batch whose output corresponds to a valid generator will vote for the target language. Unfortunately, it will also vote for all *supersets* of the target language. This is exactly the heart of the difficulty of identification: telling apart supersets of the target language from the target language, which is colloquially called overgeneralization. Taking it to the extreme, imagine that the first language of the collection contains all the strings, *i.e.*, $L_1 = \mathcal{X}$. Then, all the batches will vote for L_1 . This is problematic for two reasons: generating a fresh string from the majority-voted language is as good as random guessing, and choosing a string among the ones that voted for the majority-voted language is as good as picking one of the outputs of all batches uniformly at random.

⁷The astute reader might realize that, as stated, this strategy is not computable – as we explain, even if one could implement it, this aggregation scheme does not work.

Perhaps surprisingly, it turns out that a much simpler approach works: we show that the algorithm of Kleinberg and Mullainathan [KM24] directly enjoys exponential rates in the statistical setting, without the use of batching and boosting. This observation is based on a sufficient condition that allows one to use an algorithm that works “in the limit” to obtain exponential rates in the statistical setting, without any modification (see [Lemma 5.11](#)).

Informal Theorem 6 (see [Theorem 3.2](#)). *For any countable language collection \mathcal{L} there exists a generating algorithm that generates from \mathcal{L} at an (optimal) exponential rate.*

This pair of results for identification ([Informal Theorem 5](#)) and generation ([Informal Theorem 6](#)) allow us to get [Informal Theorem 1](#) and [2](#). The idea for [Informal Theorem 1](#) is that we will use the algorithm \mathcal{G} that generates with breadth at some rate $R(\cdot)$ for an arbitrary non-identifiable collection \mathcal{L} and membership oracle access to \mathcal{G} in order to get an identification algorithm for \mathcal{L} with some rate $R'(\cdot)$ such that $\lim_{n \rightarrow \infty} R'(n) = 0$. This is a contradiction since [Informal Theorem 5](#) shows that \mathcal{L} admits no rate in the universal setting. Finally, [Informal Theorem 2](#) follows almost immediately from our universal rates result for generation.

1.3 Additional Results With Relaxation of Consistency and Breadth

Next, we study a relaxation of consistent generation with breadth, which we call unambiguous generation, and ask: *is there a generator that unambiguously generates from a non-identifiable collection?*

In this section, we will allow the generator to repeat examples in the training data. Like all of our results with breadth, this choice is not crucial, and all of the results have analogs where the generator does not repeat training examples (see [Remark 2](#)). We make this choice for simplicity. We show that unambiguous generation (which we define later in this section) from non-identifiable collections is impossible for any generator \mathcal{G} for which $\text{MOP}(\mathcal{G})$ is decidable and that satisfies the natural property that \mathcal{G} “stabilizes” after seeing sufficiently many examples:

Definition 7 (Stability). *A generating algorithm (\mathcal{G}_n) is stable for a language collection \mathcal{L} if for any target language $K \in \mathcal{L}$ and for any enumeration of K , there is some finite $n^* \in \mathbb{N}$ such that for all $n, n' \geq n^*$, it holds that $\text{supp}(\mathcal{G}_n) = \text{supp}(\mathcal{G}_{n'})$.*

We make some initial remarks about stable generators. First, any generator \mathcal{G} that is consistent and achieves breadth is also stable, since after some finite time its support, union the training set, becomes K and remains so. (Here, whether \mathcal{G} repeats training examples or not is not crucial – the two types of generators are interchangeable; see [Remark 2](#).) Second, this notion of stability can be seen as trying to capture practical heuristics such as learning rate schedules and early stopping that reduce the amount of changes to the generator as more and more samples are seen. Moreover, the original work of Gold [Gol67] also requires the identifier to stabilize to a consistent guess, and, more recently, the stability property of learning algorithms was explored in the PEC learning setting of Malliaris and Moran [MM23].

Having defined stability, we proceed to discuss relaxations of generation with breadth. Intuitively, consistent generation with breadth requires the generator to eventually stop making mistakes – where a mistake is any element x that \mathcal{G} incorrectly includes (if $x \notin K$ or x is part of the training samples) or excludes (if $x \in K$) from its support. We now relax this and only require that, eventually, the generator \mathcal{G} makes *finitely* many mistakes. Observe that this is a non-trivial

requirement because the languages contain infinitely many strings and, so, at the start, \mathcal{G} is expected to make infinitely many mistakes. A valuable observation is that it is possible for two languages L_1 and L_2 to only differ in finitely many strings even if each contains infinitely many strings. With this observation, it is not too hard to see that the aforementioned requirement is too weak to capture a reasonable notion of generation from the target language K . Indeed, it would allow generators that, given examples from K , perpetually generate outputs (with breadth) from a language L that is not the actual target language – which is a severe form of hallucination.

Hence, to create a meaningful model, we must impose some further restrictions on the mistakes of the generator \mathcal{G} . The above example motivates that, at the least, the generator \mathcal{G} should be “closer” to generating from K than some language $L \neq K$ with $L \in \mathcal{L}$. We call such a generator *unambiguous*.

Definition 8 (Unambiguous Generator). *A generating algorithm $\mathcal{G} = (\mathcal{G}_n)$ is unambiguous for a language collection \mathcal{L} if, for any $K \in \mathcal{L}$ and every enumeration of K , its support eventually becomes closer to K than to any other language $L \neq K$ in \mathcal{L} in terms of the symmetric difference metric, i.e., there exists some $n^* \in \mathbb{N}$ such that for all $n \geq n^*$ it holds that*

$$|\text{supp}(\mathcal{G}_n) \triangle K| < \min_{L \in \mathcal{L}: L \neq K} |\text{supp}(\mathcal{G}_n) \triangle L|,$$

where recall that for two sets S and T , $S \triangle T := (S \setminus T) \cup (T \setminus S)$.

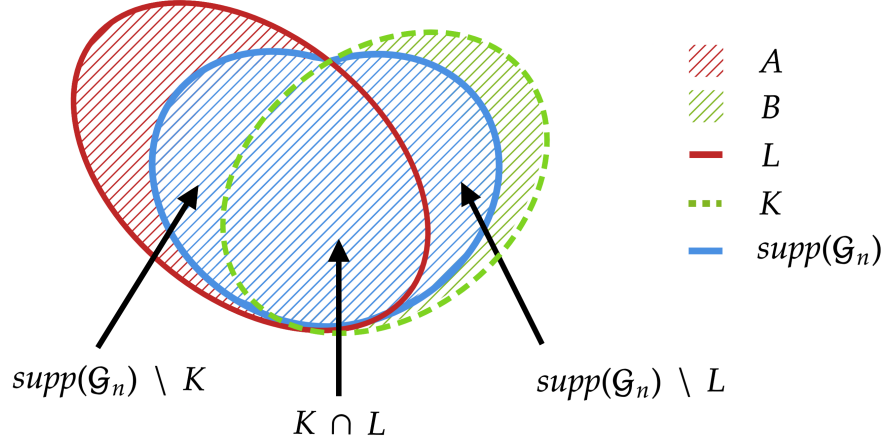


Figure 1: *An Unambiguous Generator That neither Has Consistency nor Breadth.* In this example, the language collection \mathcal{L} has two languages L and K , where K denotes the target language. The **red curve** denotes L , the **dashed green curve** denotes K , and the **blue curve** denotes the support of $\text{supp}(\mathcal{G}_n)$. The generator \mathcal{G}_n hallucinates since $\text{supp}(\mathcal{G}_n) \setminus K \neq \emptyset$ and does not achieve breadth for the target K since $B = K \setminus \text{supp}(\mathcal{G}_n)$ is non-empty. Nevertheless, this generator is unambiguous as $|\text{supp}(\mathcal{G}_n) \setminus K| + |B| < |\text{supp}(\mathcal{G}_n) \setminus L| + |A|$.

Here, we pause to observe that this notion of generation is a significant relaxation of generation with breadth that we considered earlier (Definition 4). Not only does it allow the generator to hallucinate certain strings not in the target K and omit strings actually in K for arbitrarily long, the number of hallucinations and omissions can be arbitrarily large, depending on the structure

of the language collection \mathcal{L} . Surprisingly, we show that even this very weak notion of “consistent generation with breadth” is not achievable by a large class of generators.

Informal Theorem 7 (see [Theorem 3.6](#)). *For every language collection \mathcal{L} that is not identifiable in the limit, no stable generating algorithm (\mathcal{G}_n) for which $\text{MOP}(\cdot)$ is decidable, can generate unambiguously from \mathcal{L} at any rate.*

Thus, under mild conditions, no stable algorithm can generate unambiguously from a non-identifiable collection. Moreover, we also prove an analog of [Informal Theorem 7](#) in the online setting (see [Theorem 3.7](#)), which extends our earlier result for generation with breadth in the online setting ([Informal Theorem 4](#)). This raises several questions regarding unambiguous generation, which we leave as interesting open problems (see [Section 1.4](#)). Note that while this impossibility result has a benign requirement that the generator is stable, it already considerably extends our main result [Informal Theorem 1](#), since any generator that achieves breadth must be stable – otherwise, its support cannot settle on the target language K . (Note that while [Informal Theorem 1](#) requires the generator to not repeat training examples, any generator that repeats training examples can be converted into one that does not repeat training examples and vice-versa; see [Remark 2](#).)

1.4 Takeaways, Discussion, and Open Problems

We believe that a key takeaway of our results is that the question of Kleinberg and Mullainathan [\[KM24\]](#) seems to open an avenue towards a formal modern theory of language generation bridging learning theory and traditional TCS fields, like complexity theory and formal languages. As we explain in the subsequent technical overview, our tools contribute to this direction by connecting classical lines of work on identification of formal languages tracing back to Gold [\[Gol67\]](#), Angluin [\[Ang79; Ang80; Ang88\]](#) and computability theory [\[Soa99; Sip12\]](#), to modern learning paradigms such as learning curves [\[BHM+21\]](#) and language generation [\[KV24; KM24\]](#).

Next, we emphasize that our impossibility result ([Theorem 3.3](#)) is *not* a dead end for language generation. Instead, it illustrates the need for additional human feedback during the post-training process – which provides additional information over positive samples alone – to achieve effective language models. Indeed, if both positive and negative examples are available, then generation with breadth is achievable for all countable collections of languages.⁸ In other words, our results can be seen as further theoretical evidence of the benefits of post-training with human feedback, highlighting its importance in developing language models that achieve both consistency and breadth, and adding to prior theoretical results from Kalai and Vempala [\[KV24\]](#).

Further, we underline that even though we focus on a prompt-less generation setting [\[KV24; KM24\]](#), most of our results immediately extend to a prompted setting using the approach of Kleinberg and Mullainathan [\[KM24\]](#).

Remarks and Open Questions. We now state a few remarks regarding our results and pose some interesting open questions. First, as a byproduct of our results, we establish almost tight rates for

⁸This follows from the work of Gold [\[Gol67\]](#), which showed that any countable collection of languages can be identified with such feedback. Using appropriate batching and boosting, we show that this identification algorithm (which works in the limit) can be converted to a generation algorithm with breadth that achieves an exponential rate. Concretely, [Theorem 3.11](#) shows how to identify at an exponential rate and [Proposition 6.5](#) shows how to convert this to a generation algorithm.

identification and generation with positive examples (see [Section 3.1](#) and [Section 3.4](#) for formal statements and discussion). Obtaining tight rates for these tasks is an interesting problem.

Next, our impossibility results capture a large class of language-generating algorithms but do not completely forbid consistent generation with breadth. An immediate open question is how much further we can extend the class of generating algorithms for which the impossibility result in [Informal Theorem 1](#) holds.

Open Question 1. Is there a class of generative algorithms for which the induced generators can be modeled as Turing machines and which achieve breadth and consistency for all countable collections of languages?

Further, we also proved a more robust version of our main result ([Informal Theorem 1](#)), namely, [Informal Theorem 7](#), which showed that no algorithm from a large class of generators can generate while making a “small” number of hallucinations or omissions (also see [Section 3.3](#) for another robust version of [Informal Theorem 1](#)). It is interesting to understand if one can prove a more robust version of [Informal Theorem 1](#). To this end, we propose the following problem.

Open Question 2. What is the Pareto frontier of an approximate notion of breadth and consistency? In other words, if we fix a collection of languages and allow the generator to hallucinate at some given rate, what is the minimal fraction of the mass from the target language that this generator has to miss?

Next, to the best of our knowledge, it is not possible to test if a language collection is identifiable in the limit (without access to a strong oracle); this, for instance, becomes evident by inspecting Angluin’s criterion for identifiable collections (see [Definition 10](#)). Hence, we would like to know the following:

Open Question 3. Is there a best-of-both-worlds algorithm between consistent generation and generation with breadth, *i.e.*, is there an algorithm that will always generate in the limit from the target language consistently but, whenever identification is possible, it will also achieve breadth?

We make some initial progress on this question by showing that the algorithm proposed by Kleinberg and Mullainathan [[KM24](#)] already achieves this best-of-both worlds guarantee, provided it has access to a *subset oracle* for \mathcal{L} that answers queries of the form “is $L_i \subseteq L_j$?” (see [Section B.2](#)).

Finally, our algorithm that achieves (almost) exponential rates for identification uses an algorithm for identification in the limit as a black box. However, our algorithm that achieves exponential rates for generation makes use of certain specific properties of the algorithm of Kleinberg and Mullainathan [[KM24](#)]. Thus, we ask the following question.

Open Question 4. Is there a black-box transformation from an algorithm that generates in the limit in the online setting to an algorithm that generates with *exactly* exponential rates in the statistical setting?

1.5 Further Related Works

Our setting is based on the statistical formulation of Angluin [Ang88], who studied identification from stochastic examples in the limit. However, Angluin [Ang88] does not provide any learning rates which is one of the main aspects of our work. In terms of techniques, our inspiration for the statistical rates comes from *universal learning*, initiated by Bousquet, Hanneke, Moran, van Handel, and Yehudayoff [BHM+21] and studied in Bousquet et al. [BHM+21], Hanneke et al. [HKM+22], Kalavasis, Velegkas, and Karbasi [KVK22], Bousquet et al. [BHM+23], Hanneke, Moran, and Zhang [HMZ23], and Attias et al. [AHK+24]. However, as we have already explained there are various differences between our setting and our techniques (we provide a more extensive and self-contained discussion in Section D).

Our work connects various disjoint strands of research and we discuss each one of them below.

Theory on Hallucinations. In terms of rigorous evidence about hallucinations in LLMs, we have already mentioned the work of Kalai and Vempala [KV24] at the start of Section 1. The result of Kalai and Vempala [KV24] is that calibrated⁹ language models must hallucinate. The fascinating implication of this result is that one can lower bound the rate of hallucination, *i.e.*, the quantity $\mathbb{E}_{S \sim \mathcal{P}^n, x \sim \mathcal{G}_n} \mathbb{1}\{x \notin K\}$, by the extent of a model’s calibration. Their intuition is that the root of hallucinations are *rare* patterns in the training data. Informally, their main result (under assumptions on K and \mathcal{P}) is that for any trained model \mathcal{G}_n with n samples, the hallucination rate $\mathbb{E}_{S \sim \mathcal{P}^n, x \sim \mathcal{G}_n} \mathbb{1}\{x \notin K\} \geq \hat{R} - \text{Mis}_{\mathcal{P}}(\mathcal{G}_n) - 1/\sqrt{n}$, where \hat{R} is the fraction of facts that only appear *once* in the training data and $\text{Mis}_{\mathcal{P}}(\mathcal{G}_n)$ is the amount of miscalibration of the model. Hence, if the model is calibrated, *i.e.*, $\text{Mis}_{\mathcal{P}}(\mathcal{G}_n) \approx 0$, the hallucination rate is lower bounded by the rare facts’ rate. Compared to our work, their goal is to show a quantitative lower bound, which is obtained under assumptions on the training distribution \mathcal{P} and the fact that the model is calibrated. Our goal is different: we want to understand whether a model can achieve breadth while avoiding hallucinations building on the recent work of Kleinberg and Mullainathan [KM24]. We also refer the reader to Kalai and Vempala [KV24] for an extensive overview of applied works on hallucinations.

Peng, Narayanan, and Papadimitriou [PNP24] use communication complexity to prove that the transformer layer is incapable of composing functions if the domains of the functions are large enough. This work could also be seen as rigorous evidence about the hallucinations of LLMs since function composition is a fundamental task for reasoning [GLL+24].

The work of Xu, Jain, and Kankanhalli [XJK24] is also studying hallucinations of LLMs. They define hallucination as a failure to identify the target function which belongs to an *uncountable* collection of functions. This is *significantly* stronger than the definition we and prior works [KV24; KM24] have considered (making their impossibility results *significantly* easier to prove). Their main result is that all LLMs must hallucinate. This is easy to see: consider an LLM learning to predict the next element in a sequence of 0s and 1s, after observing only a finite prefix of the enumeration, it has no way of knowing the next element in the order (since they allow both continuations) and, hence, the target sequence cannot be identified.

Finally, the work of Aithal et al. [AML+24], which is mainly empirical, aims to explain hallucinations on the other important family of generative models, namely diffusion-based models, via

⁹The exact definition of calibration is not important for this work: a language model is calibrated if, roughly speaking, the strings that the model assigns probability mass p , appear in a p fraction of the true distribution [Daw82].

mode interpolation which, in theory, relies on difficulties in approximating non-smooth parts of the score function.

Language Learning. In our results, we make no implicit assumption about the architecture of our models; this is in accordance with the works of Solomonoff [Sol64], Gold [Gol67], Angluin [Ang82], Angluin and Smith [AS83], Angluin [Ang88], Pitt [Pit89], and Kleinberg and Mullainathan [KM24]. However, there are various works aiming at understanding language learning capabilities of specific architectures, *e.g.*, [Elm90; GS01; Mer19; BAG20; EGZ20; Hah20; HHG+20; YPP+21; MS23]. For instance, Liu et al. [LAG+23] show that low-depth transformers can represent the computations of any finite-state automaton, while Sanford, Hsu, and Telgarsky [SHT23] identify a particular mathematical problem that cannot be computed by single-layer multi-head transformers. The aforementioned works share some similarities with us in the sense that they focus on whether models can be trained to generate or recognize strings in a *fixed* formal language. Akyürek et al. [AWK+24] study in-context language learning: the language model is prompted with a finite collection of strings from an unknown regular language (which changes across different tasks), and must infer the distribution over strings corresponding to the full language. In a similar spirit, Edelman et al. [EEG+24] study in-context learning of Markov chains. Other related works are those of Xie et al. [XRL+22] and Hahn and Goyal [HG23] that study conditions under which in-context learning can arise for language learning.

Allen-Zhu and Li [AL24] design context-free grammars and empirically study the consistent generation (accuracy) and breadth (diversity) of GPT models on these synthetic examples. In comparison to this work, we provide a theoretical treatment of the trade-off between consistency and breadth under a very abstract model, studied by Gold [Gol67], Angluin [Ang79; Ang88], and Kleinberg and Mullainathan [KM24]. Our results indicate that, even in a very idealized framework, achieving (perfect) consistency and breadth is impossible. We view the empirical findings of Allen-Zhu and Li [AL24] as an exciting indication that, in the real world (or more concretely in controlled experiments on “small” models and synthetic datasets), a balance between (imperfect) consistency and breadth is possible and modern LLMs can achieve it. Further understanding how much consistency and breadth one can achieve at the same time theoretically is an exciting direction.

Finally, in a concurrent and independent work, Li, Raman, and Tewari [LRT24] also study language generation, interpreting it in a learning-theoretic setting reminiscent of the PAC framework and the online learning setting of Littlestone [Lit88]. They propose “non-uniform generatability” – which relaxes “uniform generatability” [KM24] – and characterize the collections for which uniform and non-uniform generatability are achievable in the Gold-Angluin model; in particular, unlike Kleinberg and Mullainathan [KM24] they also allow the collection \mathcal{L} to contain uncountably many languages. These dimensions are analogs to the Littlestone dimension (and its extension to the non-uniform setting [Lu23]), which only holds for finite collections of languages. Moreover, they show the proposed dimension is incomparable to the VC dimension. Finally, they give analogous characterizations in the “prompted generation” setting, extending some of the results of Kleinberg and Mullainathan [KM24]. Our work is orthogonal to theirs: first, we study trade-offs between generating with and without breadth – both in a statistical setting and the Gold-Angluin model – and, second, we study the “learning curves” for generation and identification in the framework of Bousquet et al. [BHM+21].

Probably Eventually Correct Learning. As we mentioned Gold’s model is a predecessor of the famous PAC model of Vapnik [Vap13] and Valiant [Val84]. A natural question is whether there is a conceptual meeting point for the two works. Is there a notion of “PAC learning in the limit?” The answer to this question is affirmative and comes from the field of *algorithmic stability* (see e.g., [ABL+22; BGH+23; CMY23; KKM+23; MSS23] and the references therein), studied in the context of binary classification [MM23].

Malliaris and Moran [MM23] introduce the Probably Eventually Correct (PEC) model of learning. Here we fix a collection $\mathcal{L} = \{L_1, L_2, \dots\}$ of languages and a distribution \mathcal{P} over positive and negative labeled examples (in contrast to the standard identification setting of Gold). PEC learning focuses on distributions \mathcal{P} realizable by the collection \mathcal{L} in the sense of Bousquet et al. [BHM+21] (see Section D). An algorithm is said to PEC learn \mathcal{L} if for any realizable distribution \mathcal{P} , with probability 1 over i.i.d. samples $\{(x_i, y_i) : i \in \mathbb{N}\}$ drawn from \mathcal{P} , there exists time $t^* \in \mathbb{N}$ such that for all $t \geq t^*$, given $\{(x_i, y_i) : 1 \leq i \leq t\}$, the algorithm outputs an $L_t \in \mathcal{L}$ such that

$$\Pr_{(x,y) \sim \mathcal{P}} [L_t(x) \neq y] = 0.$$

Malliaris and Moran give a combinatorial characterization of the collections of languages that are PEC learnable: a collection of languages \mathcal{L} is PEC learnable if and only if it does not shatter an infinite Littlestone tree. We stress that, when the learner has access to positive and negative examples, the absence of an infinite Littlestone tree does not characterize identification in our setting. This is in stark contrast with binary classification. In particular, in Section D, we show that there exists a set of languages that have an infinite Littlestone tree, hence not learnable in the online setting of Bousquet et al. [BHM+21], but it allows for identification in the limit with positive and negative examples. In fact, the collection we use in Example 3 is identifiable in the limit even with just positive examples. This already sets the stage for a starkly different landscape of optimal learning rates between the setting of Bousquet et al. [BHM+21] and Angluin [Ang88], as we will see in Section 3.1.

As we said before, the online model of Gold [Gol67] and the classical online setting of Littlestone [Lit88] have various differences. Lu [Lu23] studies non-uniform online learning in order to bridge the gaps between the inductive inference model of Gold [Gol67] and classical online learning. In this setting, the adversary is oblivious and fixes the true language K in advance (as in Gold’s model). At each round, an example from K is revealed, the learner makes a prediction but then she observes feedback. The model is non-uniform in the sense that the mistake bound depends on K .

Learning from Positive Examples. Learning from positive examples occurs very frequently in real-world applications and has been extensively studied. A lot of work has been done on learning from positive examples in Gold’s model of learning in the limit [Gol67; Ang80; Ber86; Ang88; Shi90; ZL95]. Apart from that, an extension of Valiant’s PAC model has been also studied [Nat87; Den98]. Natarajan [Nat87] considered the setting where the learner only has access to positive examples and showed that even very simple classes such as halfspaces in two dimensions are not learnable from positive examples alone. Denis [Den98] relaxed this requirement: they study a setting where the learner has access to both positively labeled examples but also to unlabeled examples [DGL05]. At the heart of virtually all of the results in this line of work is the use of unlabeled samples in order to generate negative examples. When the original distribution is uniform,

better algorithms are known: De, Diakonikolas, and Servedio [DDS15] gave efficient learning algorithms for DNFs and LTFs, Frieze, Jerrum, and Kannan [FJK96] and Anderson, Goyal, and Rademacher [AGR13] gave efficient learning algorithms for learning d -dimensional simplices. On the other side, Rademacher and Goyal [RG09] and Eldan [Eld11] give lower bounds for learning with positive examples.

Recently, interest in learning from positive examples has sparked from work on truncated statistics (e.g., [DGT+18; DGT+19; KTZ19; FKT20; DKT+21; Ple21; DNS23; DLN+24; DKP+24; LMZ24]). Kontonis, Tzamos, and Zampetakis [KTZ19] show how to learn concept classes of bounded Gaussian surface area from positive Gaussian examples and Lee, Mehrotra, and Zampetakis [LMZ24] generalize this to show how to learn concept classes approximable by polynomials in the L_2 -norm from positive examples. However, all these works focus on computationally efficient learning/testing while we focus on statistical consistency of identification and generation without any restrictions on computation time.

2 Model and Preliminaries

In this section, we introduce notation and preliminaries that are useful in subsequent sections.

Countable Domains and Enumerations. We always assume that languages are subsets of some fixed infinite and *countable* domain \mathcal{X} . Since \mathcal{X} is infinite and countable, after a suitable bijective mapping, one can think of \mathcal{X} as \mathbb{N} . In some cases, one may also like to think of \mathcal{X} as the set of (arbitrarily long) strings over a finite alphabet Σ , i.e., Σ^* . This is again without loss of generality since \mathbb{N} is bijective to $\{0,1\}^*$ (e.g., using the standard binary encoding). Depending upon the context, we use one interpretation ($\mathcal{X} = \mathbb{N}$) or the other ($\mathcal{X} = \Sigma^*$), whichever is more intuitive. The notion of *enumeration* is important in our work; fix a set $L \subseteq \mathcal{X}$. We refer to L as a language. An enumeration of L is a complete and ordered listing of all the elements in L that allows for, potentially, repetitions of elements. In particular, an enumeration x_1, x_2, \dots of L has the property that for any element $w \in L$ there is a finite index i such that $x_i = w$. For example, $1, 2, 3, \dots$ is a valid enumeration of \mathbb{N} but $2, 4, 6, \dots$ is not (since, in the latter sequence, odd numbers do not appear at any finite position).

Additional Notation. We use \mathcal{A} , \mathcal{I} , and \mathcal{G} to denote algorithms, and often reserve \mathcal{G} for a generator, i.e., an algorithm that given examples $x_1, \dots, x_n \in \mathcal{X}$, outputs a new example from \mathcal{X} . We use \mathcal{P} and \mathcal{D} to denote distributions over the elements of some language $L \subseteq \mathcal{X}$. We use standard notation related to distributions: Fix a distribution \mathcal{P} over language L . Given an element $x \in \mathcal{X}$, $\mathcal{P}(x)$ denotes the probability mass \mathcal{P} assigns to x . The support of distribution \mathcal{P} is denoted by $\text{supp}(\mathcal{P})$, i.e., $\text{supp}(\mathcal{P}) := \{x \in L : \mathcal{P}(x) > 0\}$. As a shorthand, given a sequence x_1, x_2, \dots, x_n , for each index $1 \leq i \leq n$, we use $x_{\leq i}$ to denote the prefix $\{x_1, x_2, \dots, x_i\}$. Finally, we use standard notation for indicator functions and limits: Given an expression E (such as $h \neq K$ or $s \in K$), $\mathbb{1}\{E\}$ denotes the indicator that E is true. For a function $R : \mathbb{N} \rightarrow \mathbb{R}_{\geq 0}$, $R \downarrow 0$ denotes that $\lim_{n \rightarrow \infty} R(n) = 0$.

Language Collections and Membership Oracle to Languages. We always consider a *countable* collection of languages $\mathcal{L} = \{L_1, L_2, \dots\}$ and assume we have access to a *membership oracle* that, given an index i and a string s , outputs $\mathbb{1}\{s \in L_i\}$, as is standard in all prior works [Gol67; Ang80;

[KM24]. This is motivated by the fact that if these languages are “reasonable,” *e.g.*, they are generated by context-free grammars or decided by Turing machines [Sip12], then (1) there can be only countably many of them and (2) each of them admits a membership oracle. Finally, we reserve the letter K to denote the unknown target language $K \in \mathcal{L}$. We will say that an example x is a *positive* example for K if $x \in K$; otherwise x will be a *negative* example for K .

2.1 Language Identification and Generation in the Limit

In this section, we first present the Gold-Angluin model for identification in the limit and, then, Kleinberg and Mullainathan’s model for generation in the limit.

Language Identification in the Limit

The problem of language identification in the limit from positive examples was introduced by Gold [Gol67] and further studied by Angluin [Ang79; Ang80]. The setting is specified by a collection of languages $\mathcal{L} = \{L_1, L_2, \dots\}$. For a fixed collection \mathcal{L} , an adversary and an identifier play the following game: The adversary chooses a language K from \mathcal{L} without revealing it to the identifier, and it begins *enumerating* the strings of K (potentially with repetitions) x_1, x_2, \dots over a sequence of time steps $t = 1, 2, 3, \dots$. The adversary can repeat strings in its enumeration, but the crucial point is that for every string $x \in K$, there must be at least one time step t at which it appears.

At each time t , the identification algorithm I , given the previous examples x_1, x_2, \dots, x_t , outputs an index i_t that corresponds to its guess for the true language K .

Definition 9 (Language Identification in the Limit [Gol67]). *Fix some language K from collection \mathcal{L} . The identification algorithm I identifies K in the limit if there is some $t^* \in \mathbb{N}$ such that for all steps $t > t^*$, the identifier’s guess i_t satisfies $i_t = i_{t-1}$ and $L_{i_t} = K$. The language collection \mathcal{L} is identifiable in the limit if there is an identifier that identifies in the limit any $K \in \mathcal{L}$, for any enumeration of K .*

Gold [Gol67] showed that collections of finite cardinality languages, *i.e.*, each language in the collection \mathcal{L} is finite, can be identified in the limit from positive examples. This is true since in the limit, one will see all the elements of the target (finite) language, at which point it can be identified. The identification algorithm is the following: at time t , guess L to consist solely of the elements that have occurred in the sequence. Since L is finite, there will be a finite time after which all elements of L will have been revealed, so after that the algorithm will have identified the target. Interestingly, all finite collections of languages are also identifiable in the limit [Gol67].

A super-finite collection of languages denotes any collection which contains all languages of finite cardinality and at least one of infinite cardinality. Gold showed that super-finite collections of languages cannot be identified in the limit from positive examples. Further, he showed that negative examples help: any super-finite collection can be identified in the limit using positive and negative examples¹⁰ (the idea is simple: keep guessing the infinite language until seeing a negative example; then it reduces to the finite case).

¹⁰This means that the adversary presents an enumeration of the whole domain \mathcal{X} , with a label indicating whether the example is in the target language.

Theorem 2.1 ([Gol67]). *Let $\mathcal{L} = \{L_\infty, L_1, L_2, \dots\}$ be the language collection with $L_1 \subset L_2 \subset \dots \subset L_\infty = \cup_{i \geq 1} L_i$ and for each i , $|L_i| < \infty$. Then, there is no algorithm that identifies \mathcal{L} in the limit from positive examples. Moreover, this collection can be identified in the limit when the algorithm has access to both positive and negative examples.*

The above result already shows a separation in terms of identification between observing only positive examples and observing positive *and* negative examples in Gold’s model. Moreover, it raises the question of whether there exist non-trivial collections of languages identifiable in the limit from positive examples. In that direction, Angluin [Ang79] studied pattern languages (whose definition is not important for our work) and showed that for that collection identification in the limit is possible only with positive examples.

The next question is whether one can get a *characterization* of the language collections that can be identified from positive examples. Angluin [Ang80] resolved this problem.

Definition 10 (Angluin’s Condition [Ang80]). *Fix a language collection $\mathcal{L} = \{L_1, L_2, \dots\}$. Suppose there is a membership oracle which, given a string x and index i , answers $\mathbb{1}\{x \in L_i\}$. The collection \mathcal{L} is said to satisfy Angluin’s condition if there is an oracle that given an index i enumerates a set of finite strings T_i such that*

$$T_i \subseteq L_i \text{ and for all } j \geq 1, \text{ if } T_i \subseteq L_j \text{ then } L_j \text{ is not a proper subset of } L_i.$$

The difficulty in trying to identify a language from positive examples is the problem of *over-generalization*. If while seeing positive examples the algorithm specifies a language that is a proper superset of the true answer K , then by only seeing positive examples it will never see a counterexample to that language. This would be avoided with positive and negative examples. Angluin’s condition essentially ensures this over-generalization problem can be avoided by from just positive examples (without the help of negative examples).

Before proceeding to Angluin’s result, we stress one important point: inspecting Angluin’s definition, we can see that it requires access to a procedure that *finds* this set of strings T_i . This oracle is called a *tell-tale* oracle and is quite crucial for Angluin’s algorithm to work.

Definition 10 led to the following characterization.

Theorem 2.2 ([Ang80]). *A countable language collection \mathcal{L} is identifiable in the limit if and only if it satisfies Angluin’s criterion.*

Finally, let us consider the case of language identification with both positive and negative examples, *i.e.*, when the adversary provides an enumeration of the whole domain \mathcal{X} and every example has a label indicating whether it is in the true language K . We mention that focusing on algorithms equipped with membership oracle, the following result appears in Gold [Gol67].

Theorem 2.3 ([Gol67]). *Any countable language collection is identifiable in the limit from positive and negative examples.*

To see how the algorithm works, let $\mathcal{L} = \{L_1, L_2, \dots\}$ and denote by L_z the smallest indexed language in \mathcal{L} for which $L_z = K$. The algorithm observes an enumeration of the form $(x_t, y_t) \in \mathcal{X} \times \{0, 1\}$ for $t \geq 1$. Recall this means that $\mathbb{1}\{x_t \in K\} = y_t$. The algorithm works as follows: in every timestep $t \in \mathbb{N}$, it predicts the lowest index of a consistent language, *i.e.*, the smallest $j \in \mathbb{N}$

for which $\mathbb{1}\{x_\tau \in L_j\} = y_\tau$ for all $\tau \leq t$. Consider two cases: if $z = 1$, then the algorithm will never predict any language $L_{z'}, z' \geq 2$, so it will be correct from the first step. If $z > 1$, then for all $L_{z'}, z' < z$, that come before L_z in the enumeration of \mathcal{L} , there is a finite time $t_{z'}$ when the example $(x_{t_{z'}}, y_{t_{z'}})$ contradicts the language $L_{z'}$.

Language Generation in the Limit

We now move to language generation in the limit from positive examples, introduced by Kleinberg and Mullainathan [KM24]. The setup is exactly the same as in the Gold-Angluin model (the adversary provides an enumeration of K), but now the goal of the learner is to *generate unseen examples* from K instead of identifying the index of K . Their formal definition is the following.

Definition 11 (Language Generation in the Limit [KM24]). *Fix some language K from the collection $\mathcal{L} = \{L_1, L_2, \dots\}$ and a generating algorithm \mathcal{G} . At each step t , let $S_t \subseteq K$ be the set of all strings that the algorithm \mathcal{G} has seen so far. \mathcal{G} must output a string $x_t \notin S_t$ (its guess for an unseen string in K). The algorithm \mathcal{G} consistently generates from K in the limit if, for all enumerations of K , there is some $t^* \in \mathbb{N}$ such that for all steps $t \geq t^*$, the algorithm's guess a_t belongs to $K \setminus S_t$. The collection \mathcal{L} allows for consistent generation in the limit if there is an algorithm \mathcal{G} that, for any choice of the target language $K \in \mathcal{L}$, it consistently generates from K in the limit.*

Definition 11 straightforwardly generalizes to randomized algorithms; consider the same setup as before except that now the output string a_t may be randomized. The definition of generation is also the same except that instead of requiring $a_t \in K \setminus S_t$ one requires that the support A_t of the distribution from which a_t is sampled is non-empty and satisfies $A_t \subseteq K \setminus S_t$.

Observe that language generation requires that the algorithm's outputs are *consistent* with K (in the limit), but allows the algorithm to not generate certain strings from K . For instance, if K is the set of all strings, then the algorithm that always outputs even length strings (not in S_t), generates from K in the limit but also misses infinitely many strings in K (namely, all strings of odd length). Consistency is clearly a desirable notion: without consistency, algorithms may keep outputting strings outside the target language K which, when K is the set of all meaningful and true strings, inevitably leads to hallucinations.

A trivially consistent generator is one that outputs data already seen in the training set. As we already mentioned, we count such outputs as mistakes. This form of predicting unseen positive examples makes the task of generation interesting. At first sight, it seems that there is an easy strategy that achieves generation in the limit: given an enumeration of all hypotheses L_1, L_2, \dots , we sequentially generate from L_i ($i = 1, 2, \dots$) until it becomes inconsistent with the sample S_n ; then we move to L_{i+1} . This strategy seems natural for generation because we know that there is some index k such that the true language $K = L_k$. This idea has a fundamental issue, already reported by Kleinberg and Mullainathan [KM24]: if there exists an index i such that $i < k$ and $L_k \subsetneq L_i$, then the generator will get stuck at L_i and never update.

A non-trivial solution to this problem was given by Kleinberg and Mullainathan [KM24]. They show that all countable sets of languages in countable domains allow for generation in the limit from positive examples; this is in stark contrast with identification in the limit from positive examples.

Theorem 2.4 (Theorem 1 in Kleinberg and Mullainathan [KM24]). *There is an algorithm with the property that for any countable collection of languages $\mathcal{L} = \{L_1, L_2, \dots\}$, any target language $K \in \mathcal{L}$, and any enumeration of one of these languages K , the algorithm generates from K in the limit with positive examples.*

We now provide some intuition on how this algorithm works. Let L_1, L_2, \dots be an enumeration of the collection of languages and K be the true language. Let z be an index such that $L_z = K$. We say that a language L_i is consistent with the sample S_t at time t if S_t is contained in L_i . Now assume that we have two languages L_i and L_j with $L_i \subseteq L_j$ which are both consistent with S_t . Then, it is clear that the generating algorithm should prefer to generate from L_i rather than L_j : any $w \in L_i \setminus S_t$ satisfies $w \in L_j \setminus S_t$. This property inspired Kleinberg and Mullainathan [KM24] to define the notion of a *critical language*. Let $\mathcal{C}_n = \{L_1, L_2, \dots, L_n\}$. A language L_n is critical at step t if L_n is consistent with S_t and for every $L_i \in \mathcal{C}_n$ that is consistent with S_t , it must be $L_n \subseteq L_i$. There are some key properties upon which the generating algorithm is built:

- At any time, there is at least one language consistent with S_t , the true one $L_z = K$. Also, there is at least one critical language at any step t : for any t , the consistent language L_i with the lowest index i must be critical at step t , as it is the only consistent language in \mathcal{C}_i .
- There exists times t for which L_z (which is K) is not critical. But eventually, L_z will become critical at some step and then remain critical forever after that. Also, any critical language coming after L_z must be a subset of L_z , thus it is safe to generate from it.
- Hence the algorithm, roughly speaking, keeps track of a list of critical languages and generates from the last one in the list; this is because, after some finite index, all the critical languages are subsets of L_z and, hence, it is safe to generate from any of them.

More details about this algorithm will appear later on when we design our generation algorithms for the probabilistic setting (see [Section 5.2](#)).

3 Overview of Results

In this section, we present the formal statements of our main results. We begin with statistical rates for identification and for consistent generation (without the requirement of breadth) in [Section 3.1](#). Next, in [Section 3.2](#), we present our results for generation with breadth – showing that no generator from a large family of generators (that includes present-day LLMs) can generate with breadth from any language collection that is non-identifiable. Contrasting Kleinberg and Mullainathan [KM24]’s result for generation without breadth, these results show that generation with breadth is significantly harder – as hard as identification, for a large and natural class of generators. [Section 3.3](#) extends this impossibility result to a relaxation of generation with breadth, showing that even this relaxed definition of generation and breadth cannot be achieved by the same large class of generators. Finally, in [Section 3.4](#), we present additional results for identification when one has some additional structure (e.g., access to a stronger oracle) or information (e.g., negative examples).

3.1 Results for Identification and Generation Without Breadth

Prior work of Gold [Gol67] and Kleinberg and Mullainathan [KM24] studies language identification and generation in an online, *i.e.*, adversarial setting. In this work, we study the distributional versions of these problems. The identification problem we study is not new and, in fact, goes back to Angluin’s work in 1988 [Ang88]. However, Angluin [Ang88] does not provide any *rate* at which language identification can be achieved as the number of samples observed increases (when it is achievable).

Summary of Results in This Section. In this section, we give learning rates for both identification and generation (see Theorems 3.1 and 3.2 respectively). For both tasks, we study the learning curves – that is how the identification or generation error decays as the sample size increases. As a result, we extend the results of Gold [Gol67] and Kleinberg and Mullainathan [KM24] to the statistical setting. Our results in this section achieve a near-optimal rate for identification (Theorem 3.1) and an optimal rate for generation (Theorem 3.2).

3.1.1 Universal Rates: Model and Preliminaries

We work under the *universal rates* framework, introduced by Bousquet, Hanneke, Moran, van Handel, and Yehudayoff [BHM+21], in order to capture the notion of a learning curve for language identification and generation. Following the notation we used before, recall that we have a countable set of languages $\mathcal{L} = \{L_1, L_2, \dots\}$, where each $L \in \mathcal{L}$ is also countable and $\cup_{L \in \mathcal{L}} L \subseteq \mathcal{X}$, for some countable domain \mathcal{X} . Recall the notion of a valid distribution proposed by Angluin [Ang88] in this setting (Definition 1). Intuitively, this condition can be thought of as the equivalent of *realizability* in the classification setting.

The learning algorithm is a sequence of (universally measurable and computable) functions $\{h_n\}_{n \in \mathbb{N}}$, where n captures the size of the training set. We are interested in understanding the behavior of the *error* of the algorithm, which is defined appropriately based on the downstream task – either identification or generation for this paper. Given some rate function $R: \mathbb{N} \rightarrow [0, 1]$ we say that we can achieve rate $R(n)$ for the set of language \mathcal{L} and the loss function $\text{er}(\cdot)$ if there exists a learning algorithm $\{h_n\}_{n \in \mathbb{N}}$ whose error satisfies

$$(\forall \text{ valid } \mathcal{P}) (\exists C, c) \quad \text{such that} \quad \mathbb{E}[\text{er}(h_n)] \leq C \cdot R(cn), \quad \forall n \in \mathbb{N}.$$

Crucially, these learning curves are distribution-specific; the constants c, C depend on \mathcal{P} but the rate R holds universally for all valid distributions. Such learning curves are a well-studied topic in learning theory [Sch97; AL98; BHM+21; VL23]. The above gives rise to the following definition.

Definition 12 (Learning Rates [BHM+21]). *Given a language collection \mathcal{L} , an error function $\text{er}(\cdot)$, and a rate function $R: \mathbb{N} \rightarrow [0, 1]$ satisfying $\lim_{n \rightarrow \infty} R(n) \rightarrow 0$, we say:*

- *Rate R is achievable for \mathcal{L} if there is an algorithm $\{h_n\}_{n \in \mathbb{N}}$ such that for every valid distribution \mathcal{P} , there exist c, C for which $\mathbb{E}[\text{er}(h_n)] \leq C \cdot R(c \cdot n), \forall n \in \mathbb{N}$.*
- *No rate faster than $R(n)$ is achievable for \mathcal{L} if for all algorithms $\{h_n\}_{n \in \mathbb{N}}$ there exists a valid distribution \mathcal{P} and c, C for which $\mathbb{E}[\text{er}(h_n)] \geq C \cdot R(c \cdot n)$, for infinitely many $n \in \mathbb{N}$.*

Further, we have the following.

- (Optimal Rate) Rate R is optimal for \mathcal{L} if it is achievable and no rate faster than R is achievable.
- (No Rate) We say that \mathcal{L} admits no rate if for every algorithm $\{h_n\}_{n \in \mathbb{N}}$ there exists a valid distribution \mathcal{P} such that $\limsup_{n \rightarrow \infty} \mathbb{E}[\text{er}(h_n)] > 0$.

In the case of identification, to avoid trivial cases, we consider collections \mathcal{L} that contain at least two distinct languages that contain one common element.

Definition 13 (Non-Trivial Collections of Languages for Identification). *A language collection \mathcal{L} is non-trivial for identification if there exist two languages $L_1, L_2 \in \mathcal{L}$ such that $L_1 \neq L_2$ and $|L_1 \cap L_2| > 0$.*

Notice that if the collection \mathcal{L} does not satisfy [Definition 13](#), then one can identify the target language K immediately after observing a single element from K .

In the case of generation, the “non-triviality” condition turns out to be more nuanced, *e.g.*, compared to the case of identification above or binary classification [\[BHM+21\]](#). We give an informal definition below, and we refer to [Definition 17](#) for the formal one and a discussion about its necessity.

Informal Definition 1 (Non-Trivial Collections of Languages for Generation, see [Definition 17](#)). *A language collection \mathcal{L} is non-trivial for generation if any algorithm needs to see at least two examples from the target language to be able to generate from it.*

3.1.2 Universal Rates for Identification

For any language collection $\mathcal{L} = \{L_1, L_2, \dots\}$ and $n \in \mathbb{N}$, with true language $K \in \mathcal{L}$, and set of examples $x_1, \dots, x_n \in \mathcal{X}^n$, an identification algorithm I_n gets as input x_1, \dots, x_n and outputs an index $I_n(x_1, \dots, x_n)$. We define the *identification error* of the learner $\{I_n: \mathcal{X}^n \rightarrow \mathbb{N}\}_{n \in \mathbb{N}}$ as

$$\text{er}(I_n(x_1, \dots, x_n)) = \mathbb{1}\{L_{I_n(x_1, \dots, x_n)} \neq K\}. \quad (4)$$

Under this definition, $\mathbb{E}_{x_1, \dots, x_n \sim \mathcal{P}}[\text{er}(I_n)] = \Pr_{x_1, \dots, x_n \sim \mathcal{P}}[L_{I_n(x_1, \dots, x_n)} \neq K]$, *i.e.*, the probability that it fails to identify the correct language after it sees n examples from \mathcal{P} .¹¹

Our main result for identification is a fundamental dichotomy: every non-trivial collection of languages is identifiable with positive examples at either an (almost) exponential rate or it is not identifiable at any rate.

Theorem 3.1 (Dichotomy of Rates for Identification with Positive Examples). *For every collection of countably many languages \mathcal{L} that is non-trivial for identification exactly one of the following holds:*

- For every $g(n) = o(n)$ there exists a learner that identifies \mathcal{L} at rate $e^{-g(n)}$. Moreover, no learner can achieve a rate faster than e^{-n} .
- \mathcal{L} is not identifiable at any rate.

Concretely, the first condition holds for \mathcal{L} if and only if it satisfies Angluin’s condition ([Definition 10](#)).

¹¹One subtle point is that this definition allows the learner to output any index $j \in \mathbb{N}$ such that $L_j = K$ and there may be many such indices since we do not assume all languages in \mathcal{L} are distinct. Our identification algorithms will have the property that they output the smallest index at which K appears in $\mathcal{L} = \{L_1, L_2, \dots\}$.

This dichotomy of rates differs from prior universal rates for classification where the usual theme is a trichotomy of rates [BHM+21; KVK22; HMZ23]. Moreover, while in the universal setting for binary classification, any measurable class of functions is learnable at arbitrarily slow rates, in identification, this is not the case: there exist collections of languages that do not admit a Bayes consistent learner and these are exactly the collections that do not satisfy Angluin’s condition. For the full proof, we refer the reader to [Section 5.1](#).

3.1.3 Universal Rates for Consistent Generation

The main difference between this setting and the setting of language identification is the definition of the error rate. There exists a valid text-generating distribution \mathcal{P} , meaning one that is supported on some target language $K \in \mathcal{L}$, and the learning (or rather, generating) algorithm is a sequence of (universally measurable and computable) functions $\{\mathcal{G}_n: \mathcal{X}^n \rightarrow \mathcal{X}\}_{n \in \mathbb{N}}$, where each \mathcal{G}_n takes as input n samples generated i.i.d. from \mathcal{P} and outputs a new word, with the goal that this word belongs to the target language (see [Remark 1](#)). As in the online setting, to avoid trivial solutions, we want to generate examples that do not appear in the training set.

Remark 1 (Notation for Generating Algorithms). More formally, a generating algorithm is a collection of mappings $\{\mathcal{G}_n\}_{n \in \mathbb{N}}$, where for each n , \mathcal{G}_n is a mapping from the domain of n training samples \mathcal{X}^n to the set of “generators” or (randomized) Turing machines \mathcal{G} that, on each execution, output a sample from \mathcal{X} . For this section, it is sufficient to imagine generators as being deterministic (*i.e.*, generating samples from a point mass) and, hence, we simplify writing \mathcal{G}_n as a mapping from \mathcal{X}^n to \mathcal{X} . In the next section, where we study generation with breadth, to have any hope of achieving breadth, we need to consider \mathcal{G}_n in its full generality as a mapping from \mathcal{X}^n to \mathcal{G} .

Now, we are ready to define the *generation error*: for any $n \in \mathbb{N}$ and set of examples $x_1, \dots, x_n \in \mathcal{X}^n$ we define the generation error of the learner $\{\mathcal{G}_n: \mathcal{X}^n \rightarrow \mathcal{X}\}_{n \in \mathbb{N}}$ for this task as

$$\text{er}(\mathcal{G}_n(x_1, \dots, x_n)) = \mathbb{1}\{\mathcal{G}_n(x_1, \dots, x_n) \notin K \setminus \{x_1, \dots, x_n\}\}. \quad (5)$$

Notice that, under this definition,

$$\mathbb{E}_{x_1, \dots, x_n \sim \mathcal{P}^n}[\text{er}(\mathcal{G}_n(x_1, \dots, x_n))] = \Pr_{x_1, \dots, x_n \sim \mathcal{P}^n}[\mathcal{G}_n(x_1, \dots, x_n) \notin K \setminus \{x_1, \dots, x_n\}],$$

i.e., the probability that the learner fails to generate a new word from the target language after observing n examples from it. Our main result in this section is that we can achieve consistent generation with exponential rates.

Theorem 3.2 (Rates for Generation). *For every countable collection of languages \mathcal{L} there exists a generating algorithm that generates from \mathcal{L} at rate e^{-n} . Conversely, for every collection of languages that is non-trivial for generation ([Definition 17](#)), no generating algorithm can achieve rate faster than e^{-n} .*

Surprisingly, this shows that consistent generation can be achieved at an exponential rate for *any* countable collection of languages. We mention that the result we prove is slightly stronger: we show that, for any \mathcal{L} , with probability at least $1 - C \cdot e^{-c \cdot n}$, we can generate *infinitely* many new strings from K , after training the algorithm on n examples – not just a single word. Together, [Theorems 3.1](#) and [3.2](#) show that the stark separation between language identification and generation in the online setting, obtained by Kleinberg and Mullainathan [[KM24](#)], also extends to the statistical setting of Angluin [[Ang88](#)] and Bousquet et al. [[BHM+21](#)]. The proof of [Theorem 3.2](#) appears in [Section 5.2](#); see [Figure 3](#) for an outline of the proof.

3.2 Results for Generation With Breadth

Next, we present our results for language generation with breadth. Clearly, generation with breadth is a stronger requirement than generation. But, at least intuitively, it is weaker than identification: it only requires one to generate samples from the entire support of K and not identify the index of K . Contrary to this intuition, our results show that, for a large class of generators, generation with breadth is as hard as identification. Our results show that, while this class of generators is powerful enough to generate without breadth, no generator in this class can achieve generation with breadth for non-identifiable collections of languages.

3.2.1 Membership Oracle Problem

The family of generators we consider is implicitly determined by the decidability of a certain problem associated with the generator.

Definition 5 (Membership Oracle Problem). *Given a generator \mathcal{G} , the membership oracle problem for \mathcal{G} , denoted as $\text{MOP}(\mathcal{G})$, is defined as follows: given the description of \mathcal{G} and a string x , output Yes if $x \in \text{supp}(\mathcal{G})$ and output No otherwise.*

As mentioned before the decidability of problems is extensively studied in formal languages and complexity theory [Sip12]. Our main result (Informal Theorem 1 whose formal statement appears as Theorem 3.3) applies to any generator \mathcal{G} for which $\text{MOP}(\mathcal{G})$ is decidable. Note that our result only needs a decider of $\text{MOP}(\mathcal{G})$ to exist – this is purely a property of the generation algorithm used – and it does *not*, for instance, require the individuals training the generator or the users to have access to the decider in any fashion.

To gain some intuition about the membership oracle problem, let us consider a simple example.

Example 1 (Standard Next-Token Predictor). Let $\mathcal{G}_{\text{next-token}}$ be a text generator or language model that generates text token-by-token: at each step t , it generates certain scores $\{p_t(\sigma) : \sigma \in \Sigma\}$ and outputs token σ with probability $\propto p_t(\sigma)$. It is not important how these scores are generated. They can be generated in various ways. For instance, they can be the logit-scores of transformer-based models. They could also, be generated by thresholding logit-scores in any complicated but computable way – such as, by using beam search, top- K , or top- p sampling [HBD+20]. $\text{MOP}(\mathcal{G}_{\text{next-token}})$ is decidable and, in fact, there is a simple decider: given a string w of length n , it computes the scores for the first n iterations; where in the t -th iteration ($t > 1$), it conditions on the event that $\mathcal{G}_{\text{next-token}}$ has generated the string $w_1 w_2 \dots w_{t-1}$ so far. Then it computes the following function and outputs the result

$$\begin{cases} \text{Yes} & \text{if } \prod_{t=1}^n p_t(w_t) > 0, \\ \text{No} & \text{otherwise.} \end{cases}$$

We stress that our main result only needs the existence of such a decider, and does not require the individuals training the generator or the users to have any access to it.

3.2.2 Results for Generators for Which $\text{MOP}(\cdot)$ Is Decidable

Before stating our result about the rate at which generation with breadth can be achieved, we need to define the corresponding error function. For the error to make sense, let \mathcal{G} be the set of (randomized) Turing machines that do not take any input and output one element from \mathcal{X} (on each execution). Given a target language K and examples $x_1, \dots, x_n \in \mathcal{X}$, we define the *error for generation with breadth* for the learner $\{\mathcal{G}_n: \mathcal{X}^n \rightarrow \mathcal{G}\}_{n \in \mathbb{N}}$ as

$$\text{er}(\mathcal{G}_n(x_1, \dots, x_n)) = \mathbb{1}\{\text{supp}(\mathcal{G}_n(x_1, \dots, x_n)) \neq K \setminus \{x_1, \dots, x_n\}\},$$

where $\text{supp}(\mathcal{G}_n(x_1, \dots, x_n))$ is the set of strings $\mathcal{G}_n(x_1, \dots, x_n)$ can output with positive probability, *i.e.*, it is the support of the distribution of outputs of $\mathcal{G}_n(x_1, \dots, x_n)$. The above means that we count each step t as a mistake if the generating algorithm has a positive probability of outputting a string outside of K (*i.e.*, hallucination), a zero probability of outputting an unseen element of K (*i.e.*, mode collapse), or a positive probability of repeating a seen training example.

Remark 2 (Generating Examples From the Training Set). For generation without breadth, it is important to restrict the generator from outputting elements it has already seen. Otherwise, the futile generator, which always outputs the first training sample it sees, achieves generation without breadth. This requirement, however, is not important for generation with breadth: any generator \mathcal{G} that generates with breadth without repeating training examples can be converted to one \mathcal{G}' that generates with breadth and repeats the training examples and vice versa.¹² Hence, all of our results hold with either notion of generation with breadth.

Our main result shows a separation between the rates achievable for generation with and without breadth by any generating algorithm for which $\text{MOP}(\cdot)$ is decidable.

Theorem 3.3. *Let \mathfrak{G} be the set of all generating algorithms (\mathcal{G}_n) for which $\text{MOP}(\cdot)$ is decidable (Definitions 5 and 6). For every collection of countably many languages \mathcal{L} that is non-trivial for generation (Definition 17) and not identifiable in the limit:*

- *No generating algorithm in \mathfrak{G} generates with breadth from \mathcal{L} at any rate; and*
- *There is a generating algorithm in \mathfrak{G} that generates consistently without breadth from \mathcal{L} at rate e^{-n} . Conversely, no generating algorithm (even outside of \mathfrak{G}) can generate at a rate faster than e^{-n} .*

Further, for any collection of countably many languages \mathcal{L} that is non-trivial for generation (Definition 17) and identifiable in the limit, and for any $g(n) = o(n)$, there is a generating algorithm in \mathfrak{G} that generates with breadth from \mathcal{L} at rate $e^{-g(n)}$. Conversely, no generation algorithm can generate consistently at a rate faster than e^{-n} , even without the breadth requirement.

Thus, while generation without breadth is achievable for any countable collection of languages (whether it is identifiable or non-identifiable), generators in \mathfrak{G} can only generate with breadth from identifiable collections – which are a very restricted subset of all languages [Gol67; Ang80; KM24]. It remains to discuss which types of generators $\text{MOP}(\cdot)$ is decidable for, and we present a

¹²For instance, \mathcal{G}' can run \mathcal{G} with probability $1/2$ and with the remaining $1/2$ probability output a training sample selected uniformly at random. Given \mathcal{G}' , \mathcal{G} can be implemented by rejection sampling as follows: repeatedly execute \mathcal{G}' until it generates an unseen element x and output x .

large family in the next section. Meanwhile, due to [Example 1](#), it is already clear that [Theorem 3.4](#) applies to present-day LLMs. The proof of this result appears in [Section 6.2](#); see [Figure 5](#) for an outline of the proof.

Our negative result leaves several interesting questions open which we already discussed in [Section 1.4](#).

3.2.3 A Family of Generators for Which $\text{MOP}(\cdot)$ Is Decidable

[Example 1](#) already shows that $\text{MOP}(\cdot)$ is decidable for many existing language models. Next, we show that $\text{MOP}(\cdot)$ is decidable under even fewer restrictions on the generator \mathcal{G} – informally, we will allow for *any* generator which generates text token-by-token.

Definition 14 (Token-by-Token Generators). *Token-by-token generators \mathcal{G} are parameterized by randomized Turing machines M . M can be randomized and halts on all inputs. Given M , the corresponding token-by-token generator \mathcal{G}_M generates outputs as follows: for each $t \in \mathbb{N}$,*

1. *Let $w_1 w_2 \dots w_{t-1}$ be the tokens generated so far.*
2. *Let A_t be any auxiliary information generated so far, where A_1 is the empty string.*
3. *Generate (s_t, A_{t+1}) by running M with input $w_1 w_2 \dots w_{t-1}$ and A_t .*
4. *If $s_t = \text{EOS}$ (i.e., end of string), then output $s_1 \dots s_t$ and halt; otherwise proceed to iteration $t + 1$.*

Note that token-by-token generators are a very powerful class: for instance, any distribution over Σ^* for some finite alphabet Σ admits a token-by-token generator by the Bayes rule. That said, of course, one can also construct non-token-by-token generators.

We show that $\text{MOP}(\mathcal{G})$ is decidable for all token-by-token generators.

Theorem 3.4. *For any token-by-token generator \mathcal{G} , $\text{MOP}(\mathcal{G})$ is decidable.*

Next, we demonstrate that token-by-token generators capture several interesting language models. First, the family of token-by-token generators captures existing large language models (LLMs): for instance, to simulate an LLM L , we define the next token predictor M as a Turing machine that simulates L on the provided string until L generates one new token. Further, since we do not place computational restrictions on M , M can also simulate interactions between LLMs or auxiliary systems that select a suitable LLM to respond depending on the request—a strategy that has led to recent advances in text generation [[SDD+23](#); [JSR+24](#); [Kni24](#); [Tea24](#)]. Finally, due to a reduction to the halting problem, there are some generators for which $\text{MOP}(\cdot)$ is undecidable and give an explicit example in [Section A](#).

Remark 3 (Noisy Membership Oracle). A supposedly weaker requirement than the decidability of $\text{MOP}(\cdot)$ is the existence of a *noisy oracle* that, given a string x , correctly (and in finite time) decides the membership of x into $\text{supp}(\mathcal{G})$ with a probability at least $2/3$. However, due to the folklore result that $\text{BPP} \subseteq \text{EXP}$ [[AB09](#)], a noisy oracle is equivalent to the decidability of $\text{MOP}(\cdot)$.

3.2.4 Results for Generation With Breadth in the Limit

In this section, we state the implications of our techniques for generation with breadth in the adversarial or online setting of Gold [Gol67] and Angluin [Ang79; Ang80].

Theorem 3.5. *For every non-identifiable collection of countably many languages \mathcal{L} , no generating algorithm, for which $\text{MOP}(\cdot)$ (Definitions 5 and 6) is decidable, can generate with breadth from \mathcal{L} in the limit. If \mathcal{L} is identifiable, then there is a generator \mathcal{G} (for which $\text{MOP}(\mathcal{G})$ is decidable) that generates with breadth from \mathcal{L} .*

This result makes important progress on a question left open by Kleinberg and Mullainathan [KM24] for a fairly large family of generators, which includes all iterative generators due to Theorem 3.4. In particular, $\text{MOP}(\cdot)$ is decidable for the generation algorithm of Kleinberg and Mullainathan [KM24] (since it is deterministic and the unique element it outputs can be computed by executing the algorithm) and, hence, the above result shows that Kleinberg and Mullainathan [KM24]’s algorithm cannot generate with breadth in the limit from any non-identifiable collection. Further, in Section 3.3 we strengthen this result by showing that even a relaxed notion of generation with breadth remains unreachable for a large class of generators. The proof of this result can be found in Section 6.3.

3.3 Results for Generation With Approximate Consistency and Breadth

In this section, we study a relaxation of generation with breadth, which we call unambiguous generation, and ask: *Is there a generator that unambiguously generates from a non-identifiable collection?*

We recall that, in this section, we will allow the generator to repeat examples in the training data. Like all of our results with breadth, this choice is not crucial, and all of the results have analogs where the generator does not repeat training examples (Remark 2). We make this choice to simplify the notation.

We refer the reader to Section 1.3 for a discussion and motivation of the definition for unambiguous generation, which we restate below.

Definition 8 (Unambiguous Generator). *A generating algorithm $\mathcal{G} = (\mathcal{G}_n)$ is unambiguous for a language collection \mathcal{L} if, for any $K \in \mathcal{L}$ and every enumeration of K , its support eventually becomes closer to K than to any other language $L \neq K$ in \mathcal{L} in terms of the symmetric difference metric, i.e., there exists some $n^* \in \mathbb{N}$ such that for all $n \geq n^*$ it holds that*

$$|\text{supp}(\mathcal{G}_n) \triangle K| < \min_{L \in \mathcal{L}: L \neq K} |\text{supp}(\mathcal{G}_n) \triangle L|,$$

where recall that for two sets S and T , $S \triangle T := (S \setminus T) \cup (T \setminus S)$.

This notion is a significant relaxation of generation with breadth that we considered so far (see Section 3.2): Not only does it allow the generator to hallucinate certain strings not in the target K and omit strings actually in K for arbitrarily long, the number of hallucinations and omissions can be very large and, depending on the structure of the language collection \mathcal{L} , even arbitrarily large.

Surprisingly, even this very weak notion of “generation with breadth” turns out to be unachievable by a very large family of generators. Concretely, it is unachievable by any generator for which $\text{MOP}(\cdot)$ is decidable and that satisfies the natural property that it stabilizes after a finite time. We state the formal notion of stability below.

Definition 7 (Stability). A generating algorithm (\mathcal{G}_n) is stable for a language collection \mathcal{L} if for any target language $K \in \mathcal{L}$ and for any enumeration of K , there is some finite $n^* \in \mathbb{N}$ such that for all $n, n' \geq n^*$, it holds that $\text{supp}(\mathcal{G}_n) = \text{supp}(\mathcal{G}_{n'})$.

Before turning to our formal result, we need to construct the error function that defines unambiguous generation, and we use the natural choice: for a language K and examples $x_{i_1}, \dots, x_{i_n} \in \mathcal{X}$, we denote and we define the *error for unambiguous generation* for the generating algorithm $\{\mathcal{G}_n: \mathcal{X}^n \rightarrow \mathcal{G}\}_{n \in \mathbb{N}}$ on input $S_n = \{x_{i_1}, \dots, x_{i_n}\}$ as¹³

$$\text{er}(\mathcal{G}_n(S_n)) = 1 \left\{ |\text{supp}(\mathcal{G}_n(S_n)) \Delta K| < \min_{L \in \mathcal{L}: L \neq K} |\text{supp}(\mathcal{G}_n(S_n)) \Delta L| \right\},$$

where $\text{supp}(\mathcal{G}_n(S_n))$ is the set of strings $\mathcal{G}_n(S_n)$ can output with positive probability. Similar to the case of identification and generation, we say that an algorithm achieves unambiguous generation for a collection \mathcal{L} at some rate R , where $R: \mathbb{N} \rightarrow \mathbb{R}_{\geq 0}, R \downarrow 0$, if for any valid distribution \mathcal{P} with respect to \mathcal{L} there are $c, C > 0$ so that $\mathbb{E}_{x_{i_1}, \dots, x_{i_n} \sim \mathcal{P}^n} [\text{er}(\mathcal{G}_n(x_{i_1}, \dots, x_{i_n}))] \leq C \cdot R(c \cdot n)$. The following result shows that this notion of generation is not achievable, for a large and natural class of generating algorithms.

Theorem 3.6 (Impossibility of Unambiguous Generation). *For every non-identifiable collection of countably many languages \mathcal{L} , no stable generating algorithm, for which $\text{MOP}(\cdot)$ (Definitions 5 and 6) is decidable, can unambiguously generate from \mathcal{L} at any rate.*

Note that while this result has a benign requirement that the generator is stable, it already considerably extends our main result [Theorem 3.3](#), since any generator that achieves breadth must be stable – otherwise, its support cannot settle on the target language K . (To be precise, [Theorem 3.3](#) required generators to not repeat their training examples, but this requirement is not crucial and any generator that does repeat its training examples can be converted into one that does not repeat its training examples, and vice-versa; see [Remark 2](#).)

In addition to [Theorem 3.6](#), we also prove its analog in the online setting – significantly extending our earlier impossibility result in the online setting ([Theorem 3.5](#)). Before stating the result in the online, we introduce unambiguity in the limit, which is a natural counterpart to its statistical definition:

- A generating algorithm $\mathcal{G} = (\mathcal{G}_n)$ is said to be unambiguous for a collection $\mathcal{L} = \{L_1, L_2, \dots\}$ if, for any $K \in \mathcal{L}$ and enumeration x_{i_1}, x_{i_2}, \dots of K , there is an $n_0 \geq 1$, such after seeing $n \geq n_0$ elements $S_n = x_{i_1}, \dots, x_{i_n}$,

$$|\text{supp}(\mathcal{G}_n(S_n)) \Delta K| < \min_{L \in \mathcal{L}: L \neq K} |\text{supp}(\mathcal{G}_n(S_n)) \Delta L|.$$

Theorem 3.7 (Impossibility of Unambiguous Generation in the Limit). *For every non-identifiable collection of countably many languages \mathcal{L} , no generating algorithm stable in the limit for which $\text{MOP}(\cdot)$ (Definitions 5 and 6) is decidable can unambiguously generate from \mathcal{L} in the limit.*

¹³Recall that \mathcal{G} is the set of (randomized) Turing machines that do not take any input and output one element from \mathcal{X} (on each execution).

The proofs of [Theorems 3.6](#) and [3.7](#) appear in [Section 7.1](#) and [7.2](#), respectively. To develop some intuition, we recommend reading the proof of [Theorem 3.7](#) before the proof of [Theorem 3.6](#).

Remark 4. In [Section C](#), we study another notion of generation with approximate breadth which, informally, requires that the generating algorithm is consistent and puts zero mass only on *finitely* many points of the target language K . This is also a weakening of generation with breadth and turns out to be incomparable to the notion of unambiguous generation studied in this section.

3.4 Further Results for Identification

In this section, we present identification algorithms that achieve *exact* exponential rate when one has some additional structure – access to a stronger oracle, or a finite collection \mathcal{L} , or a countable collection \mathcal{L} of finite languages – or additional information – negative examples.

In [Section 3.4.1](#), we allow the identifier to make queries of the form “is $L_i \subseteq L_j$?” Next, in [Section 3.4.2](#), we consider generation from collections \mathcal{L} containing finitely many languages $\mathcal{L} = \{L_1, L_2, \dots, L_k\}$. (Note that each language in \mathcal{L} can still be infinite.) Finally, in [Section 3.4.4](#), in addition to positive examples, we also give the identifier access to negative examples (*i.e.*, elements $x \in \mathcal{X}$ not in the target language K).

3.4.1 Exponential Rates for Identification Using Subset Oracle

Our first result shows that when \mathcal{L} satisfies Angluin’s condition and the learning algorithm has access to a subset oracle for \mathcal{L} (which answers queries of the form “ $L_i \subseteq L_j$?”) then it is possible to achieve exact exponential rates.

Proposition 3.8. *For every countable language collection \mathcal{L} that satisfies Angluin’s condition ([Definition 10](#)), there exists a learning algorithm that has access to a subset oracle for \mathcal{L} and identifies \mathcal{L} at a rate e^{-n} . Formally, a subset oracle is a primitive that, given two indices i and j , outputs Yes if $L_i \subseteq L_j$; otherwise, it outputs No.*

Recall that our algorithm that achieves almost exponential rates requires merely black-box access to an algorithm that identifies \mathcal{L} in the limit. In other words, it does not make use of the particular structure of the online identification algorithm. To achieve exact exponential rates, we make use of a particular algorithm: the one proposed by Kleinberg and Mullainathan [[KM24](#)]. At a high level, the proof consists of the following steps:

- C1 First, we show that Kleinberg and Mullainathan [[KM24](#)]’s algorithm with access to a subset oracle for \mathcal{L} can, in fact, *identify* the target language (see [Section B.1](#)).
- C2 Next, we identify a sufficient condition that allows one to use any identification algorithm that identifies \mathcal{L} in the limit to obtain exponential rates (see [Lemma 8.1](#)). Interestingly, this conversion does not need *any* changes to the identification algorithm.
- C3 Finally, we show that the algorithm of Kleinberg and Mullainathan [[KM24](#)] satisfies this condition.

The full proof of [Proposition 3.8](#) appears in [Section 8.1](#).

3.4.2 Exponential Rates for Identification of Finite Collections

We now shift our attention to finite collections of languages. Gold [Gol67] and Angluin [Ang80] showed that all finite collections are identifiable in the limit. We show that for such collections we can get exact exponential rates, *without* the need of the subset oracle we used in the previous result (Proposition 3.8).

Proposition 3.9. *For every finite language collection \mathcal{L} , there exists a learning algorithm which identifies \mathcal{L} at a rate e^{-n} .*

The proof of this result builds on the proof of Proposition 3.8. In particular, we show that the algorithm of prior work satisfies the sufficient condition that allows an algorithm that identifies \mathcal{L} in the limit to obtain exponential rates (Condition C2). The full proof of Proposition 3.9 appears in Section 8.2.

3.4.3 Exponential Rates for Identification of Collections of Finite Languages

We now move on to a result about identifying countable collections of *finite* languages with exactly exponential rates. Gold [Gol67] showed such collections are identifiable in the limit through a very simple algorithm: predict the first language that contains the set of all examples seen so far. We show that for such collections we can get exact exponential rates.

Proposition 3.10. *For every countable language collection \mathcal{L} that only contains languages of finite size, there exists a learning algorithm which identifies \mathcal{L} at a rate e^{-n} .*

The idea of the proof is simple. Since any valid distribution has finite support, for large enough n , the sample will contain all the elements of the support with probability $1 - C \cdot e^{-c \cdot n}$. The formal proof of Proposition 3.10 appears in Section 8.3.

3.4.4 Exponential Rates for Identification from Positive and Negative Examples

We now shift our attention to a setting – introduced by Gold [Gol67] – where, in addition to an enumeration of the target language K , one also receives an enumeration of $\mathcal{X} \setminus K$.

Let us first recall the difference between the different types of information in the two settings. In the case of just positive examples (considered so far), the adversary picks a target language K from \mathcal{L} along with an enumeration of this language, and presents the examples from this enumeration sequentially to the learner. In the case of positive and negative examples, the adversary again picks a target language K from \mathcal{L} , but now it chooses a *labeled* enumeration of the whole domain \mathcal{X} , where now the label of each element indicates whether it belongs to the target language K or not. It is known that every countable collection of languages is identifiable in the limit with positive and negative examples [Gol67].

Naturally, we need a different notion of a *valid* distribution in this setting. We adopt a definition that was proposed by Angluin [Ang88].

Definition 15 (Valid Distributions Under Positive and Negative Examples [Ang88]). *A distribution \mathcal{P} over $\mathcal{X} \times \{0, 1\}$ is valid with respect to a collection of languages \mathcal{L} if and only if $\text{supp}(\mathcal{P}) = \mathcal{X}$ and there exists some $K \in \mathcal{L}$ such that for all $x \in \mathcal{X}$ it holds that $\Pr_{(X,Y) \sim \mathcal{P}}[Y = 1 \mid X = x] = \mathbb{1}\{x \in K\}$.*

Our main result in this setting is that every countable collection of languages is identifiable with positive and negative examples at an optimal exponential rate.

Theorem 3.11 (Identification with Positive and Negative Examples). *For every countable collection of languages \mathcal{L} , there exists a learner that identifies \mathcal{L} at rate e^{-n} . Conversely, for every countable collection of languages \mathcal{L} that is non-trivial for identification, no learner can identify \mathcal{L} at rate faster than e^{-n} .*

The proof of [Theorem 3.11](#) appears in [Section 8.4](#). Our proof of this result is inspired by the approach of Bousquet et al. [[BHM+21](#)]. First, we show the exponential rates lower bound by directly using a result of Bousquet et al. [[BHM+21](#)]. In order to get the upper bound, we use a black-box transformation from any learner that identifies \mathcal{L} in the limit, to a learner that achieves exponential rates in the statistical setting.

The approach shares similarities to the one with just positive examples (see [Section 1.2](#), Paragraph B). The crucial reason why we can obtain exactly exponential rates here, instead of almost exponential rates as in the previous setting, is that we can use the negative examples to accurately estimate the correct sizes of the batches we use, instead of having to use “guesses” of increasing size as we did in the setting of just positive examples.

To give a more concrete comparison to the binary classification setting of Bousquet et al. [[BHM+21](#)], let us first explain some results from this work. Bousquet et al. [[BHM+21](#)] define the following infinite game sequential, which is appropriately rephrased using the terminology from our work looks as follows:

- In every round, the adversary presents a word $x_t \in \mathcal{X}$ to the learner.
- Subsequently, the learner predicts a label from $\{0, 1\}$ for this word, denoted by \hat{y}_t .
- Then, the adversary reveals the true label y_t to the learner.

The only constraint on the adversary is that at any given point $t \in \mathbb{N}$, there has to be some language $K \in \mathcal{L}$ such that $y_{t'} = \mathbb{1}\{x_{t'} \in K\}$, for all $t' \leq t$. In other words, the choices of the labels have to be consistent with some language $K \in \mathcal{L}$. Crucially, the consistent language does not need to be fixed in advance and it can keep changing throughout the interaction. In their setting, the learner “wins” the game if it makes only finitely many mistakes. They provide a necessary and sufficient condition on the structure of \mathcal{L} which determines the existence of a winning strategy for the learner: the learner can win this game if and only if \mathcal{L} does not have an infinite Littlestone tree (see [Definition 22](#)). Interestingly, this condition does not capture the existence of a winning strategy for the learner in Gold’s setting: we have constructed a language family \mathcal{L} which has an infinite Littlestone tree, but it is identifiable in the limit from positive and negative examples. Perhaps more surprisingly, this language is identifiable even with just positive examples. The construction appears in [Section D](#).

4 Organization of the Rest of the Paper

We next describe the organization of the rest of the paper.

- The proofs of [Section 3.1](#) (statistical rates for identification and generation) can be found in [Section 5](#). The proof for the identification universal rates appears in [Section 5.1](#) and for generation in [Section 5.2](#).

- The proofs of [Section 3.2](#) can be found in [Section 6](#). In [Section 6.1](#), we discuss the decidability of $\text{MOP}(\cdot)$. In [Section 6.2](#) we provide our main result that generation with breadth is not possible for generating algorithms for which $\text{MOP}(\cdot)$ is decidable. Finally, in [Section 6.3](#), we see the implications of this result for generation in the limit.
- The proofs of [Section 3.3](#) appear in [Section 7](#). In [Section 7.1](#) we give the proof of the result in the online setting and in [Section 7.2](#) the proof of the result in the statistical setting.
- The proofs of [Section 3.4](#) appear in [Section 8](#). In [Section 8.1](#) we give the proof of exponential rates for identification using a subset oracle, in [Section 8.2](#) the proof of exponential rates for identification of finite collections using a membership oracle, and in [Section 8.3](#) the proof of exponential rates for identification of countable collections of finite languages. The proof for the identification rates with positive and negative examples appears in [Section 8.4](#).

5 Proofs from [Section 3.1](#) (Rates for Identification and Generation)

5.1 Proof of [Theorem 3.1](#) (Rates for Identification)

In this section, we give the full proof of [Theorem 3.1](#); see [Figure 2](#) for an outline. As we alluded to before, the first step in the proof is to show that all non-trivial collections are not learnable at rate faster than e^{-n} .

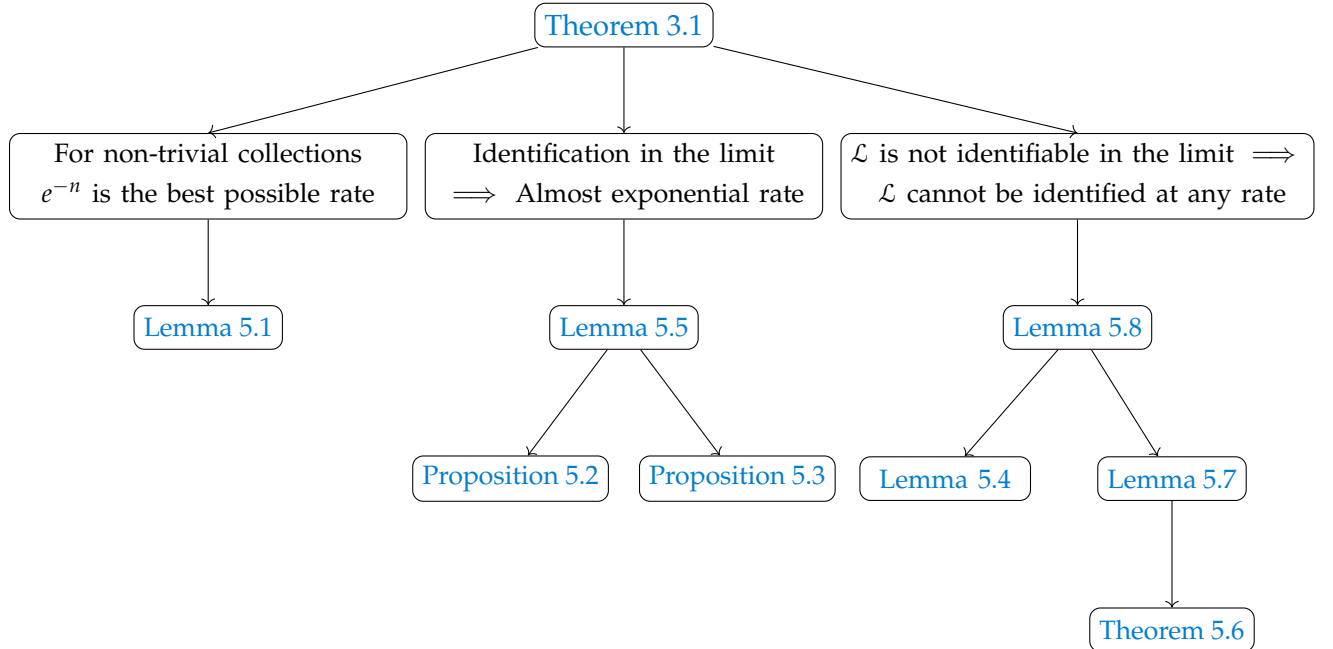


Figure 2: Outline of Proof of [Theorem 3.1](#)

Lemma 5.1 (Exponential Rate Is Best Possible for Identifying Any Non-trivial Collection). *Let \mathcal{L} be a non-trivial collection of countably many languages. Then, for any identification algorithm $\mathcal{A} = \{h_n\}_{n \in \mathbb{N}}$ there exists a valid distribution \mathcal{P} such that $\mathbb{E}[\text{er}(h_n)] \geq e^{-2n}$, for infinitely many $n \in \mathbb{N}$.*

Proof. Since \mathcal{L} is non-trivial, there exist two distinct languages $L_i, L_j \in \mathcal{L}$ and $x \in \mathcal{X}$ such that $x \in L_i, x \in L_j$. Let $\mathcal{P}_{L_i}, \mathcal{P}_{L_j}$ be valid distributions for L_i, L_j that place at least $1/2$ on x and if the languages have more elements, they spread the remaining mass on the rest of the elements arbitrarily; otherwise they put the remaining mass on x . Notice that since $L_i \neq L_j$ at least one of them has at least one more element other than x . For any $n \in \mathbb{N}$, under both distributions, with probability at least 2^{-n} the algorithm will only see the element x appearing in the samples. Let \mathcal{E}_n be that event and condition on it. Notice that

$$\Pr [L_{h_n(x, \dots, x)} = L_i \mid \mathcal{E}_n] + \Pr [L_{h_n(x, \dots, x)} = L_j \mid \mathcal{E}_n] \leq 1,$$

where the probability is with respect to the randomness of the identification algorithm. Thus, we have that either $\Pr [L_{h_n(x, \dots, x)} \neq L_i \mid \mathcal{E}_n] \geq 1/2$ or $\Pr [L_{h_n(x, \dots, x)} \neq L_j \mid \mathcal{E}_n] \geq 1/2$ for each $n \in \mathbb{N}$. Hence, by the pigeonhole principle, for at least one of L_i, L_j , the previous inequality holds for infinitely many $n \in \mathbb{N}$. Assume, without loss of generality, that it holds for L_i and let \hat{N} denote the set of $n \in \mathbb{N}$ for which it holds. Then, for each $n \in \hat{N}$, we have that

$$\begin{aligned} \mathbb{E}_{X_1, \dots, X_n \sim \mathcal{P}_{L_i}^n} [\text{er}(h_n(X_1, \dots, X_n))] &= \Pr_{X_1, \dots, X_n \sim \mathcal{P}_{L_i}^n} [L_{h_n(X_1, \dots, X_n)} \neq L_i] \\ &\geq \Pr_{X_1, \dots, X_n \sim \mathcal{P}_{L_i}^n} [L_{h_n(X_1, \dots, X_n)} \neq L_i \mid \mathcal{E}_n] \cdot \Pr_{X_1, \dots, X_n \sim \mathcal{P}_{L_i}^n} [\mathcal{E}_n] \\ &\geq \frac{1}{2^n} \cdot \Pr [L_{h_n(x, \dots, x)} \neq L_i \mid \mathcal{E}_n] \quad (\text{by the definition of } \mathcal{E}_n) \\ &\geq \frac{1}{2^{n+1}}, \quad (\text{due to the assumption on } L_i) \end{aligned}$$

which concludes the proof. \square

We now move on to the (almost) exponential rates upper bound for identification. This will be done via a transformation from learners that achieve identification in the limit in Gold's model [Gol67] to learners that achieve (almost) exponential rates in our setting. The first step in this result is to show that when we draw countably many samples from \mathcal{P} all the elements of the target language will appear in the sample.

Proposition 5.2 (Infinite Draws Are Enumerations). *Let \mathcal{P} be a probability distribution supported on a countable domain and $\{X_i\}_{i \in \mathbb{N}}$, where every X_i is i.i.d. from \mathcal{P} . Then,*

$$\Pr_{\{X_i\}_{i \in \mathbb{N}} \sim \mathcal{P}^\infty} [\text{supp}(\mathcal{P}) = \cup_{i \in \mathbb{N}} \{X_i\}] = 1.$$

Proof. For the direction $\Pr_{\{X_i\}_{i \in \mathbb{N}} \sim \mathcal{P}^\infty} [\text{supp}(\mathcal{P}) \supseteq \cup_{i \in \mathbb{N}} \{X_i\}]$ notice that for any element $x \notin \text{supp}(\mathcal{P})$

$$\Pr_{\{X_i\}_{i \in \mathbb{N}} \sim \mathcal{P}^\infty} [x \in \cup_{i \in \mathbb{N}} \{X_i\}] \leq \sum_{i \in \mathbb{N}} \Pr_{X_i \sim \mathcal{P}} [x = X_i] = 0. \quad (6)$$

Hence,

$$\begin{aligned} \Pr_{\{X_i\}_{i \in \mathbb{N}} \sim \mathcal{P}^\infty} [\text{supp}(\mathcal{P}) \supseteq \cup_{i \in \mathbb{N}} \{X_i\}] &= 1 - \Pr_{\{X_i\}_{i \in \mathbb{N}} \sim \mathcal{P}^\infty} [\exists x \notin \text{supp}(\mathcal{P}), x \in \cup_{i \in \mathbb{N}} \{X_i\}] \\ &\geq 1 - \sum_{x \notin \text{supp}(\mathcal{P})} \Pr_{\{X_i\}_{i \in \mathbb{N}} \sim \mathcal{P}^\infty} [x \in \cup_{i \in \mathbb{N}} \{X_i\}] \\ &\stackrel{(6)}{=} 1. \end{aligned}$$

For the other direction, *i.e.*, $\Pr_{\{X_i\}_{i \in \mathbb{N}} \sim \mathcal{P}^\infty}[\text{supp}(\mathcal{P}) \subseteq \cup_{i \in \mathbb{N}} \{X_i\}]$, notice that for any element $x \in \text{supp}(\mathcal{P})$, $\Pr_{X \sim \mathcal{P}}[X = x]$ is a positive constant $p_x > 0$ and let $(\mathcal{E}_n := \{X_n = x\})_{n \in \mathbb{N}}$ be a sequence of events. Notice that these events are independent and that

$$\sum_{n \in \mathbb{N}} \Pr[\mathcal{E}_n] = \sum_{n \in \mathbb{N}} p_x = \infty.$$

Hence, we can apply the second Borel-Cantelli lemma (see [Lemma E.2](#)) and get that

$$\Pr \left[\limsup_{n \rightarrow \infty} \mathcal{E}_n \right] = 1. \quad (7)$$

In other words, the element x will appear infinitely often in the stream X_1, \dots , with probability one. Therefore,

$$\begin{aligned} \Pr_{\{X_i\}_{i \in \mathbb{N}} \sim \mathcal{P}^\infty}[\text{supp}(\mathcal{P}) \subseteq \cup_{i \in \mathbb{N}} \{X_i\}] &= 1 - \Pr_{\{X_i\}_{i \in \mathbb{N}} \sim \mathcal{P}^\infty}[\exists x \in \text{supp}(\mathcal{P}) : x \notin \cup_{i \in \mathbb{N}} \{X_i\}] \\ &\geq 1 - \sum_{x \in \text{supp}(\mathcal{P})} \Pr_{\{X_i\}_{i \in \mathbb{N}} \sim \mathcal{P}^\infty}[x \notin \cup_{i \in \mathbb{N}} \{X_i\}] \\ &\stackrel{(7)}{=} 1. \end{aligned}$$

□

Next, we show that for any algorithm \mathcal{A} that identifies the target language in the limit in the adversarial (online) setting and for any valid distribution \mathcal{P} there is some number $t^* := t^*(\mathcal{A}, \mathcal{P}) \in \mathbb{N}$ such that, when we draw t^* many i.i.d. samples from \mathcal{P} and use them to simulate the adversarial game with \mathcal{A} , it will identify the target language with probability at least $6/7$. We denote the time of the last mistake of the algorithm $\mathcal{A} = \{h_n\}_{n \in \mathbb{N}}$ on a sequence x_1, x_2, \dots by $T_{\mathcal{A}}(x_1, x_2, \dots)$, *i.e.*,

$$T_{\mathcal{A}}(x_1, x_2, \dots) = \inf \left\{ n_0 \in \mathbb{N} : L_{h_n(x_1, \dots, x_n)} \neq K, \forall n \geq n_0 \right\}.$$

Proposition 5.3 (Tail Bound on the Distribution of Last Mistake). *Fix any countable collection of languages \mathcal{L} and let $K \in \mathcal{L}$ be the true language. For any algorithm $\mathcal{A} = \{h_n\}_{n \in \mathbb{N}}$ that identifies \mathcal{L} in the limit in the online setting from positive examples and any valid distribution \mathcal{P} for K ([Definition 1](#)), there exists a number $t^* \in \mathbb{N}$ such that*

$$\Pr_{\{X_i\}_{i \in \mathbb{N}} \sim \mathcal{P}^\infty} [T_{\mathcal{A}}(X_1, X_2, \dots) \leq t^*] \geq \frac{6}{7}.$$

Proof. Let X_1, X_2, \dots , be a countable i.i.d. sample from \mathcal{P} . From [Proposition 5.2](#) we get that this sample is a valid input to \mathcal{A} since, with probability one, it consists only of elements of K and eventually every element of K appears in this sequence. Consider the execution of \mathcal{A} on prefixes of the sequence and denote by $T_{\mathcal{A}} := T_{\mathcal{A}}(X_1, X_2, \dots)$ the time it made its last mistake. We have that $\Pr_{\{X_i\}_{i \in \mathbb{N}} \sim \mathcal{P}^\infty} [T_{\mathcal{A}} \in \mathbb{N}] = 1$. Thus,

$$\lim_{t \rightarrow \infty} \Pr_{\{X_i\}_{i \in \mathbb{N}} \sim \mathcal{P}^\infty} [T_{\mathcal{A}}(X_1, X_2, \dots) \geq t] = 0.$$

Thus, as required, there exists some $t^* \in \mathbb{N}$ such that

$$\Pr_{\{X_i\}_{i \in \mathbb{N}} \sim \mathcal{P}^\infty} [T_{\mathcal{A}}(X_1, X_2, \dots) \geq t^*] \leq \frac{1}{7}.$$

□

Thus far we have shown that for every valid distribution \mathcal{P} there exists some number $t^* \in \mathbb{N}$ so that if we simulate the online learning process with t^* samples i.i.d. from \mathcal{P} , then the algorithm identifies the true language K correctly with probability at least $6/7$. However, the number t^* depends on the distribution \mathcal{P} , and hence we cannot immediately devise a learning strategy based on it. To make the exposition easier to follow, let us first assume that we do know t^* ; we will shortly relax this assumption. For n sufficiently large consider the following algorithm:

- We split the input sequence into n/t^* non-overlapping batches, where the i -th batch consists of the elements $X_{(i-1) \cdot t^* + 1}, \dots, X_{i \cdot t^*}$.
- We use each of these sequences as an input to a copy of \mathcal{A} and we get n/t^* many predictors $\left\{ h_n^i \left(X_{(i-1) \cdot t^* + 1}, \dots, X_{i \cdot t^*} \right) \right\}_{i \in [n/t^*]}$.
- Since these predictors might be outputting different indices (descriptions) of the same language, we find the smallest indexed language the output of each classifier can be mapped to. In other words, if j_i is the index outputted by the i -th batch, we find the smallest number $j' \in \mathbb{N}$ such that $L_{j_i} = L_{j'}$, and we set $j_i := j'$. Since we only have query access to the languages, we can only approximate this step. In particular, for every $n \in \mathbb{N}$, we set $j_i := j'$ if $j' \in \mathbb{N}$ is the smallest number for which $\mathbb{1}\{x_\ell \in L_{j_i}\} = \mathbb{1}\{x_\ell \in L_{j'}\}$, for all $\ell \in [n]$. The details are handled in [Lemma 5.4](#).
- We predict the index that at least $(5/7) \cdot (n/t^*)$ of the predictors agree upon; if no such language exists we output one arbitrarily.

Before moving to the general case where t^* is unknown, it is instructive to explain why the previous approach achieves exponential rates. Using standard concentration bounds, it is not hard to see that with probability at least $1 - c \cdot e^{-C \cdot n}$, where c and C are \mathcal{P} -dependent constants, at least a $5/7$ fraction of the predictors will output an index that describes the true language. Conditioned on that event, it is immediate that a $5/7$ -majority is well-defined and predicting based on it yields the correct answer.

Let us now explain how to handle the actual problem setting, in which as we mentioned, we do not have knowledge of t^* . Let $f: \mathbb{N} \rightarrow \mathbb{N}$ be some (very slowly) increasing function of the input size n , which we will specify shortly. Given that function, we use the following modified approach, where t^* is replaced by $f(n)$.

- We split the input sequence into $n/f(n)$ non-overlapping batches, where the i -th batch consists of the elements $X_{(i-1) \cdot f(n) + 1}, \dots, X_{i \cdot f(n)}$.
- We use each of these sequences as an input to a copy of \mathcal{A} and we get $n/f(n)$ many predictors $\left\{ h_n^i \left(X_{(i-1) \cdot f(n) + 1}, \dots, X_{i \cdot f(n)} \right) \right\}_{i \in [n/f(n)]}$.
- We use the post-processing approach from [Lemma 5.4](#), that we also explained above.
- We predict the index that at least $(5/7) \cdot (n/f(n))$ of the predictors agree upon; if no such language exists we output one arbitrarily.

Since $f(\cdot)$ is increasing, there is some $n_0 \in \mathbb{N}$ such that $f(n_0) = t^*$. Thus, for $n \geq n_0$ we can repeat the previous argument; with probability at least $1 - c \cdot e^{-C \cdot n/f(n)}$, at least $(5/7) \cdot (n/f(n))$ of the predictors will be outputting the correct target language, so taking the majority vote over them yields the desired result. Notice that now we do not achieve exactly exponential rates, but for every sublinear function $g(\cdot)$ we can achieve rates $e^{-g(n)}$.

We first state and prove the post-processing lemma to map the outputs of predictors that correspond to different indices of the target language K to the same index. For this result, it is useful to define the notion of “projection” of a language onto a subset of \mathcal{X} .

Definition 16 (*m*-Projection of a Language). *Let $\mathcal{X} = \{x_1, x_2, \dots\}$ be a countable domain and let $L \subseteq \mathcal{X}$ be a language. For any $m \in \mathbb{N}$ we denote by $L[m] := L \cap \{x_1, x_2, \dots, x_m\}$ the projection of the language onto the first m elements of the domain.*

The point of the next lemma is the following: in the enumeration of the language collection \mathcal{L} , we allow repetitions of K (as in Gold’s model). Hence, when running multiple copies of our identification algorithms, the majority of them will identify K ; yet we cannot guarantee that they will identify the *same* index for K (due to multiple appearances of K in the enumeration). This lemma guarantees that there exists a sufficiently large prefix of the enumeration of the domain \mathcal{X} so that the projection of predicted languages will be mapped to the smallest index version of K in \mathcal{L} , which we denote by L_z below.

Lemma 5.4 (Post-processing to Map to Lowest-Index Occurrence of K). *Let $\mathcal{L} = \{L_1, L_2, \dots\}$ be a countable collection of languages over $\mathcal{X} = \{x_1, x_2, \dots\}$ and $K \in \mathcal{L}$. Let $z := \min\{j \in \mathbb{N} : L_j = K\}$ be the first index at which the target language appears in \mathcal{L} . Let $\mathcal{J} = (i_1, \dots, i_m)$ be a multiset of indices and for all $1 \leq \ell \leq \max_{j \in [m]} i_j$, $n \in \mathbb{N}$, let*

$$\hat{i}_j^n = \min\{1 \leq \ell \leq i_j : L_\ell[n] = L_{i_j}[n]\},$$

be the index of the first language that has the same projection as L_{i_j} . Then, there exists a number $n_0 := n_0(K, \mathcal{L}, \mathcal{X})$ that depends on $K, \mathcal{L}, \mathcal{X}$, but not \mathcal{J} , such that for all $n \geq n_0$

$$L_{i_j} = K \implies \hat{i}_j^n = z, \forall j \in [m].$$

Before we give the formal proof, let us explain the main idea and the implication of this result. For every language L that precedes L_z , there exists some element $x \in \mathcal{X}$ such that $\mathbb{1}\{x \in L\} \neq \mathbb{1}\{x \in L_z\}$. This will enable us to detect and remove all languages different from K preceding L_z . Then, by taking projections onto large enough prefixes we indeed map any $L_j = K, j > z$, to L_z . This result will be useful for our constructions that require aggregating outputs from different executions of the algorithm which, without this post-processing step, can output different indices.

Proof of Lemma 5.4. Assume without loss of generality that for some $j \in [m]$ we have that $L_{i_j} = K$, otherwise the statement holds vacuously. We will handle the cases $z = 1, z > 1$ separately.

Case A ($z = 1$): For any $j \in [m]$ for which $L_{i_j} = K$ and any $n \in \mathbb{N}$ we have that $L_{i_j}[n] = L_z[n]$, and since $z = 1$ this is the first index for which the equality holds. Hence, in this case, the claim holds with $n_0 = 1$.

Case B ($z > 1$): Since L_z is the first occurrence of K in \mathcal{L} , for all languages L_ℓ with $1 \leq \ell < z$, there exists some $x \in \mathcal{X}$ such that $\mathbb{1}\{x \in L_j\} \neq \mathbb{1}\{x \in L_z\}$. Let x_{z_ℓ} be the smallest indexed element of \mathcal{X} for which the previous holds. Moreover, let $n_0 = \max_{1 \leq \ell < z} z_\ell$. Notice that for all $n \geq n_0$ and all $1 \leq j < z$, holds that $L_j[n] \neq L_z[n]$. Furthermore, for all $n \in \mathbb{N}$ and all $j \in [m]$ such that $L_{i_j} = K$ it holds that $L_{i_j}[n] = L_z[n]$. Combining these two claims, we can deduce that for all $j \in [m]$ such that $L_{i_j} = K$ and for all $n \geq \hat{n}_0$ the first index $i \in \mathbb{N}$ such that $L_{i_j}[n] = L_i[n]$ is indeed z . Notice that n_0 depends only on the enumeration of \mathcal{L}, \mathcal{X} , and the target language K . \square

We are now ready to state and prove the formal result regarding the identification rates of collections that are identifiable in the limit.

Lemma 5.5 (Reduction From Identification at Almost-Exponential Rate to Online Identification). *Let $\mathcal{L} = \{L_1, L_2, \dots\}$ be a countable collection of languages and $g: \mathbb{N} \rightarrow \mathbb{N}$ be a sublinear function. For any algorithm $\mathcal{A} = \{h_n\}_{n \in \mathbb{N}}$ that identifies \mathcal{L} in the limit in the online setting with positive examples and any valid distribution \mathcal{P} there exists an algorithm $\mathcal{A}' = \{h'_n\}_{n \in \mathbb{N}}$ such that for all $n \in \mathbb{N}$*

$$\mathbb{E}_{X_1, \dots, X_n \sim \mathcal{P}^n} [\text{er}(h'_n(X_1, \dots, X_n))] \leq c \cdot e^{-C \cdot g(n)}.$$

Proof. Let $\mathcal{K} := \{L_{i_1}, L_{i_2}, \dots\} \subseteq \mathcal{L}$ be the set of all languages in \mathcal{L} that correspond to representations of K , i.e., for all $L \in \mathcal{K}$ it holds that $L = K$. First, notice that since $g(n) = o(n)$ we can construct some non-decreasing function $f: \mathbb{N} \rightarrow \mathbb{N}$ with $\lim_{n \rightarrow \infty} f(n) = \infty$ and $n/f(n) \geq g(n)$. Let $t^* \in \mathbb{N}$ be a number such that

$$\Pr_{X_1, \dots, X_{t^*} \sim \mathcal{P}^{t^*}} [L_{h_{t^*}}(X_1, \dots, X_{t^*}) \in \mathcal{K}] \geq \frac{6}{7}.$$

From [Proposition 5.3](#) such a number t^* is guaranteed to exist. Hence, there is some n_0 such that for all $n \geq n_0$, $f(n) \geq t^*$. Thus, for all $n \geq n_0$

$$\Pr_{X_1, X_2, \dots, X_{f(n)} \sim \mathcal{P}^{f(n)}} [L_{h_{f(n)}}(X_1, \dots, X_{f(n)}) \in \mathcal{K}] \geq \frac{6}{7}.$$

Recall the error of the classifier $h_{f(n)}(\cdot)$ as defined in [Equation \(4\)](#). We have that for all $n \geq n_0$, it holds that $\text{er}(h_{f(n)})$ is a Bernoulli random variable with $p \leq 1/7$. Thus, if we have a collection of $\hat{t}_n := n/f(n)$ such i.i.d. random variables, using Hoeffding's bound [[DP09](#)] we get that

$$\Pr_{X_1, \dots, X_n} \left[\frac{1}{\hat{t}_n} \sum_{i=1}^{\hat{t}_n} \text{er} \left(h_{f(n)} \left(X_{(i-1) \cdot \hat{t}_n + 1}, \dots, X_{i \cdot \hat{t}_n} \right) \right) \geq \frac{2}{7} \right] \leq e^{-2\hat{t}_n/49} \leq e^{-2g(n)/49}.$$

Thus, for $n \geq n_0$, at least a $5/7$ -fraction of the predictors outputs an index that corresponds to the target language. We condition on that event \mathcal{E}_0^n for the rest of the proof.

Let $\mathcal{J}^n = (i_1, \dots, i_{\hat{t}_n})$ be the multiset of the indices of the languages outputted by the predictors in the previous step and $z = \min\{\ell \in \mathbb{N}: L_\ell = K\}$. Then, using [Lemma 5.4](#) we know that there exists some $\hat{n}_0 := \hat{n}_0(K, \mathcal{L}, \mathcal{X})$ such that for all $n \geq \hat{n}_0$

$$L_{i_j} = K \implies \hat{i}_j^n = z, \forall j \in [\hat{t}_n],$$

where \hat{i}_j^n is defined in [Lemma 5.4](#). Thus, letting $\hat{\mathcal{I}}^n = (\hat{i}_1^n, \dots, \hat{i}_{t_n}^n)$ we have that for all $n \geq \max\{n_0, \hat{n}_0\}$, at least a $5/7$ -fraction of the indices in $\hat{\mathcal{I}}^n$ are exactly the index z .

Thus, for all $n \geq \max\{n_0, \hat{n}_0\}$ and conditioned on the event \mathcal{E}_0^n , we have that the $5/7$ -majority vote over the indices in $\hat{\mathcal{I}}^n$ corresponds to the first occurrence of K in \mathcal{L} . Since \mathcal{E}_0^n occurs with probability at least $1 - e^{-2g(n)/49}$, this concludes the proof. \square

We now move on to the final ingredient we require for the proof of [Theorem 3.1](#). It remains to show that Angluin's condition characterizes the collections of languages that can be identified at an (almost) exponential rate. First, we discuss a result from Angluin [[Ang88](#)] which our proof builds upon. Let us first briefly describe the convergence criterion in Angluin's paper. There is a valid distribution \mathcal{P} that is supported over some $K \in \mathcal{L}$ and the algorithm is presented with an infinite sequence of i.i.d. draws from \mathcal{P} . After seeing each example, the learner must output some $i \in \mathbb{N}$ with the goal being that $L_i = K$. In that setting, an algorithm learns the target language if for all but finitely many $n \in \mathbb{N}$ it outputs the same index $i \in \mathbb{N}$ for which $L_i = K$. Notice that the learning requirement is two-fold: **i)** the learner needs to stabilize, and **ii)** the index it predicts needs to correspond to the target language. In that setting, Angluin [[Ang88](#)] showed the following result.

Theorem 5.6 (Corollary 10 [[Ang88](#)]). *Let \mathcal{L} be a countable collection of languages over a countable domain \mathcal{X} that does not satisfy Angluin's condition ([Definition 10](#)). Then, for every learning algorithm $\mathcal{A} = \{h_n\}_{n \in \mathbb{N}}$ there exists a valid distribution \mathcal{P} supported on some $K \in \mathcal{L}$ such that, with probability at least $1/3$ over the i.i.d. draw of $\{X_n\}_{n \in \mathbb{N}}$, the learner does not identify K .*

In particular, Angluin's result shows that, with probability at least $1/3$, the learner will either not stabilize to any number $i \in \mathbb{N}$ or it will stabilize to a number $j \in \mathbb{N}$ with $L_j \neq K$. Our next result provides a strengthening of Angluin's result since it shows that with probability at least $1/3$, the learner will, in fact, predict infinitely many times indices that do not correspond to K .

Lemma 5.7. *Let \mathcal{L} be a countable collection of languages over a countable domain \mathcal{X} that does not satisfy Angluin's condition ([Definition 10](#)). Then, for every learning algorithm $\mathcal{A} = \{h_n\}_{n \in \mathbb{N}}$ there exists a valid distribution \mathcal{P} supported on some $K \in \mathcal{L}$ such that*

$$\Pr_{\{X_i\}_{i \in \mathbb{N}} \sim \mathcal{P}^\infty} \left[\exists i_1 < i_2 < i_3 < \dots : L_{h_{i_j}(X_1, \dots, X_{i_j})} \neq K, \forall j \in \mathbb{N} \right] \geq \frac{1}{3}.$$

Proof. Let \mathcal{L} be a countable collection of languages that does not satisfy Angluin's condition. Assume towards contradiction that there is learner $\{h_n\}_{n \in \mathbb{N}}$ that, for any valid distribution \mathcal{P} , with probability $c_{\mathcal{P}} < 1/3$ misidentifies K infinitely often, i.e., for any valid distribution \mathcal{P}

$$\Pr_{\{X_i\}_{i \in \mathbb{N}} \sim \mathcal{P}^\infty} \left[\exists i_1 < i_2 < i_3 < \dots : L_{h_{i_j}(X_1, \dots, X_{i_j})} \neq K, \forall j \in \mathbb{N} \right] = c_{\mathcal{P}} < \frac{1}{3}. \quad (8)$$

We will construct a different learner $\{h'_n\}$ which, for all valid distributions \mathcal{P} , learns the corresponding target language K in Angluin's setting [[Ang88](#)] with probability at least $2/3$. This will create the desired contradiction with [Theorem 5.6](#). Let h'_n be a learner that works as follows.

Learner h'_n

Input: Access to any infinite draw X_1, X_2, \dots from \mathcal{P}^∞ and oracle access to learner $h(\cdot)$

Description:

1. **For each** $n \in \mathbb{N}$ **do:**

(a) Compute $i^* = h_n(X_1, \dots, X_n)$

(b) Compute $h'_n(X_1, \dots, X_n)$ as the smallest index j^* such that L_{j^*} classifies the first n elements x_1, x_2, \dots, x_n of the domain \mathcal{X} in the same way as L_{i^*} , i.e.,

$$h'_n(X_1, \dots, X_n) := \min \{j \in [i^*] : \mathbb{1}\{x_i \in L_j\} = \mathbb{1}\{x_i \in L_{i^*}\}, \forall i \in [n]\}.$$

Notice that this can be done with $O(n \cdot i^*)$ many membership queries; where we send n queries to each of the first i^* languages in \mathcal{L} .

Let $z \in \mathbb{N}$ be the smallest number for which $L_z = K$. We will handle the cases $z = 1$ and $z > 1$ separately.

Case A ($z = 1$): If $L_{h(X_1, \dots, X_n)} = K$, we have that $\mathbb{1}\{x \in L_{h(X_1, \dots, X_n)}\} = \mathbb{1}\{x \in L_1\}$ for all $x \in \mathcal{X}$. Hence, since 1 is the smallest index, we have that for all $n \in \mathbb{N}$

$$L_{h_n(X_1, \dots, X_n)} = K \implies h'_n(X_1, \dots, X_n) = z.$$

In this case, we define $n_0 := 1$.

Case B ($z > 1$): Let $1 \leq z' < z$. Then, since L_z is the first language for which $L_z = K$ it must be the case that $L_{z'} \neq L_z$. Hence, the set $S_{z'} := \{x \in \mathcal{X} : \mathbb{1}\{x \in L_{z'}\} \neq \mathbb{1}\{x \in L_z\}\}$ is non-empty. We let $\ell_{z'} := \min \{i \in \mathbb{N} : x_i \in S_{z'}\}$, i.e., let $\ell_{z'}$ be the smallest number that $x_{\ell_{z'}}$ certifies that $L_{z'}$ is different from L_z . Notice that $\ell_{z'} < \infty$. We also define $n_0 := \max \{\ell_1, \dots, \ell_{z-1}\}$, Notice that for all $n \geq n_0$ we have that

$$L_{h_n(X_1, \dots, X_n)} = K \implies h'_n(X_1, \dots, X_n) = z.$$

Let \mathcal{E} denote the event that

$$|\{n \in \mathbb{N} : h_n(X_1, \dots, X_n) \neq K\}| < \infty.$$

Notice that, by definition of $c_{\mathcal{P}}$,

$$\Pr_{\{X_i\}_{i \in \mathbb{N}} \sim \mathcal{P}^\infty} [\mathcal{E}] = 1 - c_{\mathcal{P}}.$$

In other words, conditioned on the event \mathcal{E} , the learner $\{h_n\}_{n \in \mathbb{N}}$ makes finitely many mistakes. Thus, conditioned on \mathcal{E} , for every draw $D := \{X_i\}_{i \in \mathbb{N}} \sim \mathcal{P}^\infty$ there is some number $n_D \in \mathbb{N}$ such that $h_n(X_1, \dots, X_n) = K, \forall n \geq n_D$. Let $n' = \max \{n_D, n_0\}$. Then, conditioned on \mathcal{E} , for every $n \geq n'$ we have that

$$h'_n(X_1, \dots, X_n) = z.$$

Since $c_{\mathcal{P}} < 1/3$ (see [Equation \(8\)](#)), $1 - c_{\mathcal{P}} > 2/3$ which implies that the learner $\{h'_n\}_{n \in \mathbb{N}}$ converges in Angluin's model with probability greater than $2/3$, for any choice of a valid data-generating distribution \mathcal{P} . This contradicts [Theorem 5.6](#) and concludes the proof. \square

Let us now explain why [Lemma 5.7](#) does not immediately imply a lower bound in our setting. First, notice that the previous result says that any learner $\{h_n\}_{n \in \mathbb{N}}$ must make infinitely many errors with probability at least $1/3$, under some valid data-generating distribution \mathcal{P} . Expressing this using a $\limsup(\cdot)$ implies that¹⁴

$$\Pr_{\{X_i\}_{i \in \mathbb{N}} \sim \mathcal{P}^\infty} \left[\limsup_{n \rightarrow \infty} \{L_{h_n}(X_1, \dots, X_n) \neq K\} \right] \geq \frac{1}{3}.$$

Since $R: \mathbb{N} \rightarrow \mathbb{R}_{\geq 0}$ is a rate function, it satisfies $\lim_{n \rightarrow \infty} R(n) = 0$. Thus, in order to show that \mathcal{L} is not learnable at any rate, it is enough to show that for any learner $\{h_n\}_{n \in \mathbb{N}}$, there exists a valid distribution \mathcal{P} such that $\Pr_{X_1, \dots, X_n \sim \mathcal{P}^n} [L_{h_n}(X_1, \dots, X_n) \neq K]$ does not converge as $n \rightarrow \infty$, or if it does converge it holds that

$$\lim_{n \rightarrow \infty} \Pr_{X_1, \dots, X_n \sim \mathcal{P}^n} [L_{h_n}(X_1, \dots, X_n) \neq K] \neq 0.$$

For this, it suffices to show that

$$\limsup_{n \rightarrow \infty} \Pr_{X_1, \dots, X_n \sim \mathcal{P}^n} [L_{h_n}(X_1, \dots, X_n) \neq K] > 0, \quad (9)$$

for some valid distribution \mathcal{P} . It follows from the reverse of Fatou's lemma that for every sequence of events $\{\mathcal{E}_n\}_{n \in \mathbb{N}}$

$$\Pr \left[\limsup_{n \rightarrow \infty} \mathcal{E}_n \right] \geq \limsup_{n \rightarrow \infty} \Pr[\mathcal{E}_n],$$

which is not sufficient to deduce the result we need. In fact, it is not hard to construct a family of events such that $\{\mathcal{E}_n\}_{n \in \mathbb{N}}$ such that $\Pr[\limsup_{n \rightarrow \infty} \mathcal{E}_n] = 1$, but $\limsup_{n \rightarrow \infty} \Pr[\mathcal{E}_n] = 0$: consider an infinite stream of *independent* coin flips, where the probability of success of the n -th try is $1/n$. The second Borel-Cantelli lemma (see [Lemma E.2](#)) implies the result. Hence, we need to study the particular structure of our problem to show that $\limsup_{n \rightarrow \infty} \Pr[\mathcal{E}_n] > 0$.

In fact, to deduce that $\limsup_{n \rightarrow \infty} \Pr[\mathcal{E}_n] > 0$, we show a stronger result: the \limsup of the probability of error of the learner is not merely bounded away from zero, but, it is at least $1/2$ ([Lemma 5.8](#)). To that end, we follow a strategy which consists of the following two main steps:

- First, we assume that there exists a learner $\{h_n\}_{n \in \mathbb{N}}$ such that for every valid distribution \mathcal{P} there is some $c > 0$ such that

$$\limsup_{n \rightarrow \infty} \Pr_{X_1, \dots, X_n \sim \mathcal{P}^n} \left[\left\{ L_{h_n}(X_1, \dots, X_n) \neq K \right\} \right] \leq \frac{1}{2} - c.$$

Then, we show that using the learner $\{h_n\}_{n \in \mathbb{N}}$ we can construct a learner $\{h'_n\}_{n \in \mathbb{N}}$ such that for all valid distributions $\mathbb{E}_{X_1, \dots, X_n \sim \mathcal{P}^n} [\mathbb{1}\{L_{h'_n}(X_1, \dots, X_n) \neq K\}] \leq C \cdot e^{-c \cdot n / \log n}$, where c, C are distribution-dependent constants. This can be viewed as a boosting argument for identification. To make this argument work, we also need to use our post-processing result ([Lemma 5.4](#)) to map different outputs that correspond to K to the same index.

¹⁴Informally, \limsup of a sequence of events captures the events that occur infinitely often. For instance, $\Pr[\limsup_{n \rightarrow \infty} \mathcal{E}_n]$ represents the probability that infinitely many of the events \mathcal{E}_n occur. On the other hand, $\limsup_{n \rightarrow \infty} \Pr[\mathcal{E}_n]$ roughly speaking denotes the largest value that the probabilities $\Pr[\mathcal{E}_1], \Pr[\mathcal{E}_2], \dots$ approach infinitely often as $n \rightarrow \infty$.

- Subsequently, using the Borel-Cantelli lemma (see [Lemma E.1](#)) we show that for $\{h'_n\}_{n \in \mathbb{N}}$ it holds that for any valid distribution \mathcal{P}

$$\Pr_{\{X_i\}_{i \in \mathbb{N}} \sim \mathcal{P}^\infty} \left[\limsup_{n \rightarrow \infty} \left\{ L_{h'_n}(X_1, \dots, X_n) \neq K \right\} \right] = 0,$$

which, combined with [Lemma 5.7](#), leads to a contradiction.

The formal statement and the proof of the result follow.

Lemma 5.8. *For every countable collection of languages \mathcal{L} that does not satisfy Angluin's condition, and every learning algorithm $\mathcal{A} = \{h_n\}_{n \in \mathbb{N}}$ there exists a valid distribution \mathcal{P} supported on $K \in \mathcal{L}$ such that*

$$\limsup_{n \rightarrow \infty} \Pr_{X_1, \dots, X_n \sim \mathcal{P}^n} \left[L_{h_n}(X_1, \dots, X_n) \neq K \right] \geq \frac{1}{2}.$$

Proof. Assume towards contradiction that there exists a countable collection of languages \mathcal{L} that does not satisfy Angluin's condition and a learning algorithm $\mathcal{A} = \{h_n\}_{n \in \mathbb{N}}$ such that for all target languages $K \in \mathcal{L}$ and for all valid distributions \mathcal{P} supported over K there exists some $c < 1/2$ such that

$$\limsup_{n \rightarrow \infty} \Pr_{X_1, \dots, X_n \sim \mathcal{P}^n} \left[L_{h_n}(X_1, \dots, X_n) \neq K \right] = c.$$

Let also $\tilde{c} := 1/2 - c > 0$. By definition of the limit superior, it holds that

$$\left| \left\{ n \in \mathbb{N} : \Pr_{X_1, \dots, X_n \sim \mathcal{P}^n} [L_{h_n}(X_1, \dots, X_n) \neq K] > c + \frac{\tilde{c}}{2} \right\} \right| < \infty.$$

For the rest of the proof, let us fix some valid distribution \mathcal{P} . Let n_0 be the largest number such that $\Pr_{X_1, \dots, X_n \sim \mathcal{P}^n} [L_{h_n}(X_1, \dots, X_n) \neq K] > c + (\tilde{c}/2)$. Notice that n_0 depends on \mathcal{P} . The previous argument shows that $n_0 < \infty$. For all $n > n_0$ we have that

$$\Pr_{X_1, \dots, X_n \sim \mathcal{P}^n} [L_{h_n}(X_1, \dots, X_n) \neq K] \leq c + \frac{\tilde{c}}{2} = \frac{1}{2} - \frac{\tilde{c}}{2}.$$

Consider the algorithm $\mathcal{A}' = \{h'_n\}_{n \in \mathbb{N}}$ that works as follows: for every n , it splits the dataset into $\hat{t}_n := n/\log n$ consecutive and non-overlapping batches, each of size $\log n$. Then, it runs algorithm $h_{\log n}$ on each of the batches. Let i_j denote the output of the j -th batch, $\mathcal{J}^n = (i_1, \dots, i_{\hat{t}_n})$ denote the multiset of all these indices and $z = \min\{\ell \in \mathbb{N} : L_\ell = K\}$. Then, using [Lemma 5.4](#) we know that there exists some $\hat{n}_0 := \hat{n}_0(K, \mathcal{L}, \mathcal{X})$ such that for all $n \geq \hat{n}_0$

$$L_{i_j} = K \implies \hat{i}_j^n = z, \quad \forall j \in [\hat{t}_n],$$

where \hat{i}_j^n is defined in [Lemma 5.4](#). Thus, letting $\hat{\mathcal{J}}^n = (\hat{i}_1^n, \dots, \hat{i}_{\hat{t}_n}^n)$ we have that for all $n \geq \max\{n_0, \hat{n}_0\}$, all the indices of \mathcal{J}^n that correspond to some index of K are mapped to z in the collection $\hat{\mathcal{J}}^n$.

Finally, the algorithm outputs the majority vote over the indices in $\hat{\mathcal{J}}^n$. Using a standard Chernoff bound, we see that

$$\Pr \left[\sum_{j \in [\hat{t}_n]} \frac{\mathbb{1}_{\{\hat{i}_j^n \neq z\}}}{n/\log n} \geq \frac{1}{2} - \frac{\tilde{c}}{4} \right] \leq e^{-\tilde{c}^2 2n/\log(n)}, \quad \forall n \geq \max\{n_0, \hat{n}_0\}.$$

This implies that

$$\Pr_{X_1, \dots, X_n} [h'_n(X_1, \dots, X_n) \neq z] \leq e^{-\tilde{c}^2 2n / \log(n)}, \quad \forall n \geq \max\{n_0, \hat{n}_0\}.$$

Thus, we have that

$$\begin{aligned} \sum_{n \in \mathbb{N}} \Pr_{X_1, \dots, X_n \sim \mathcal{P}^n} [L_{h'_n}(X_1, \dots, X_n) \neq K] &= \sum_{n \leq \max\{n_0, \hat{n}_0\}} \Pr_{X_1, \dots, X_n \sim \mathcal{P}^n} [L_{h'_n}(X_1, \dots, X_n) \neq K] \\ &+ \sum_{n > \max\{n_0, \hat{n}_0\}} \Pr_{X_1, \dots, X_n \sim \mathcal{P}^n} [L_{h'_n}(X_1, \dots, X_n) \neq K] \\ &\leq \max\{n_0, \hat{n}_0\} + \sum_{n > \max\{n_0, \hat{n}_0\}} \Pr_{X_1, \dots, X_n \sim \mathcal{P}^n} [L_{h'_n}(X_1, \dots, X_n) \neq K] \\ &\leq \max\{n_0, \hat{n}_0\} + \sum_{n > \max\{n_0, \hat{n}_0\}} \Pr_{X_1, \dots, X_n \sim \mathcal{P}^n} [h'_n(X_1, \dots, X_n) \neq z] \\ &\leq \max\{n_0, \hat{n}_0\} + \sum_{n > \max\{n_0, \hat{n}_0\}} e^{-\tilde{c}^2 2n / \log(n)} \\ &< \infty. \end{aligned}$$

Using the Borel-Cantelli lemma (see [Lemma E.1](#)), we get that

$$\Pr_{\{X_i\}_{i \in \mathbb{N}} \sim \mathcal{P}^\infty} \left[\limsup_{n \rightarrow \infty} \left\{ L_{h'_n}(X_1, \dots, X_n) \neq K \right\} \right] = 0.$$

Since this holds for all valid distributions \mathcal{P} , it contradicts [Lemma 5.7](#), which states that, for some valid \mathcal{P}' (which depends on $\{h'_n\}_{n \in \mathbb{N}}$), it holds that

$$\Pr_{\{X_i\}_{i \in \mathbb{N}} \sim \mathcal{P}'^\infty} \left[\limsup_{n \rightarrow \infty} \left\{ L_{h'_n}(X_1, \dots, X_n) \neq K \right\} \right] \geq \frac{1}{3}.$$

This concludes the proof. □

We now have all the components to prove [Theorem 3.1](#) by following the outline in [Figure 2](#).

Proof of Theorem 3.1. Let \mathcal{L} be any non-trivial collection of languages. Then, [Lemma 5.1](#) implies that no learner can learn \mathcal{L} at a rate faster than e^{-n} .

Let us first consider the case that \mathcal{L} satisfies Angluin's condition. This implies that \mathcal{L} is identifiable in the limit (see [Theorem 2.2](#)). Let $g: \mathbb{N} \rightarrow \mathbb{R}$, be some sublinear function, i.e., $g(n) = o(n)$. Then, [Lemma 5.5](#) shows that there exists a learner that achieves rates $e^{-g(n)}$ for \mathcal{L} .

Lastly, we consider the case where \mathcal{L} does not satisfy Angluin's condition. Then, [Lemma 5.8](#) shows that for every learner $\{h_n\}_{n \in \mathbb{N}}$, there exists a valid distribution \mathcal{P} for which

$$\limsup_{n \rightarrow \infty} \Pr_{X_1, \dots, X_n \sim \mathcal{P}^n} [L_{h_n}(X_1, \dots, X_n) \neq K] \geq \frac{1}{2}.$$

Hence, \mathcal{L} is not learnable at any rate. This concludes the proof. □

5.2 Proof of Theorem 3.2 (Rates for Generation)

In this section, we prove Theorem 3.2 following the outline in Figure 3.

First, in Section 5.2.1 we formally define the family of collections that are non-trivial for generation (Definition 17) and show that (1) for any trivial collection, it is possible to generate after seeing a *constant* number of examples (Lemma 5.9) and (2) an exponential rate is the best-possible for any non-trivial collection (Lemma 5.10). Next, in Section 5.2.2, we present a sufficient condition under which a generation algorithm that works “in-the-limit” achieves exponential rate (without any modifications). Finally, in Sections 5.2.3 and 5.2.4, we present algorithms that achieve exponential rates given access to a subset oracle and membership oracle for the collection \mathcal{L} respectively.

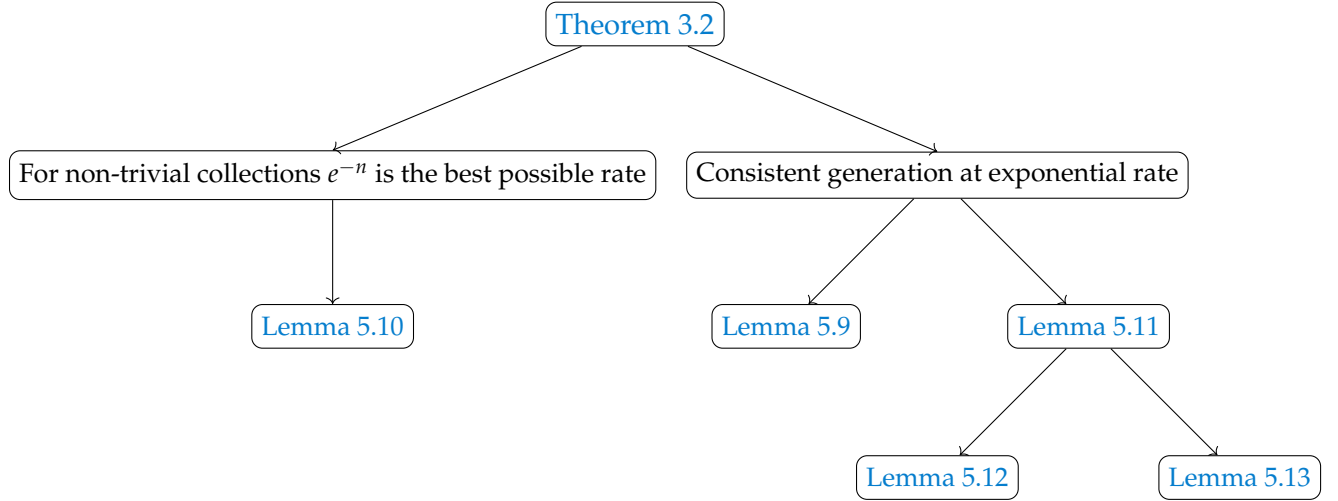


Figure 3: Outline of Proof of Theorem 3.2

5.2.1 Optimal Rate for Non-Trivial Collections for Generation

For the case of generation, we need a different notion of non-trivial languages from the one we did for identification (see Definition 13). Indeed, assume that $\mathcal{L} = \{L_1, L_2\}$, and $L_1 \neq L_2, L_1 \cap L_2 = \emptyset$. This collection satisfies Definition 13, thus no algorithm can identify at a rate faster than exponential. However, it is not hard to see there is a consistent generation algorithm that does not need *any* samples: just generate a string from $L_1 \cap L_2$. Instead, the following condition turns out to characterize collections that are non-trivial for generation.

Definition 17 (Non-trivial for Generation). *A language collection \mathcal{L} is non-trivial for generation if there exists some $x \in \mathcal{X}$ and a finite set of languages $\mathcal{L}' \subseteq \mathcal{L}$ such that:*

- *each $L \in \mathcal{L}'$ contains x ; and*
- *the intersection of all languages in \mathcal{L}' is finite, i.e., $|\cap_{L \in \mathcal{L}'} L| < \infty$.*

To verify that the definition of non-trivial collections is meaningful, we show in the following result that for all trivial collections, there exists an algorithm that generates correctly with probability 1, for all valid distributions, when n is sufficiently large even if the training dataset contains

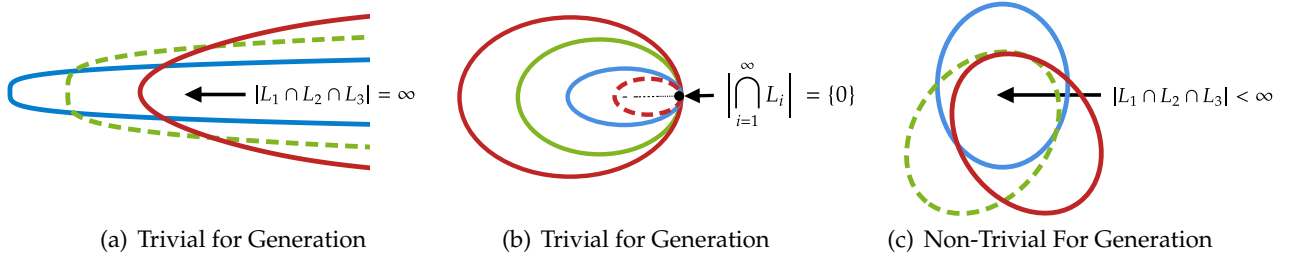


Figure 4: Illustrations of language collections that are (a,b) trivial for generation and (c) non-trivial for generation. In cases (a) and (c), the collection \mathcal{L} has three languages – L_1 , L_2 , and L_3 – denoted by different colors. Case (b) illustrates the example in [Example 2](#); here, the collection \mathcal{L} has infinitely many languages which follow a nested structure $L_1 \supsetneq L_2 \supsetneq \dots \supsetneq \{0\}$.

only one distinct element. Interestingly, our result uses an algorithm that generates in the limit in a setting that is slightly different from the one considered by Kleinberg and Mullainathan [KM24]: we fix some target language $K \in \mathcal{L}$, we give a single input $x \in K$ to the algorithm and then we run it for infinitely many steps, without giving any further inputs. We show that there exists some $n_{K,\mathcal{L}} \in \mathbb{N}$ which depends only on K and the enumeration of \mathcal{L} , but, crucially, not on x , such that the algorithm generates correctly for every $n \geq n_{K,\mathcal{L}}$. The algorithm is described below.

Generating in the limit from a trivial collection $\mathcal{L} = \{L_1, L_2, \dots\}$

Input: A set of n elements $\{X_1, \dots, X_n\}$ from \mathcal{X} , potentially containing repetitions

Description:

1. Select any arbitrary element x from $\{X_1, \dots, X_n\}$
2. Initialize the index $j = 1$
3. **while** $x \notin L_j$ **do:** increment j by 1
When X_1, \dots, X_n are drawn from a valid distribution this step terminates with probability 1
4. Compute $V^n(x) := \{L_i \in \mathcal{L} : x \in L_i, 1 \leq i \leq n\} \cup \{L_j\}$
5. Let $k := 1$
6. **while** $x_k \notin \bigcap_{L \in V^n(x)} L \setminus \{X_1, \dots, X_n\}$ **do:** increment k by 1
When \mathcal{L} is trivial for generation (see [Definition 17](#)) this step terminates with probability 1
7. **return** x_k

Notice that the previous algorithm is indeed computable given access to a membership oracle for each language in \mathcal{L} .

Lemma 5.9 (Algorithm For Generation from Trivial Collections). *For every collection of languages \mathcal{L} that is trivial for generation, there exists a generation algorithm $(\mathcal{G}_n)_{n \in \mathbb{N}}$ such that for every valid distribution \mathcal{P} with respect to \mathcal{L} it terminates with probability 1 for all $n \in \mathbb{N}$, and there exists some constant C that depends on \mathcal{P}, \mathcal{L} , such that for all $n \geq C$, it holds that*

$$\mathbb{E}_{X_1, \dots, X_n \sim \mathcal{P}^n} [\mathbb{1} \{ \mathcal{G}_n(X_1, \dots, X_n) \notin \text{supp}(\mathcal{P}) \setminus \{X_1, \dots, X_n\} \}] = 0.$$

Proof. Let \mathcal{L} be a trivial collection for generation (see [Definition 17](#)). Fix some valid distribution \mathcal{P} supported over target language K . Let $z \in \mathbb{N}$ be the smallest number such that $L_z = K$. Then, the triviality condition states that for every $x \in \mathcal{X}$ and every finite set of languages $\mathcal{L}' \subseteq \mathcal{L}$ it holds that either $x \notin L$, for some $L \in \mathcal{L}'$, or $|\cap_{L \in \mathcal{L}'} L| = \infty$. We have that, with probability 1, every i.i.d. draw of n samples from \mathcal{P} satisfies $X_1, \dots, X_n \in L_z$. Choose an arbitrary $x \in \{X_1, \dots, X_n\}$. Notice again that, with probability 1, $x \in K$. Consider the execution of the algorithm described above by fixing this x . Let us now verify that this algorithm generates some $x' \in K \setminus \{X_1, \dots, X_n\}$ for all $n \geq z$, and terminates with probability 1 for all $n \in \mathbb{N}$. First, notice that by definition of z , when $n \geq z$ it holds that $L_z \in V^n(x)$. Notice that, since for all $L \in V^n(x)$ it holds that $x \in L$ and $|V^n(x)| \leq n < \infty$, the triviality definition implies that for all $n \geq z$,

$$\left| \bigcap_{L \in V^n(x)} L \right| = \infty.$$

Hence, for all $n \geq z$ we have that

- $K \in V^n(x)$, thus $\bigcap_{L \in V^n(x)} L \subseteq K$.
- $\left| \bigcap_{L \in V^n(x)} L \right| = \infty$.

Thus, it holds that

$$\left| \bigcap_{L \in V^n(x)} L \setminus \{X_1, \dots, X_n\} \right| = \infty.$$

Hence, the algorithm can generate unseen strings from the target language K for all $n \geq z$, with probability 1. Notice that z indeed only depends on K, \mathcal{L} .

We now prove the termination property of our algorithm. To that end, it suffices to show that both while loops terminate with probability 1. As we argued before, $x \in L_z$ with probability 1, hence, the first while loop terminates after at most z steps. We now consider the termination of the second while loop. As we argued above, $\left| \bigcap_{L \in V^n(x)} L \setminus \{X_1, \dots, X_n\} \right| = \infty$ with probability 1, hence the loop will terminate after a finite number of steps. This concludes the proof. \square

We remark that the requirement that the set \mathcal{L}' in [Definition 17](#) is finite is crucial. The next example gives a collection of languages that is trivial for generation, yet it satisfies a modification of [Definition 17](#) that allows \mathcal{L}' to be infinite.

Example 2 (A Trivial Collection for Generation That “Almost” Satisfies [Definition 17](#)). Define the domain \mathcal{X} and the language collection \mathcal{L} as follows

$$\mathcal{X} = [0, 1] \cap \mathbb{Q} \quad \text{and} \quad \mathcal{L} = \left\{ \left[0, \frac{1}{n}\right] \cap \mathbb{Q} : n \in \mathbb{N} \right\},$$

where \mathbb{Q} is the set of rational numbers. Notice that both \mathcal{X} and \mathcal{L} are countable, and each $L \in \mathcal{L}$ is also countable. Consider the element $x = 0$ and the set $\mathcal{L}' = \mathcal{L}$. First, notice that $x \in L, \forall L \in \mathcal{L}'$ (by definition of every $L \in \mathcal{L}$). Moreover, it is not hard to see that $\bigcap_{L \in \mathcal{L}'} L = \{0\}$, hence $|\bigcap_{L \in \mathcal{L}'} L| = 1 < \infty$. It is also not hard to see that every finite sub-collection $\mathcal{L}'' \subseteq \mathcal{L}$, satisfies $|\bigcap_{L \in \mathcal{L}''} L| = \infty$. Hence, the conditions of [Definition 17](#) can only be satisfied by infinite sub-collections. We can show that

there is an algorithm that generates from \mathcal{L} , without seeing any example. Indeed, consider the algorithm that in every round $n \in \mathbb{N}$ outputs the element $1/n$. Let K be any target language. By definition, there is some $n_K \in \mathbb{N}$ such that $1/n \in K$, for all $n \geq n_K$. Hence, this algorithm can generate from K .

We now move on the proof the main result in this section. Similar to the identification setting before, we show the main result in two parts. First, we show that for any non-trivial collection of languages, no algorithm can generate at a rate faster than exponential. The approach shares some high-level ideas with the identification setting, but the more complicated condition that characterizes non-trivial generation makes the technical details more nuanced. In particular, leveraging the non-triviality condition, we can deduce that there exists a finite set of elements $\{x_{\ell_1}, \dots, x_{\ell_B}\}$ and a finite collection of languages \mathcal{L}' so that $\cap_{L \in \mathcal{L}'} L = \{x_{\ell_1}, \dots, x_{\ell_B}\}$. Then, whenever the training set consists exactly of the elements $\{x_{\ell_1}, \dots, x_{\ell_B}\}$ (containing also duplicates of them), no matter what element $x \in \mathcal{X} \setminus \{x_{\ell_1}, \dots, x_{\ell_B}\}$ the algorithm generates, there exists some $L \in \mathcal{L}'$ so that $x \notin L$. This allows us to find some “hard” distribution, which depends on the generating algorithm, and for which this event happens with exponentially small probability. The formal statement of the result and the technical details of the proof follow.

Lemma 5.10 (Exponential Rate Is Optimal for Generating From Any Non-trivial Collection). *Let \mathcal{L} be a non-trivial collection of languages for generation. Then, for any generating algorithm $(\mathcal{G}_n)_{n \in \mathbb{N}}$ there exists a valid distribution \mathcal{P} such that $\mathbb{E}[\text{er}(\mathcal{G}_n)] \geq C \cdot e^{-c \cdot n}$, for infinitely many $n \in \mathbb{N}$.*

Proof. Since \mathcal{L} is non-trivial for generation, there exists some $x \in \mathcal{X}$ and a finite $\mathcal{L}' \subseteq \mathcal{L}$ such that $x \in \cap_{L \in \mathcal{L}'} L$ and $|\cap_{L \in \mathcal{L}'} L| = B < \infty$. Let $\{x_{\ell_1}, \dots, x_{\ell_B}\} := \cap_{L \in \mathcal{L}'} L$ be the distinct elements that appear in the intersection of the sub-collection \mathcal{L}' . Define a collection of distributions $\{\mathcal{P}_L\}_{L \in \mathcal{L}'}$ that has two properties:

- For every $L \in \mathcal{L}'$ it holds that \mathcal{P}_L is valid for L .
- All the distributions $\{\mathcal{P}_L\}_{L \in \mathcal{L}'}$ put exactly the same mass on every element of the set $\{x_{\ell_1}, \dots, x_{\ell_B}\}$.

Notice that, by definition of $\{x_{\ell_1}, \dots, x_{\ell_B}\}$, there are collections of distributions that satisfy these two constraints.

For any $n \geq B$, let \mathcal{E}_n be the event that the training set is $(x_{\ell_1}, \dots, x_{\ell_B}, x_{\ell_1}, \dots, x_{\ell_1})$. Notice that under any $\mathcal{P} \in \{\mathcal{P}_L\}_{L \in \mathcal{L}'}$,

$$\Pr_{X_1, \dots, X_n \sim \mathcal{P}^n} [\mathcal{E}_n] \geq C \cdot e^{-c \cdot n},$$

for the same constants C, c for all $\mathcal{P} \in \{\mathcal{P}_L\}_{L \in \mathcal{L}'}$. Recall that for any valid distribution supported on a target language K

$$\text{er}(\mathcal{G}_n(X_1, \dots, X_n)) = \mathbb{1} \{ \mathcal{G}_n(X_1, \dots, X_n) \notin K \setminus \{X_1, \dots, X_n\} \}.$$

Since $|L| = \infty, \forall L \in \mathcal{L}$ and $|\cap_{L \in \mathcal{L}'} L| < \infty$, it follows that $|\mathcal{L}'| \geq 2$. Let $k := |\mathcal{L}'|$. By definition of \mathcal{L} , $k < \infty$, and by the previous argument $k \geq 2$. Moreover, notice that, for all $x \in \mathcal{X} \setminus \{x_{\ell_1}, \dots, x_{\ell_B}\}$ there exists $L \in \mathcal{L}'$ such that $x \notin L$. Indeed, if $x \in L, \forall L \in \mathcal{L}'$ then $x \in \cap_{L \in \mathcal{L}'} L \setminus \{x_{\ell_1}, \dots, x_{\ell_B}\}$, but $\cap_{L \in \mathcal{L}'} L \setminus \{x_{\ell_1}, \dots, x_{\ell_B}\} = \emptyset$. For every distribution p over \mathcal{X} it holds that

$$\Pr_{X \sim p} [X \in \{x_{\ell_1}, \dots, x_{\ell_B}\} \text{ or } \exists L \in \mathcal{L}' \text{ such that } X \notin L] = 1.$$

For every $n \in \mathbb{N}, n \geq B$, conditioned on the event \mathcal{E}_n , since the algorithm is a randomized mapping from the training set to \mathcal{X} , we have that $\mathcal{G}_n(x_{\ell_1}, \dots, x_{\ell_B}, x_{\ell_1}, \dots, x_{\ell_1}) = p_n$. We consider two cases:

- For infinitely many $n \in \mathbb{N}, n \geq B$ it holds that

$$\Pr_{X \sim p_n} [X \in \{x_{\ell_1}, \dots, x_{\ell_B}\}] \geq \frac{1}{2}.$$

Let \hat{N} be the infinite set for which the previous holds. Then, for all $\mathcal{P} \in \{\mathcal{P}_L\}_{L \in \mathcal{L}'}$ and for all $n \in \hat{N}$ it holds that

$$\begin{aligned} \mathbb{E}_{X_1, \dots, X_n \sim \mathcal{P}^n} [\text{er}(\mathcal{G}_n(X_1, \dots, X_n))] &= \Pr_{X_1, \dots, X_n \sim \mathcal{P}^n} [\mathcal{G}_n(X_1, \dots, X_n) \notin K \setminus \{X_1, \dots, X_n\}] \\ &\geq \Pr_{X_1, \dots, X_n \sim \mathcal{P}^n} [\mathcal{G}_n(X_1, \dots, X_n) \notin K \setminus \{X_1, \dots, X_n\} \mid \mathcal{E}_n] \cdot \Pr_{X_1, \dots, X_n \sim \mathcal{P}^n} [\mathcal{E}_n] \\ &\geq C \cdot e^{-c \cdot n} \cdot \Pr[\mathcal{G}_n(x_{\ell_1}, \dots, x_{\ell_B}, x_{\ell_1}, \dots, x_{\ell_1}) \notin K \setminus \{x_{\ell_1}, \dots, x_{\ell_B}\}] \\ &\quad \text{(by the definition of } \mathcal{E}_n\text{)} \\ &\geq C \cdot e^{-c \cdot n} \cdot \Pr_{X \sim p_n} [X \in \{x_{\ell_1}, \dots, x_{\ell_B}\}] \\ &\quad \text{(by the definition of } p_n\text{)} \\ &\geq C \cdot e^{-c \cdot n} \cdot \frac{1}{2}. \quad \text{(by the assumption on } \hat{N}\text{)} \end{aligned}$$

Thus, taking as the target distribution any $\mathcal{P} \in \{\mathcal{P}_L\}_{L \in \mathcal{L}'}$ we see that the algorithm indeed has an exponential rates lower bound.

- For infinitely many $n \in \mathbb{N}, n \geq B$, it holds that

$$\Pr_{X \sim p_n} [\exists L \in \mathcal{L}' \text{ such that } X \notin L] \geq \frac{1}{2}.$$

Then, due to the pigeonhole principle, there is some $L \in \mathcal{L}'$ such that for infinitely many $n \in \mathbb{N}$ it holds that

$$\Pr_{X \sim p_n} [X \notin L] \geq \frac{1}{2k}.$$

Let \hat{N} be the infinite set for which the previous holds. Then, for the data-generating distribution \mathcal{P}_L we have that

$$\begin{aligned} \mathbb{E}_{X_1, \dots, X_n \sim \mathcal{P}_L^n} [\text{er}(\mathcal{G}_n(X_1, \dots, X_n))] &= \Pr_{X_1, \dots, X_n \sim \mathcal{P}_L^n} [\mathcal{G}_n(X_1, \dots, X_n) \notin L \setminus \{X_1, \dots, X_n\}] \\ &\geq \Pr_{X_1, \dots, X_n \sim \mathcal{P}_L^n} [\mathcal{G}_n(X_1, \dots, X_n) \notin L \setminus \{X_1, \dots, X_n\} \mid \mathcal{E}_n] \cdot \Pr_{X_1, \dots, X_n \sim \mathcal{P}^n} [\mathcal{E}_n] \\ &\geq C \cdot e^{-c \cdot n} \cdot \Pr[\mathcal{G}_n(x_{\ell_1}, \dots, x_{\ell_B}, x_{\ell_1}, \dots, x_{\ell_1}) \notin K \setminus \{x_{\ell_1}, \dots, x_{\ell_B}\}] \\ &\quad \text{(by the definition of } \mathcal{E}_n\text{)} \\ &\geq C \cdot e^{-c \cdot n} \cdot \Pr_{X \sim p_n} [X \notin L] \quad \text{(by the definition of } p_n\text{)} \\ &\geq C \cdot e^{-c \cdot n} \cdot \frac{1}{2k}. \quad \text{(by the assumption on } \hat{N}\text{)} \end{aligned}$$

Then, we can pick the target distribution to be \mathcal{P}_L , and the exponential lower bound follows.

The proof is concluded by noticing that, from the pigeonhole principle, at least one of the previous two cases holds for any sequence of $\{p_n\}_{n \in \mathbb{N}}$. \square

5.2.2 A Sufficient Condition To Achieve Exponential Rate

Let us now shift our attention to the upper bound. Following the approach of Kleinberg and Mullainathan [KM24], we consider two settings: first, we assume access to a subset oracle which can answer questions $L_i \subseteq L_j$, for all $i, j \in \mathbb{N}$. Then, we consider the setting where we only have access to a membership oracle for each language in \mathcal{L} .

Before describing our approach let us explain why a direct adaptation of the approach of Bousquet et al. [BHM+21] does not seem to work in this setting. Recall that Bousquet et al. [BHM+21] transform in a black-box manner a learner which is eventually “correct” in the adversarial setting, to a learner that achieves exponential rates in the statistical setting, by running multiple copies of it on independent samples of the dataset, and then aggregating their results through a majority vote. A crucial property of the learner of Bousquet et al. [BHM+21] is that the majority vote is taken over objects that have binary values, namely the predicted label of the test point. One immediate obstacle to applying this approach here, is that the eventually correct generators will be outputting different valid strings in every iteration. Further, these valid strings might be even coming from a different subsets of the true language. Thus, it is not clear at all which aggregation strategy could lead to the desired result. One potential approach to circumvent this obstacle is to have all the generated strings give “votes” to the different languages of \mathcal{L} (potentially up to a cap n) that they belong to. It is clear that after some finite n_0 , with probability at least $1 - C \cdot e^{-c \cdot n}$, the target language K would be collecting votes from the majority of the strings. Unfortunately, it is not hard to see that for infinitely many $n_0 \in \mathbb{N}$ there must be another $L' \in \mathcal{L}, L' \neq K$, that is accumulating more votes than K ; if this was not the case we would have been able to identify K , for all countable \mathcal{L} , which contradicts our established lower bounds. Thus, it is not clear how to make this aggregation strategy work either.

Nevertheless, we show that, perhaps surprisingly, a much simpler strategy works: we only need to run one copy of the algorithm proposed by Kleinberg and Mullainathan [KM24] on the entire dataset to get exponential rates. In fact, we identify a sufficient condition that allows us to use any algorithm that works “in-the-limit” in the statistical setting without making any modifications to it. We believe that this idea might find other applications in the universal rates literature.

The following elementary result will be crucial for the analysis of both settings, *i.e.*, the one with the subset oracle and the one with just the membership oracle.

Lemma 5.11. *Let \mathcal{L} be a countable collection of languages. Let $\mathcal{A} = \{h_n\}_{n \in \mathbb{N}}$ be an algorithm that generates from \mathcal{L} in the limit with positive examples with the following additional property:*

- *for every target language $K \in \mathcal{L}$ there exists a finite set of examples $\{x_{i_1}, \dots, x_{i_\ell}\} \subseteq K$ that depends only on K and the enumeration of \mathcal{L}, \mathcal{X} , and*
- *a finite number $n_0 \in \mathbb{N}$ that depends on K and the enumeration of \mathcal{L}, \mathcal{X} ,*

such that \mathcal{A} always generates correctly if its input has size at least n_0 and it contains $x_{i_1}, \dots, x_{i_\ell}$. Then, \mathcal{A} generates from K with exponential rates in the statistical setting.

Proof. Let \mathcal{P} be a valid data-generating distribution. Then, by definition, $\text{supp}(\mathcal{P}) = K$, for some $K \in \mathcal{L}$. Let $x_{i_1}, \dots, x_{i_\ell} \subseteq K$ be a set of points such that after \mathcal{A} takes as input this set it starts generating correctly, *i.e.*, for any S such that $x_{i_1}, \dots, x_{i_\ell} \subseteq S$ and $|S| \geq n_0$ it holds that $h_{|S|}(S) \in K \setminus$

S. Since \mathcal{P} is a valid data-generating distribution it holds that $x_{i_1}, \dots, x_{i_\ell} \subseteq \text{supp}(\mathcal{P})$. Let $p_{i_1}, \dots, p_{i_\ell}$ be the mass of points $x_{i_1}, \dots, x_{i_\ell}$ under \mathcal{P} . Suppose we draw n samples i.i.d. from \mathcal{P} . Then, the probability that we do not observe all $x_{i_1}, \dots, x_{i_\ell}$ in the sample is bounded as

$$\begin{aligned}
\Pr_{X_1, \dots, X_n \sim \mathcal{P}^n} [\exists j \in [\ell]: x_{i_j} \notin \{X_1, \dots, X_n\}] &\leq \sum_{j \in [\ell]} \Pr_{X_1, \dots, X_n \sim \mathcal{P}^n} [x_{i_j} \notin \{X_1, \dots, X_n\}] \quad (\text{by a union bound}) \\
&= \sum_{j \in [\ell]} (1 - p_{i_j})^n \quad (\text{since we have i.i.d. draws}) \\
&\leq \sum_{j \in [\ell]} e^{-p_{i_j} \cdot n} \quad (\text{as } 1 - z \leq e^{-z} \text{ for all } z \in \mathbb{R}) \\
&\leq \ell \cdot e^{-\min_{j \in [\ell]} p_{i_j} \cdot n}.
\end{aligned}$$

Thus, the algorithm generates correctly in the statistical setting after taking as input $n \geq n_0 \in \mathbb{N}$ examples, with a probability at least $1 - C \cdot e^{-c \cdot n}$, for some distribution dependent constants C, c . This concludes the proof. \square

In the next two sections, we will show that the algorithms proposed by Kleinberg and Mullainathan [KM24] in the setting with access to a subset oracle or membership oracle already satisfy this property. For completeness, we present their algorithms and the related definitions.

5.2.3 Algorithm With Access To Subset Oracle

We start with the algorithm of Kleinberg and Mullainathan [KM24] which requires access to a subset oracle for \mathcal{L} , i.e., an oracle that for any two languages $L_i, L_j \in \mathcal{L}$ answers whether $L_i \subseteq L_j$. To that end, we first define the notion of critical language [KM24].

Definition 18 (Critical Language [KM24]). *Let $\mathcal{L} = \{L_1, L_2, \dots\}$ be a countable collection of languages. Let $S_n = \{x_{i_1}, \dots, x_{i_n}\} \subseteq \mathcal{X}$. For any $j \in \mathbb{N}$, we say that L_j is critical with respect to S_n if $S_n \subseteq L_j$ and for all $i < j$ if $S_n \subseteq L_i$ then $L_j \subseteq L_i$.*

The intuition behind this definition is that if two languages L_i, L_j are both critical and $i < j$, then it is “safer” to generate from L_j . This is exactly the way the algorithm from Kleinberg and Mullainathan [KM24] operates. To be more precise, in every iteration $n \in \mathbb{N}$ it performs the following steps:

- Let $\mathcal{L}_n = \{L_1, L_2, \dots, L_n\}$ be the first n languages of \mathcal{L} and $S_n = \{x_{i_1}, \dots, x_{i_n}\}$ be the set of examples observed so far.
- Let $\mathcal{C}_n \subseteq \mathcal{L}_n$ be the set of the critical languages with respect to S_n within \mathcal{L}_n (Definition 18). If $\mathcal{C}_n = \emptyset$ output an arbitrary $x \in \mathcal{X}$ and proceed to getting the $(n + 1)$ -th input. This step makes use of the subset oracle.¹⁵
- Let $L_k \in \mathcal{C}_n$ be the critical language with the highest index.

¹⁵Observe that it makes sense to output something arbitrary since the first consistent (in the sense that it contains the observed training examples) language in \mathcal{L} is critical by definition and hence if $\mathcal{C}_n = \emptyset$, we have not yet encountered a consistent language.

- Output the first unseen example from L_k , i.e., $x_j \in \mathcal{X}$ such that $j = \min\{i \in \mathbb{N} : x_i \in L_k, x_i \notin S_n\}$.

It is implicit in the analysis of Kleinberg and Mullainathan [KM24] that for every target language $K \in \mathcal{L}$, there exists a set $x_{i_1}, \dots, x_{i_\ell} \subseteq K$ and $n_0 \in \mathbb{N}$ that depend only on K and the enumeration of \mathcal{X}, \mathcal{L} , such that after n_0 steps if the above algorithm takes as input any set S that contains $\{x_{i_1}, \dots, x_{i_\ell}\}$, then it always generates a new example correctly. We make this explicit in the following lemma and provide a proof for completeness.

Lemma 5.12 (Adaptation of (4.3) from Kleinberg and Mullainathan [KM24]). *Let $\mathcal{L} = \{L_1, L_2, \dots\}$ be a countable collection of languages, let $K \in \mathcal{L}$ and let $z \in \mathbb{N}$ be the smallest number such that $L_z = K$. Then, there exist $x_{i_1}, \dots, x_{i_\ell} \in K$ that depend only on K and the enumeration of \mathcal{L} , such that if the algorithm of Kleinberg and Mullainathan [KM24] takes as input any set S for which $x_{i_1}, \dots, x_{i_\ell} \in S$ and $|S| \geq z$, where z depends only on K and the enumeration of \mathcal{L} , then it generates correctly from K .*

Proof. Let $L_{i_1}, \dots, L_{i_\ell} \subseteq \mathcal{L}$ with $i_1, \dots, i_\ell < z$ be the set of all languages that precede L_z in \mathcal{L} for which $L_z \not\subseteq L_{i_j}, j \in [\ell]$. Then, for each such L_{i_j} there exists some $x \in L_z$ so that $x \notin L_{i_j}$. Let x_{i_j} be the smallest indexed element in \mathcal{X} for which the previous holds. Notice that whenever $x_{i_1}, \dots, x_{i_\ell}$ is part of the input sample S , then L_z is critical; this follows immediately from the definition of criticality and the fact that the set S contradicts all the languages $L_i, i < z$, such that $L_z \not\subseteq L_i$. Moreover, when $|S| \geq z$, the algorithm outputs an unseen word from a critical language $L_{z'}$ with $z' \geq z$. By definition of the critical language, this means that $L_{z'} \subseteq L_z$. Hence, the algorithm generates correctly. \square

An immediate consequence of Lemma 5.11 is that the algorithm of Kleinberg and Mullainathan [KM24] with access to a subset query oracle generates with exponential universal rates.

5.2.4 Algorithm With Access To Membership Oracle

We now move on to the more involved version of the algorithm of Kleinberg and Mullainathan [KM24] that only requires membership access to every $L \in \mathcal{L}$. Recall this means that for every $x \in \mathcal{X}, L \in \mathcal{L}$ the algorithm can ask whether $x \in L$.

Before we describe the algorithm, we provide the definition of a modified notion of a critical language [KM24], which is based on a notion of a *projection of a language*, which we defined in Definition 16.¹⁶ Recall that, given some language L , we denote $L[m] = L \cap \{x_1, \dots, x_m\}$ (Definition 16).

Definition 19 (*m*-Critical Language [KM24]). *Let $\mathcal{L} = \{L_1, L_2, \dots\}$ be a countable collection of languages. Let $S_n = \{x_{i_1}, \dots, x_{i_n}\} \subseteq \mathcal{X}$. For any $j \in \mathbb{N}$, we say that L_j is *m*-critical with respect to S_n if $S_n \subseteq L_j$ and, for all $i < j$, if $S_n \subseteq L_i$, then $L_j[m] \subseteq L_i[m]$.*

We first give an intuitive description of the key modifications of the algorithm from the previous section that are required to make it work only with access to a membership oracle. First, notice that even though the algorithm cannot ask queries of the form $L_i \subseteq L_j$, it can ask queries of the

¹⁶Kleinberg and Mullainathan [KM24] do not explicitly define this term; we use it to simplify our discussion.

form $L_i[m] \subseteq L_j[m]$, for any finite $m \in \mathbb{N}$, by just asking $2m$ membership queries. Thus, the high-level idea is to replace subset queries with queries of the form $L_i[m] \subseteq L_j[m]$, for a sufficiently large $m \in \mathbb{N}$. The exact details are provided below.

- Let $S_n = \{x_{i_1}, \dots, x_{i_n}\}$ be the set of elements that have been presented to the learner up to step n . At the beginning of step n , set $m_n = \max\{m_{n-1}, i_n\}$.¹⁷
- Let $\mathcal{V}_n \subseteq \{L_1, L_2, \dots, L_n\}$ be the set of languages whose index is at most n and are consistent with the input S_n , i.e., $S_n \subseteq L, \forall L \in \mathcal{V}_n$. If no such languages exist, output an arbitrary $x \in \mathcal{X}$ and proceed to reading the $(n+1)$ -th input example. Notice that this can be done with n^2 membership queries.
- Let $m = m_n + 1$ and \mathcal{C}_n^m be the set of the m -critical languages within \mathcal{V}_n . Notice that since $\mathcal{V}_n \neq \emptyset$, for all $m \in \mathbb{N}$ there exists at least one m -critical language (the lowest indexed language within \mathcal{C}_n is m -critical for all $m \in \mathbb{N}$).
- Let $c_n^m \in \mathbb{N}$ be the largest index of a language in \mathcal{C}_n^m . If for some $i \leq m$, it holds that $x_i \in L_{c_n^m}$ and $x_i \notin S_n$, output x_i ¹⁸ and let $m_n = m$. Otherwise, let $m_n = m_n + 1$ and repeat the previous bullet point.

Kleinberg and Mullainathan [KM24] showed that the previous algorithm terminates in finitely many steps for every $n \in \mathbb{N}$ (Result (5.5) from Kleinberg and Mullainathan [KM24]). Moreover, they proved that for any enumeration of any target language $K \in \mathcal{L}$, there exists some $n_0 \in \mathbb{N}$ so that the algorithm generates correctly for all steps $n \geq n_0$ (Result (5.7) from Kleinberg and Mullainathan [KM24]). In fact, it is implicit in their analysis that for all $K \in \mathcal{L}$ there exist $x_{i_1}, \dots, x_{i_\ell} \in K$ that depend only on K and the enumeration of \mathcal{L}, \mathcal{X} , as well as a finite $n_0 \in \mathbb{N}$ that depends only on K and the enumeration of \mathcal{L}, \mathcal{X} , such that if an input sample S satisfies that **i)** $x_{i_1}, \dots, x_{i_\ell} \in S$ and **ii)** $|S| \geq n_0$, then the algorithm generates correctly. We make this fact explicit in the next result.

Lemma 5.13 (Adaptation of (5.7) from Kleinberg and Mullainathan [KM24]). *Let $\mathcal{L} = \{L_1, L_2, \dots\}$ be a countable collection of languages, let $K \in \mathcal{L}$ be the target language, and let $z \in \mathbb{N}$ be the smallest number such that $L_z = K$. Then, there exist $x_{i_1}, \dots, x_{i_\ell} \in K$ that depend only on K and the enumeration of \mathcal{L}, \mathcal{X} , such that if the algorithm of Kleinberg and Mullainathan [KM24] takes as input any set S for which $x_{i_1}, \dots, x_{i_\ell} \in S$ and $|S| \geq z$, where z depends only on K and the enumeration of \mathcal{L} , then it generates correctly from K .*

Proof. Let $z \in \mathbb{N}$ be the smallest number for which $L_z = K$. By definition, z has to be finite. Let $L_{k_1}, \dots, L_{k_\ell}$ be the set of all languages that precede L_z in \mathcal{L} for which $L_z \not\subseteq L_{k_j}, j \in [\ell]$. Then, for any such language L_{k_j} there exists some $x \in K$ such that $x \notin L_{k_j}$. Define x_{i_j} to be the smallest indexed element for which the previous holds. Hence, when the input sample S contains $x_{i_1}, \dots, x_{i_\ell}$ none of the languages $L_{i_j}, j \in [\ell]$, are consistent with S . Consider any iteration $n \in \mathbb{N}$, where $x_{i_1}, \dots, x_{i_\ell} \subseteq S_n$, and $n \geq z$. It follows immediately that L_z is m -critical for all $m \in \mathbb{N}$, and hence it is contained in the set \mathcal{C}_n^m . Thus, for all $m \in \mathbb{N}$, for the largest index c_n^m of a language in \mathcal{C}_n^m it holds that $c_n^m \geq z$. Recall that since the algorithm terminates (Result (5.5) from Kleinberg and Mullainathan [KM24]),

¹⁷Set $m_0 = 0$.

¹⁸If there are multiple such elements, output the one with the smallest index.

it will output some $x_m \in \mathcal{X}$ such that $x_m \notin S_n, x_m \in L_{z'}, z' \geq z$, and $L_{z'}[m] \subseteq L_z[m]$. This is because for all $m \in \mathbb{N}$, the largest index of a language in \mathcal{C}_n^m cannot drop below z . Thus, it follows that $x_m \in K \setminus S_n$. Hence, the algorithm generates correctly. \square

We are now ready to prove [Theorem 3.2](#).

Proof of Theorem 3.2. Let \mathcal{L} be some non-trivial collection for generation. An immediate corollary of [Lemma 5.11](#) and [Lemma 5.13](#) is that the algorithm of Kleinberg and Mullainathan [KM24] with access to a membership query oracle generates with exponential universal rates.

The exponential rates lower bound for generation follows immediately from [Lemma 5.10](#). \square

6 Proofs from [Section 3.2](#) (Generation With Breadth)

6.1 Proof of [Theorem 3.4](#) ($\text{MOP}(\cdot)$ Is Decidable For Iterative Generators)

In this section, we prove [Theorem 3.4](#) which we restate below.

Theorem 3.4. *For any token-by-token generator \mathcal{G} , $\text{MOP}(\mathcal{G})$ is decidable.*

Recall that a token-by-token generator \mathcal{G} is parameterized by a randomized Turing machine M , where M has the property that it halts on all inputs. \mathcal{G} generates as follows: in each iteration t , it queries M to get the next token s_t and iterates until M outputs EOS (*i.e.*, end of string). The algorithm to decide $\text{MOP}(\mathcal{G})$ is also simple: given a string s of length n , check token-by-token whether \mathcal{G} can output s_i conditioned on a prefix s_1, s_2, \dots, s_{i-1} generated so far. If at any point, s_i is not in the support of \mathcal{G} (or rather M) then, output No. Otherwise, output Yes. At each step, we can check if s_i can be generated by M using the folklore fact that membership oracles are decidable for Turing machines that always hold ([Lemma 6.1](#)). Note that we cannot use this folklore result directly for the generator \mathcal{G} , since even though M halts in each iteration, \mathcal{G} may not halt as the number of iterations is not bounded.

Lemma 6.1. *Consider a (randomized) Turing Machine M that halts on all inputs. The following problem is decidable: given strings s and p and a description of M , output Yes if M can output s given input p and output No otherwise.*

The proof of [Lemma 6.1](#) uses the following straightforward but subtle folklore lemmas.

Lemma 6.2. *Consider a (randomized) Turing Machine M that halts on all inputs. M has the following property: for each input string p , M performs at most a finite number of steps between any consecutive reads of their (internal) tape containing random bits.*

Lemma 6.3. *Consider a (randomized) Turing Machine M that halts on all inputs. For each input string p , there is a finite number $n_p \geq 1$ such that M reads at most n_p random bits always (regardless of the realization of the random bits).*

These enable us to prove [Lemma 6.1](#).

Proof of Lemma 6.1. Consider a string $s \in \Sigma^*$ of length m . We will check if M generates s with positive probability by iteratively checking if, for each $1 \leq t \leq m$, M generates token s_t with positive probability conditioned on having generated $s_1 \dots s_{t-1}$ so far.

Fix any $1 \leq t \leq m$. Suppose M has passed all earlier checks and, hence, it generates $s_1 \dots s_{t-1}$ with positive probability. Now, to complete the check for step t , it suffices to check that M generates s_t with positive probability having generated $s_1 \dots s_{t-1}$ so far. Since M halts on all inputs, Lemma 6.3 implies that there is a finite n_t such that M reads at most n_t bits of its internal random tape when given the corresponding input. Moreover, Lemma 6.2 implies that M performs finitely many operations between each of the n_t consecutive reads of the internal random tape. Hence, one can simulate the execution of M in finite time by checking all 2^{n_t} possible values of the random bits of M . If, for any of these 2^{n_t} values, M outputs s_t then we know that M passes the test and, otherwise, we know that M never generates s_t when provided the corresponding input.

One subtlety is that we do not know n_t . This is easy to overcome: since n_t is known to be finite, we can iterate over $n_t \in \mathbb{N}$ until we reach a value k where for each of the 2^k values of the first k random bits, M halts before reading the $(k+1)$ -th random bit. \square

Proof of Lemma 6.2

Proof of Lemma 6.2. The statement is vacuously true for M and p if M reads its internal tape at most once on input p always. Suppose with positive probability (over the randomness on M 's internal random tape), M reads its internal random tape at least twice given input p . Fix a value $r_1 = v$ of the first random bit such that M will (eventually) read the second bit r_2 on the random tape. Consider the step after M has read r_1 . If M performs a non-finite amount of computation before reading r_2 , then we have a contradiction to the fact that M is total since we have found an assignment v of the first random bit on which M performs an infinite number of steps. Hence, the result follows by contradiction. \square

Proof of Lemma 6.3

Proof of Lemma 6.3. Fix any input p to M . Toward a contradiction suppose that for any finite $n \geq 1$, there is (at least) one assignment v_1, v_2, \dots, v_n of the first n bits on M 's internal random tape on which M will read the $(n+1)$ -th random bit before halting. Therefore, for any $n \geq 1$, we have an assignment of the random bits for which M reads at least $n+1$ random bits and, hence, perform at least $n+1$ steps before halting. This is a contradiction to the fact that M halts always, for each value of the random bits on its internal tape. \square

6.2 Proof of Theorem 3.3 (Impossibility for Generation With Breadth)

In this section, we present the proof of Theorem 3.3 in two main parts; see Figure 5 for an outline.

First, we prove that if \mathcal{L} is not identifiable in the limit, then no algorithm in \mathcal{G} generates with breadth from \mathcal{L} at any rate. Recall that \mathcal{G} is the class of generating algorithms for which $\text{MOP}(\cdot)$ is decidable.

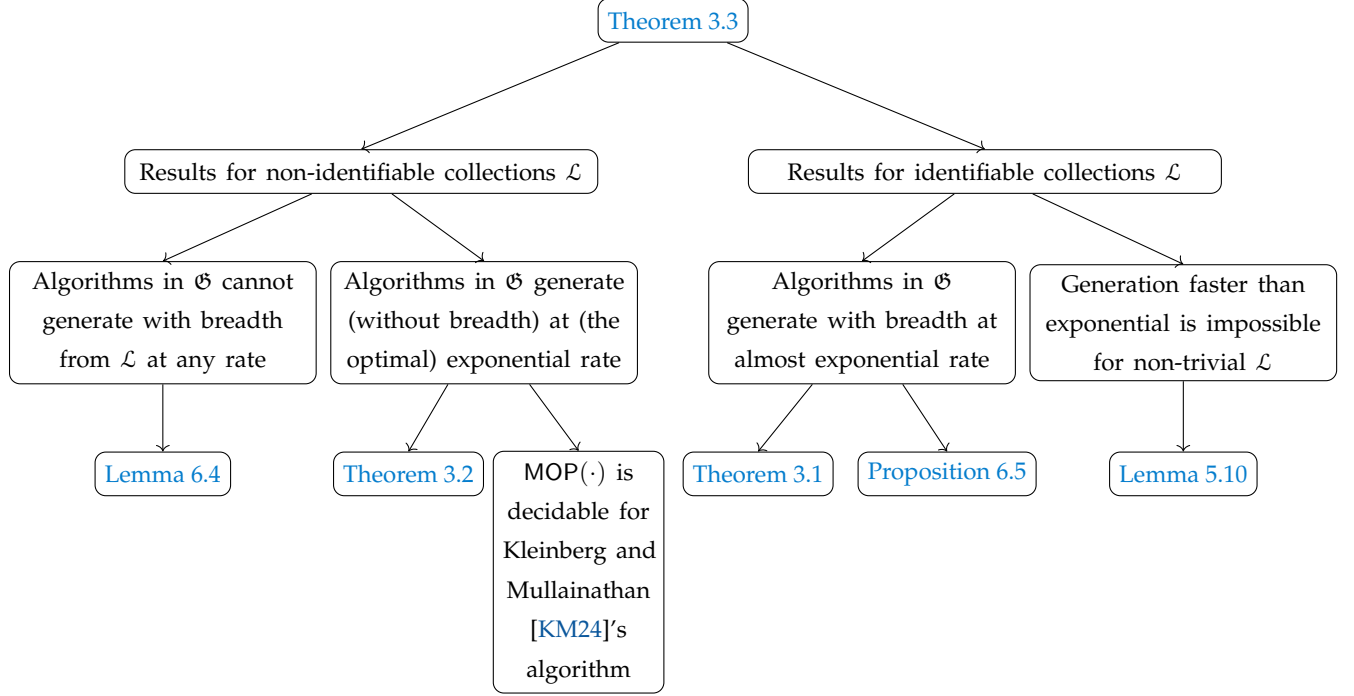


Figure 5: Outline of Proof of Theorem 3.3

Lemma 6.4. *Let \mathcal{L} be a countable collection of languages that is not identifiable in the limit. Then, for every rate R , there is no generating algorithm in \mathfrak{G} that can generate from \mathcal{L} with consistency and breadth at rate R .*

Proof. Let \mathcal{L} be a countable collection of languages that is not identifiable in the limit. Assume towards a contradiction that there exists some generating algorithm $(g_n) \in \mathfrak{G}$ that achieves consistency and breadth at some rate $R(n)$. Fix also some valid distribution \mathcal{P} supported over a target language K . This means that there exist c, C , that depend on \mathcal{P} , such that

$$\mathbb{E}_{X_1, \dots, X_n \sim \mathcal{P}^n} [\mathbb{1} \{ \text{supp}(g_n) \neq K \setminus \{X_1, \dots, X_n\} \}] \leq C \cdot R(c \cdot n).$$

This can be equivalently written as

$$\Pr_{X_1, \dots, X_n \sim \mathcal{P}^n} [\text{supp}(g_n) \neq K \setminus \{X_1, \dots, X_n\}] \leq C \cdot R(c \cdot n).$$

For every $n \in \mathbb{N}$, we denote by \mathcal{E}_n the event that $\text{supp}(g_n) = K \setminus \{X_1, \dots, X_n\}$. Let $z \in \mathbb{N}$ be the smallest number such that $L_z = K$. Recall that the elements of the universe are $\mathcal{X} = \{x_1, x_2, \dots\}$. Consider the following algorithm $(I_n)_{n \in \mathbb{N}}$ for identification:

- For every $n \in \mathbb{N}$, denote by $\{X_i\}_{i \in [n]}$ the sample i.i.d. from \mathcal{P} . Output the smallest index $j \in [n]$ such that

$$\mathbb{1} \{x_i \in L_j\} = \mathbb{1} \{x_i \in \text{supp}(g_n) \cup \{X_1, \dots, X_n\}\}, \forall i \in [n].$$

(Since $\mathcal{G}_n \in \mathfrak{G}$, $\text{MOP}(\mathcal{G}_n)$ is decidable and, hence, the above j can be computed.) If no such index exists, output an index arbitrarily.

We consider two cases.

Case A ($z = 1$): In this case, notice that if $\text{supp}(\mathcal{G}_n) = K \setminus \{X_1, \dots, X_n\}$, then, $I_n(X_1, \dots, X_n) = z$. This is because $\text{supp}(\mathcal{G}_n) \cup \{X_1, \dots, X_n\} = K$ and $L_z = K$, so for all $x \in \mathcal{X}$ it holds $\mathbb{1}\{x \in L_z\} = \mathbb{1}\{x \in \text{supp}(\mathcal{G}_n) \cup \{X_1, \dots, X_n\}\}$. Thus,

$$\begin{aligned} \Pr_{X_1, \dots, X_n \sim \mathcal{P}^n} [L_{I_n(X_1, \dots, X_n)} \neq K] &\leq \Pr_{X_1, \dots, X_n \sim \mathcal{P}^n} [I_n(X_1, \dots, X_n) \neq z] \\ &\leq \Pr_{X_1, \dots, X_n \sim \mathcal{P}^n} [I_n(X_1, \dots, X_n) \neq z \mid \mathcal{E}_n^c] \cdot \Pr_{X_1, \dots, X_n \sim \mathcal{P}^n} [\mathcal{E}_n^c] \\ &\quad \text{(since under } \mathcal{E}_n \text{ the algorithm identifies)} \\ &\leq 1 \cdot \Pr_{X_1, \dots, X_n \sim \mathcal{P}^n} [\mathcal{E}_n^c] \\ &\leq C \cdot R(c \cdot n). \end{aligned}$$

Case B ($z > 1$): For every language $L_j, j \in [z-1]$, let $i_j \in \mathbb{N}$ be the smallest number such that $\mathbb{1}\{x_{i_j} \in L_j\} \neq \mathbb{1}\{x_{i_j} \in L_z\}$. By definition of L_z , we have that i_j is well-defined. Moreover, let $n^* := \max_{j \in [z-1]} i_j$. Notice that for all $n \geq n^*$, under the event \mathcal{E}_n , we have that $I_n(X_1, \dots, X_n) = z$. To see why this is the case, notice that

1. Under the event \mathcal{E}_n it holds that $\mathbb{1}\{x \in L_z\} = \mathbb{1}\{x \in \text{supp}(\mathcal{G}_n) \cup \{X_1, \dots, X_n\}\}$ for all $x \in \mathcal{X}$.
2. Since $n \geq n^*$, for all $j \in [z-1]$ we have that $i_j \leq n$. Thus, under the event \mathcal{E}_n it cannot be the case that:

$$\mathbb{1}\{x_i \in L_j\} = \mathbb{1}\{x_i \in \text{supp}(\mathcal{G}_n) \cup \{X_1, \dots, X_n\}\}, \forall i \in [n].$$

Hence, using an identical argument as in the case $z = 1$ we have that

$$\Pr_{X_1, \dots, X_n \sim \mathcal{P}^n} [L_{I_n(X_1, \dots, X_n)} \neq K] \leq C \cdot R(c \cdot n), \forall n \geq n^*.$$

Since this holds for any valid distribution \mathcal{P} , using different \mathcal{P} -dependent constants, we see that the algorithm $(I_n)_{n \in \mathbb{N}}$ can identify \mathcal{L} at a rate R . Since \mathcal{L} is not identifiable in the limit, this contradicts [Theorem 3.1](#), and, hence, concludes the proof. □

The last ingredient we need to prove [Theorem 3.3](#) is an algorithm that given the index of a language, samples from it with breadth.

Proposition 6.5. *There exists a randomized computable algorithm \mathcal{A} for which $\text{MOP}(\cdot)$ is decidable and that, given as input a number $z \in \mathbb{N}$ and access to a collection of languages $\mathcal{L} = \{L_1, L_2, \dots\}$, satisfies $\text{supp}(\mathcal{A}(z)) = L_z$.*

Proof. The algorithm works as follows. Given the index z of the target language:

- Sample a natural number $\hat{n} \in \mathbb{N}$ from some distribution supported over \mathbb{N} .
- If $x_{\hat{n}} \in L_z$, (this can be checked by querying the membership oracle) return $x_{\hat{n}}$. Otherwise, repeat the previous bullet.

It follows immediately that this algorithm is computable and satisfies the requirements of the statement, since it is implemented via rejection sampling using a membership oracle to L_z (which is decidable), it is indeed in \mathfrak{G} . \square

We are now ready to prove [Theorem 3.3](#).

Proof of Theorem 3.3. First, consider the case that \mathcal{L} is not identifiable in the limit. Then, for every rate R , [Lemma 6.4](#) shows that no generating algorithm from \mathfrak{G} can generate from \mathcal{L} with consistency and breadth at rate R .

We now show that there is a consistent generation algorithm for \mathcal{L} at an optimal exponential rate. Notice that since \mathcal{L} is non-trivial for generation, by [Lemma 5.10](#), it holds that no algorithm can achieve rate faster than exponential. Moreover, by [Theorem 3.2](#) there exists an algorithm (namely the one by Kleinberg and Mullainathan [[KM24](#)]) that achieves exponential rates. Moreover, since this algorithm samples from a distribution that is a point mass, it is indeed in \mathfrak{G} .

Let us now consider the case that \mathcal{L} is identifiable in the limit. Then, by [Theorem 3.1](#), for every $g(n) = o(n)$, there exists an algorithm that identifies \mathcal{L} at rate $e^{-g(n)}$. It is not hard to turn this identification algorithm into an algorithm that is in \mathfrak{G} and generates with breadth via rejection sampling. This happens as described in [Proposition 6.5](#). Conditioned on the event that $L_z = K$, the previous algorithm indeed generates with breadth. Finally, since \mathcal{L} is non-trivial, no algorithm (even outside of \mathfrak{G}) can achieve a faster than exponential rate for consistent generation (even without breadth), by [Lemma 5.10](#). \square

Remark 5. One subtlety in the above proof is that to use Kleinberg and Mullainathan [[KM24](#)]'s algorithm for generation, we require it to output an arbitrarily large number of samples at each step. While the vanilla version of Kleinberg and Mullainathan [[KM24](#)]'s algorithm only outputs one sample at a time, it can easily be extended so that, given a number $m \geq 1$, it outputs m samples at each step. Moreover, the resulting algorithm is in \mathfrak{G} .

6.3 Proof of [Theorem 3.5](#) (Impossibility for Generation With Breadth in the Limit)

In this section, we prove [Theorem 3.5](#), which we restate below.

Theorem 3.5. *For every non-identifiable collection of countably many languages \mathcal{L} , no generating algorithm, for which $\text{MOP}(\cdot)$ ([Definitions 5 and 6](#)) is decidable, can generate with breadth from \mathcal{L} in the limit. If \mathcal{L} is identifiable, then there is a generator \mathcal{G} (for which $\text{MOP}(\mathcal{G})$ is decidable) that generates with breadth from \mathcal{L} .*

Proof of Theorem 3.5. The proof of [Theorem 3.5](#) is by a contradiction: we will show that if such a generator exists, it can be used to build an identification algorithm I for \mathcal{L} contradicting the fact that \mathcal{L} is non-identifiable. In addition to the generator \mathcal{G} , this identification algorithm uses

another sub-routine: an algorithm I_{PN} that, given a positive and negative enumeration of the target, identifies it in the limit. Such an identification algorithm always exists due to a result by Gold [Gol67]. The identifier I , which we construct, is as follows:

Input: Access to a generator \mathcal{G} for \mathcal{L} that (1) achieves consistency and breadth in the limit and (2) for which $\text{MOP}(\mathcal{G})$ is decidable, and access to the algorithm I_{PN} that identifies \mathcal{L} in the limit from a positive and negative enumeration of the target language.

Description:

1. **For each** $t \in \mathbb{N}$ **do:**

- (a) Observe the t -th sample s_t and let S_t be the set of samples seen so far
- (b) Train the generator \mathcal{G} from scratch over the t samples in S_t
- (c) Label the first t strings x_1, \dots, x_t of the domain as $\text{MOP}(\mathcal{G})(x_1), \dots, \text{MOP}(\mathcal{G})(x_t)$ ^a
- (d) Train I_{PN} from scratch on samples x_1, \dots, x_t with labels $\text{MOP}(\mathcal{G})(x_1), \dots, \text{MOP}(\mathcal{G})(x_t)$
- (e) **output** the index guessed by I_{PN} and go to the next iteration

^aHere, $\text{MOP}(\mathcal{G})(x)$ is the answer to the membership oracle problem for \mathcal{G} given input x .

Since $\text{MOP}(\mathcal{G})$ is decidable, the above algorithm can be implemented using a Turing machine. We claim that the above algorithm identifies the target language K after a finite number of iterations. Let z be the first index at which K appears. To formalize this, fix any enumeration s_1, s_2, \dots of the target language K . Since \mathcal{G} generates with breadth in the limit, there is a finite iteration $t_{\mathcal{G}}$ after which K generates with breadth from K and, hence, $\text{supp}(\mathcal{G}) = K$. Hence, after iteration $t_{\mathcal{G}}$, for any string x , $\text{MOP}(\mathcal{G})(x) = \mathbb{1}\{x \in K\}$. In other words, I_{PN} is provided with accurate positive and negative labels in all subsequent iterations $t \geq t_{\mathcal{G}}$. Since I_{PN} identifies in the limit, there is a finite t_{PN} such that I_{PN} identifies K once it is given labels for the first $t \geq t_{PN}$ examples in the domain. It follows that I_{PN} and, hence, our algorithm identifies K after $\max\{t_{\mathcal{G}}, t_{PN}\} < \infty$ iterations. This gives the desired contradiction, proving [Theorem 3.5](#). Note that the above identification algorithm does not need to know either $t_{\mathcal{G}}$ or t_{PN} . (Of course, as a consequence, our algorithm does not know when it has identified K .) \square

7 Proofs from [Section 3.3](#) (Generation With Approximate Consistency and Breadth)

7.1 Proof of [Theorem 3.7](#) (Impossibility in the Limit)

In this section, we prove [Theorem 3.7](#), which we restate below.

Theorem 3.7 (Impossibility of Unambiguous Generation in the Limit). *For every non-identifiable collection of countably many languages \mathcal{L} , no generating algorithm stable in the limit for which $\text{MOP}(\cdot)$ ([Definitions 5 and 6](#)) is decidable can unambiguously generate from \mathcal{L} in the limit.*

Proof of Theorem 3.7. By the way of contradiction, suppose that there is an algorithm $\mathcal{G} = (\mathcal{G}_n)$ for which $\text{MOP}(\cdot)$ is decidable (at each n) and which is an unambiguous generator for \mathcal{L} . We will use \mathcal{G} to construct an algorithm that identifies \mathcal{L} in the limit, hence, contradicting the non-identifiability of \mathcal{L} .

Fix any enumeration x_1, x_2, \dots of the domain \mathcal{X} . For each language $L \in \mathcal{L}$ and number $t \geq 1$, define the t -prefix of L as the subset $L[t]$ of the first t -elements of the domain $\{x_1, \dots, x_t\}$ in \mathcal{L} , i.e.,

$$L[t] := \{x_1, \dots, x_t\} \cap L.$$

To complete the above outline, consider the following algorithm, which we claim identifies \mathcal{L} .

Input: Access to a generator \mathcal{G} for \mathcal{L} that (1) that is unambiguous in the limit and (2) for which $\text{MOP}(\cdot)$ is decidable at each step

Description:

1. **For each** $t \in \mathbb{N}$ **do:**

- (a) Observe the t -th sample s_t and let S_t be the set of samples seen so far
- (b) Train the generator \mathcal{G}_{t-1} on s_t to get \mathcal{G}_t
- (c) Create a set of languages consistent with observed samples $C_S(t) \subseteq \{L_1, \dots, L_t\}$ that includes each $L \in \{L_1, \dots, L_t\}$ that is consistent with S_t (i.e., $L \supseteq S_t$)
- (d) Construct a set of languages consistent with the generator $C_G(t) \subseteq \{L_1, \dots, L_t\}$ that languages L from $\{L_1, \dots, L_t\}$ except if $L[t] \not\subseteq \text{supp}(\mathcal{G}_t) \cup S_t$, which can be checked in finite time using a decider for $\text{MOP}(\mathcal{G}_t)$
- (e) **output** the index of the smallest-indexed language in $C_S(t) \cap C_G(t)$ (or an arbitrary index if $C_S(t) \cap C_G(t)$ is empty)

Since the algorithm outputs the smallest index in $C_S(t) \cap C_G(t)$, it identifies K if it is the smallest-indexed language in $C_S(t) \cap C_G(t)$. The following conditions ensure this:

- (A) $L_z \in C_S(t)$ and $L_z \in C_G(t)$, where z is the smallest index at which K appears in \mathcal{L} ; and
- (B) For any $i < z$, either $L_i \notin C_S(t)$ or $L_i \notin C_G(t)$

We claim that there are finite times t_a and t_b where, for any $t \geq t_a$, Condition (A) holds and, for any $t \geq t_b$, Condition (B) holds. This claim implies that the above algorithm identifies K in the limit, leading to the desired contradiction.

Condition A holds after a finite time. Since S_t only contains samples from K , $K \in C_S(t)$ for all $t \geq 1$. Further, since $\mathcal{G} = (\mathcal{G}_t)$ is an unambiguous generator for \mathcal{L} in the limit, there exists a finite $t_0 \geq 0$, such that for all $t \geq t_0$,

$$|\text{supp}(\mathcal{G}_n) \triangle K| < \min_{L \in \mathcal{L}: L \neq K} |\text{supp}(\mathcal{G}_n) \triangle L|. \quad (10)$$

Hence, in particular, for all $t \geq t_0$

$$|\text{supp}(\mathcal{G}_n) \setminus K|, |K \setminus \text{supp}(\mathcal{G}_n)| < \infty.$$

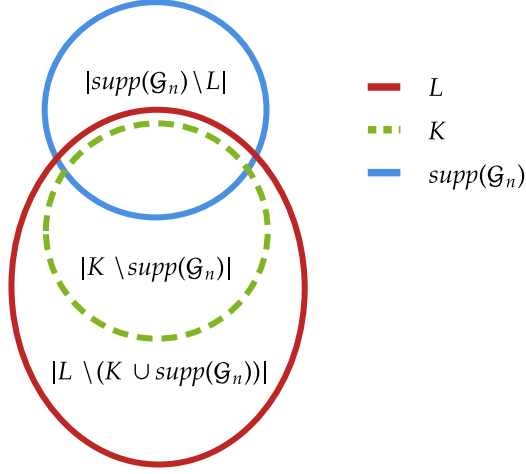


Figure 6: Figure illustrating the decomposition of $\text{supp}(\mathcal{G}_t) \triangle K$ and $\text{supp}(\mathcal{G}_t) \triangle L$.

Furthermore, as \mathcal{G} is stable, after some time t_1 , $\text{supp}(\mathcal{G}_n)$ stops changing. Consider any $t \geq \max\{t_0, t_1\}$. Since $K \setminus \text{supp}(\mathcal{G}_t)$ has finitely many elements and $K \setminus \text{supp}(\mathcal{G}_t) = K \setminus \text{supp}(\mathcal{G}_{t'})$ for any $t' \geq \max\{t_0, t_1\}$, there is a finite time t_2 after which all elements of $K \setminus \text{supp}(\mathcal{G}_t)$ have been observed. Therefore, for any $t \geq \{t_0, t_1, t_2\}$, it must hold that $K \subseteq (\text{supp}(\mathcal{G}_t) \cup S_t)$ and, hence, that $K \in C_G(t)$. Thus, it suffices to fix $t_a = \max\{t_0, t_1, t_2\}$.

Condition B holds after a finite time. Since there are only finitely many $i < z$, it suffices to show that for each $i < z$, there is a finite time t_i after which either $L_i \notin C_S(t)$ or $L_i \notin C_G(t)$. Fix any $i < z$. Consider two cases:

- **Case A** ($K \setminus L_i \neq \emptyset$): In this case, there exists an $x \in K \setminus L_i$ and, hence, after some finite time t_i when x has been observed $L_i \not\supseteq S_t$ and, hence, $L_i \notin C_S(t)$.
- **Case B** ($K \setminus L_i = \emptyset$): In this case, $L_i \supsetneq K$, and our proof is based on the following observation.

Lemma 7.1. *For any $t \geq t_a$ and $L \supsetneq K$, it holds that $L \setminus \text{supp}(\mathcal{G}_t) \neq \emptyset$.*

Proof. Since $t \geq t_0$, Equation (10) holds. From Figure 6, observe that

$$\begin{aligned} |\text{supp}(\mathcal{G}_t) \triangle K| &\geq |\text{supp}(\mathcal{G}_t) \setminus L| + |K \setminus \text{supp}(\mathcal{G}_t)| \\ |\text{supp}(\mathcal{G}_t) \triangle L| &= |\text{supp}(\mathcal{G}_t) \setminus L| + |K \setminus \text{supp}(\mathcal{G}_t)| + |L \setminus (K \cup \text{supp}(\mathcal{G}_t))|. \end{aligned}$$

Chaining the above with Equation (10) and canceling like terms implies

$$|L \setminus (K \cup \text{supp}(\mathcal{G}_t))| > 0.$$

Hence, in particular, $|L \setminus \text{supp}(\mathcal{G}_t)| > 0$, which is the desired result. \square

Hence, in this case, $L_i \setminus \text{supp}(\mathcal{G}_t) \neq \emptyset$. Let $j(i)$ be the smallest natural number such that $x_{j(i)} \in L$ but $x_{j(i)} \notin \text{supp}(\mathcal{G}_t)$. (Note that the value $j(i)$ does not depend on t , since as discussed in the proof of Condition A after $t = t_a$, the $\text{supp}(\mathcal{G}_t)$ becomes stable and not change in subsequent iterations.) Therefore, it follows that $L_i[j(i)] \not\subseteq \text{supp}(\mathcal{G}_t) \cup S_t$ for $t \geq t_a$ and, hence, for $t \geq \max\{i(j), t_a\}$ by construction, $L_i \notin C_G(t)$. This completes the proof of Case B and by earlier discussion, also the proof of [Theorem 3.7](#).

□

7.2 Proof of [Theorem 3.6](#) (Impossibility in the Statistical Setting)

In this section, we prove the impossibility result for unambiguous generation in the statistical setting. Our approach is to establish a connection to the online setting and leverage the impossibility result we have already shown there ([Theorem 3.7](#)). Namely, we will show that given such an unambiguous generator that works in the statistical setting, we can construct a generator that works in the online setting, with high probability. Using the construction from [Section 7.1](#), we can turn this generator to one that *identifies*. The details of our approach follow.

First, we describe some constructions due to Angluin [[Ang88](#)] that will be useful for our derivation. The following can be found in Example 3 from Angluin [[Ang88](#)].

Definition 20 (Distribution Induced by Sequence). *Let $\sigma = (x_{i_1}, x_{i_2}, \dots)$ be some countable sequence of elements in \mathcal{X} , and $\sigma_j = x_{i_j}, j \in \mathbb{N}$. Define \mathcal{P}_σ to be a distribution such that its mass $\mathcal{P}_\sigma(x)$ on any point $x \in \mathcal{X}$ is*

$$\mathcal{P}_\sigma(x) := \sum_{j \in \mathbb{N}: \sigma_j = x} \frac{1}{2^{j+1}},$$

with a sum over an empty set of indices interpreted as 0.

Angluin [[Ang88](#)] describes a way to draw i.i.d. samples from \mathcal{P}_σ , given only access to finite prefixes of σ and to an oracle that simulates a fair coin.¹⁹ The idea is natural: flip the fair coin until a head is observed, let I be the random variable denoting the number of trials it needed, and output the string x_{I+1} . This process gives i.i.d. draws from \mathcal{P}_σ .

Proposition 7.2 (Example 4 from Angluin [[Ang88](#)]). *Let $\sigma = (x_{i_1}, x_{i_2}, \dots)$ be some countable sequence of elements in \mathcal{X} . Given access to an oracle that simulates fair coin flips and an oracle which given input any $j \in \mathbb{N}$ returns σ_j , there exists a computable algorithm that samples from \mathcal{P}_σ ([Definition 20](#)) and terminates with probability 1.*

The next result shows that a stable generating algorithm for which $\text{MOP}(\cdot)$ is decidable and achieves unambiguous generation at some rate $R(\cdot)$, “stabilizes” to an unambiguous generator when executed on an infinite i.i.d. stream of data drawn from a valid distribution.

Lemma 7.3. *Let $R: \mathbb{N} \rightarrow \mathbb{R}_{\geq 0}$ be a rate function, i.e., $\lim_{n \rightarrow \infty} R(n) = 0$, let \mathcal{L} be a countable language collection, and $(\mathcal{G}_n: \mathcal{X}^n \rightarrow \mathfrak{G})_{n \in \mathbb{N}}$ be a generating algorithm for which $\text{MOP}(\cdot)$ is decidable and which satisfies the following two properties:*

¹⁹In fact, Angluin’s construction generalizes to oracles that simulate any (non-deterministic) coin in a straightforward way.

- $(\mathcal{G}_n)_{n \in \mathbb{N}}$ is a stable generator (Definition 7), and
- for its unambiguous generation error $\text{er}(\cdot)$, it holds that, for every valid distribution \mathcal{P} with respect to \mathcal{L} there exist $c, C > 0$ such that $\mathbb{E}_{X_1, \dots, X_n \sim \mathcal{P}^n} [\text{er}(\mathcal{G}_n(X_1, \dots, X_n))] \leq C \cdot R(c \cdot n)$.

Then, for every valid distribution \mathcal{P} with respect to \mathcal{L} it holds that

$$\Pr_{\{X_i\}_{i \in \mathbb{N}} \sim \mathcal{P}^\infty} [\exists n^* \in \mathbb{N}: \forall n \geq n^* \text{ it holds that } \text{er}(\mathcal{G}_n(X_1, \dots, X_n)) = 0] = 1.$$

In other words, the generating algorithm $\mathcal{G} = (\mathcal{G}_n)$ stabilizes to an unambiguous generation in the online sense with probability 1. Roughly speaking, given the above result, Theorem 3.6 will follow from a contradiction to the impossibility result in the online setting Theorem 3.7.

Proof of Lemma 7.3. Assume towards contradiction that there exists some valid \mathcal{P} with respect to \mathcal{L} so that

$$\Pr_{\{X_i\}_{i \in \mathbb{N}} \sim \mathcal{P}^\infty} [\exists n^* \in \mathbb{N}: \forall n \geq n^* \text{ it holds that } \{\text{er}(\mathcal{G}_n(X_1, \dots, X_n)) = 0\}] = c' < 1.$$

Let us also denote $c'' := 1 - c'$. Notice that $c'' > 0$. Since \mathcal{P} is a valid distribution with respect to \mathcal{L} , it is supported over some $K \in \mathcal{L}$, so we have that, with probability 1, an infinite i.i.d. draw from \mathcal{P} is an enumeration of K (see Proposition 5.2). Let us call this event \mathcal{E}_1 .

Moreover, since \mathcal{G}_n is a stable generator (in an online sense), under the event \mathcal{E}_1 (i.e., when the samples from \mathcal{P} form an enumeration of K), there exists some smallest number $t^* := t^*(X_1, \dots) \in \mathbb{N}$ such that for all $n \geq t^*$

$$\text{supp}(\mathcal{G}_n(X_1, \dots, X_n)) = \text{supp}(\mathcal{G}_{n+1}(X_1, \dots, X_{n+1})).$$

Now, t^* depends on the specific enumeration drawn and, hence, the distribution \mathcal{P} induces a distribution over t^* . Further, note that with probability 1, $t^* < \infty$. Hence, $\Pr_{\{X_i\}_{i \in \mathbb{N}} \sim \mathcal{P}^\infty} [t^*(X_1, \dots) > n]$ approaches 0 as $n \rightarrow \infty$. In particular, there is some number $n_1 \in \mathbb{N}$ such that for all $n \geq n_1$

$$\Pr_{\{X_i\}_{i \in \mathbb{N}} \sim \mathcal{P}^\infty} [t^*(X_1, \dots) > n] \leq \frac{c''}{3}.$$

Moreover, since the generator achieves rate $R(\cdot)$ and $\lim_{n \rightarrow \infty} R(n) = 0$, it holds that

$$\lim_{n \rightarrow \infty} \Pr_{X_1, \dots, X_n \sim \mathcal{P}^n} [\text{er}(\mathcal{G}_n(X_1, \dots, X_n)) \neq 0] = 0.$$

Thus, there is some $n_2 \in \mathbb{N}$ such that, for all $n \geq n_2$

$$\Pr_{X_1, \dots, X_n \sim \mathcal{P}^n} [\text{er}(\mathcal{G}_n(X_1, \dots, X_n)) \neq 0] \leq \frac{c''}{3}.$$

Let $n_3 := \max\{n_1, n_2\}$. Hence, taking a union bound, we see that with probability at least $1 - 2c''/3$ over the draw of $\{X_i\}_{i \in \mathbb{N}}$ it holds that

- $\text{er}(\mathcal{G}_{n_3}(X_1, \dots, X_{n_3})) = 0$, and
- $\text{supp}(\mathcal{G}_n(X_1, \dots, X_n)) = \text{supp}(\mathcal{G}_{n_3}(X_1, \dots, X_{n_3}))$, for all $n \geq n_3$.

By the definition of $\text{er}(\cdot)$ (Equation (6)), for any $n, n' \in \mathbb{N}$, samples x_{i_1}, \dots, x_{i_n} and $x_{j_1}, \dots, x_{j_{n'}}$ it holds that

$$\begin{aligned} \text{supp}(\mathcal{G}_n(x_{i_1}, \dots, x_{i_n})) &= \text{supp}(\mathcal{G}_{n'}(x_{j_1}, \dots, x_{j_{n'}})) \implies \\ \text{er}(\mathcal{G}_n(x_{i_1}, \dots, x_{i_n})) &= \text{er}(\mathcal{G}_{n'}(x_{j_1}, \dots, x_{j_{n'}})) \end{aligned}$$

These two conditions immediately imply that, with probability at least $1 - 2c''/3 > c'$, for all $n \geq n_3$ it holds that

- $\text{er}(\mathcal{G}_n(X_1, \dots, X_n)) = 0$, and
- $\text{supp}(\mathcal{G}_{n+1}(X_1, \dots, X_{n+1})) = \text{supp}(\mathcal{G}_n(X_1, \dots, X_n))$.

Hence,

$$\Pr_{\{X_i\}_{i \in \mathbb{N}} \sim \mathcal{P}^\infty} [\exists n^* \in \mathbb{N}: \forall n \geq n^* \text{ it holds that } \{\text{er}(\mathcal{G}_n(X_1, \dots, X_n)) = 0\}] > c',$$

which gives the desired contradiction. This concludes the proof. \square

Having established the previous result, we are ready to show how to use such a generator that works in the statistical setting to get a generator in the online setting. The idea of the proof is to use the enumeration σ provided from the adversary to define a valid distribution \mathcal{P}_σ (see Proposition 7.2) and then run the aforementioned generator on this distribution.

Lemma 7.4. *Let $R: \mathbb{N} \rightarrow \mathbb{R}_{\geq 0}$ be a rate function, i.e., $\lim_{n \rightarrow \infty} R(n) = 0$, let \mathcal{L} be a language collection, and $(\mathcal{G}_n: \mathcal{X}^n \rightarrow \mathfrak{G})_{n \in \mathbb{N}}$ be generating algorithm for which MOP is decidable and satisfies the following two properties:*

- $(\mathcal{G}_n)_{n \in \mathbb{N}}$ is a stable generator (Definition 7), and
- for its unambiguous generation error, it holds that, for every valid distribution \mathcal{P} with respect to \mathcal{L} there exist $c, C > 0$ such that $E_{X_1, \dots, X_n \sim \mathcal{P}^n} [\text{er}(\mathcal{G}_n(X_1, \dots, X_n))] \leq C \cdot R(c \cdot n)$.

Then, there is a randomized generating algorithm $(\mathcal{G}'_n: \mathcal{X}^n \xrightarrow{r} \mathfrak{G})_{n \in \mathbb{N}}$ for which, for any target language $K \in \mathcal{L}$ and every enumeration σ of K , it holds that

- $(\mathcal{G}'_n)_{n \in \mathbb{N}}$ is a stable generator (Definition 7), and
-

$$\Pr [\exists n^* \in \mathbb{N}: \forall n \geq n^* \text{ it holds that } \text{er}(\mathcal{G}'_n(\sigma_1, \dots, \sigma_n)) = 0] = 1,$$

where the probability is with respect to the internal randomness of the algorithm.

Proof. Let $K \in \mathcal{L}$ be any target language and σ be any enumeration of K . Let \mathcal{P}_σ be the distribution defined in Definition 20. We know that, by definition, \mathcal{P}_σ is valid with respect to \mathcal{L} , since it is supported on K . Let $(\mathcal{G}'_n)_{n \in \mathbb{N}}$ be a generating algorithm which, for every $n \in \mathbb{N}$, runs \mathcal{G}_n on \mathcal{P}_σ . In order to draw samples from \mathcal{P}_σ the generator \mathcal{G}'_n uses its internal randomness and the process

described in [Proposition 7.2](#). Since \mathcal{P}_σ is a valid distribution with respect to \mathcal{L} , [Lemma 7.3](#) gives us that

$$\Pr_{\{X_i\}_{i \in \mathbb{N}} \sim \mathcal{P}_\sigma^\infty} [\exists n^* \in \mathbb{N}: \forall n \geq n^* \text{ it holds that } \{\text{er}(\mathcal{G}_n(X_1, \dots, X_n)) = 0\}] = 1.$$

Hence, this implies that

$$\Pr [\exists n^* \in \mathbb{N}: \forall n \geq n^* \text{ it holds that } \{\text{er}(\mathcal{G}'_n(\sigma_1, \dots, \sigma_n)) = 0\}] = 1,$$

where the probability is taken with respect to the internal randomness of the algorithm. Moreover, since $(\mathcal{G}_n)_{n \in \mathbb{N}}$ is a stable generator it also holds that $(\mathcal{G}'_n)_{n \in \mathbb{N}}$ is a stable generator. This concludes the proof. \square

We are now ready to prove [Theorem 3.6](#), which follows as corollary of [Lemma 7.4](#) and the impossibility result from the online setting ([Theorem 3.7](#)).

Proof of Theorem 3.6. Let \mathcal{L} be a countable collection of languages. Assume that such a stable generating algorithm exists. Then, using the construction from [Lemma 7.4](#) we get a stable generator that generates unambiguously in the limit, for every target language $K \in \mathcal{L}$ and every enumeration σ of K , with probability 1. This contradicts the impossibility result from [Theorem 3.7](#). \square

8 Proofs from [Section 3.4](#) (Further Results for Identification)

8.1 Proof of [Proposition 3.8](#) (Identification Using Subset Oracle)

We first give a sufficient condition on the algorithm that identifies in the limit that allows one to directly use it in the statistical setting and get exponential rates.

Lemma 8.1. *Let \mathcal{L} be a countable collection of languages. Let $\mathcal{A} = \{h_n\}_{n \in \mathbb{N}}$ be an algorithm that identifies \mathcal{L} in the limit with positive examples with the following additional property:*

- *for every target language $K \in \mathcal{L}$ there exists a finite set of examples $\{x_{i_1}, \dots, x_{i_\ell}\} \subseteq K$ that depends only on K , and the enumeration of \mathcal{L}, \mathcal{X} ,*
- *and a finite number $n_0 \in \mathbb{N}$ that depends on K , the enumeration of \mathcal{L}, \mathcal{X} ,*

such that \mathcal{A} always identifies correctly if its input has size at least n_0 and it contains $x_{i_1}, \dots, x_{i_\ell}$. Then, \mathcal{A} identifies K with exponential rates in the statistical setting.

Proof. Let \mathcal{P} be a valid data-generating distribution. Then, by definition, $\text{supp}(\mathcal{P}) = K$, for some $K \in \mathcal{L}$. Let $x_{i_1}, \dots, x_{i_\ell} \subseteq K$ be a set of points such that after \mathcal{A} takes as input this set it starts identifying correctly, i.e., for any S such that $x_{i_1}, \dots, x_{i_\ell} \subseteq S$ and $|S| \geq n_0$ it holds that $L_{h_{|S|}}(S) = K$. Since \mathcal{P} is a valid data-generating distribution it holds that $x_{i_1}, \dots, x_{i_\ell} \subseteq \text{supp}(\mathcal{P})$. Let $p_{i_1}, \dots, p_{i_\ell}$

be the mass of points $x_{i_1}, \dots, x_{i_\ell}$ under \mathcal{P} . Suppose we draw n samples i.i.d. from \mathcal{P} . Then, the probability that we do not observe all $x_{i_1}, \dots, x_{i_\ell}$ in the sample is bounded as

$$\begin{aligned}
\Pr_{X_1, \dots, X_n \sim \mathcal{P}^n} [\exists j \in [\ell]: x_{i_j} \notin \{X_1, \dots, X_n\}] &\leq \sum_{j \in [\ell]} \Pr_{X_1, \dots, X_n \sim \mathcal{P}^n} [x_{i_j} \notin \{X_1, \dots, X_n\}] \quad (\text{by a union bound}) \\
&= \sum_{j \in [\ell]} (1 - p_{i_j})^n \quad (\text{since we have i.i.d. draws}) \\
&\leq \sum_{j \in [\ell]} e^{-p_{i_j} \cdot n} \quad (\text{using that } 1 - z \leq e^{-z} \text{ for all } z \in \mathbb{R}) \\
&\leq \ell \cdot e^{-\min_{j \in [\ell]} p_{i_j} \cdot n}.
\end{aligned}$$

Thus, the algorithm identifies correctly in the statistical setting after taking as input $n \geq n_0 \in \mathbb{N}$ examples, with probability at least $1 - C \cdot e^{-c \cdot n}$, for some distribution dependent constants C, c . This concludes the proof. \square

We are now ready to prove [Proposition 3.8](#).

Proof of Proposition 3.8. Let \mathcal{L} be a collection that is identifiable in the limit and assume access to a subset oracle. From [Section B.1](#) we know that the algorithm of Kleinberg and Mullainathan [KM24] given access to a subset oracle identifies any identifiable collection \mathcal{L} in the limit. Moreover, by [Lemma 5.12](#) we get that this algorithm satisfies the condition of [Lemma 8.1](#), thus this result immediately gives us that the algorithm of Kleinberg and Mullainathan [KM24] obtains exponential rates for identification in the statistical setting (assuming access to a subset oracle). \square

8.2 Proof of [Proposition 3.9](#) (Identification of Finite Collections)

Similar to the previous section, we will show that there exists an algorithm that satisfies [Lemma 8.1](#), i.e., for any finite collection of (potentially infinite) languages, it identifies with exponential rates. Recall the domain \mathcal{X} has an enumeration $\mathcal{X} = \{x_1, \dots\}$. Consider the following algorithm for identification.

Algorithm [Algorithm 8.2](#) - Identifying a finite collection $\mathcal{L} = \{L_1, \dots, L_k\}$ in the limit

Description:

1. **for each** $t \in \mathbb{N}$ **do:**

- (a) Let $S_t = \{x_{i_1}, \dots, x_{i_t}\}$, where x_{i_ℓ} is the element the algorithm sees in round ℓ
- (b) Construct a version space V_t containing all languages $L \supseteq S_t$
- (c) Let $V'_t = \{L \in V_t: \forall j \in [t], \forall L' \in V_t \text{ it holds that } x_j \in L \implies x_j \in L'\}$
- (d) **if** $V'_t \neq \emptyset$ **then:** output the smallest index j such that L_j is in V'_t
- (e) **else:** output an arbitrary index j

Proof of Proposition 3.9. Let us first show that the previous algorithm identifies in the limit. Let $k := |\mathcal{L}|$ denote the size of \mathcal{L} . Consider any target language $K \in \mathcal{L}$. Notice that \mathcal{L} can be partitioned into three sets: the languages $L \in \mathcal{L}$ such that $L = K$, the languages $L \in \mathcal{L}$ such that $K \subsetneq L$ and

the languages $L \in \mathcal{L}$ such that $K \not\subseteq L$. Then, for every language $L_j \in \mathcal{L}$ such that $K \not\subseteq L_j$ there exists some $x \in K$ such that $x \notin L_j$. Let $i_j \in \mathbb{N}$ be the smallest number for which $x_{i_j} \in K, x_{i_j} \notin L_j$. Let $\mathcal{L}' \subseteq \mathcal{L}$ be the set of all such languages and \mathcal{X}' the set of all such smallest indexed $x \in \mathcal{X}$. Notice that, since we consider a fixed enumeration of \mathcal{X} throughout, the set \mathcal{X}' depends only on the target language K and the enumerations of \mathcal{L}, \mathcal{X} . Since the collection \mathcal{L} is finite we have that $|\mathcal{X}'| < k < \infty$.

Now consider any language $L \in \mathcal{L}$ such that $K \subsetneq L$. Let $\mathcal{L}'' \subseteq \mathcal{L}$ be the set of all such languages. Then, for every $L_j \in \mathcal{L}''$ there is some $x \in L_j$ such that $x \notin K$. Let i'_j be the smallest such index. Define

$$n_0 := \max_{j \in \mathbb{N}} \{i'_j : L_j \in \mathcal{L}''\},$$

i.e., the largest index among these elements. Since the collection \mathcal{L} is finite it holds that $n_0 < \infty$. Consider any execution of the algorithm in any round $t \in \mathbb{N}$ for which the input sample S satisfies **i)** $\mathcal{X}' \subseteq S$, and **ii)** $|S| \geq n_0$. The first condition on S implies that for this round $V_t = \{L \in \mathcal{L} : K \subseteq L\}$. Moreover, the second condition implies that $V'_t = \{L \in \mathcal{L} : L = K\}$. Thus, the smallest $j \in \mathbb{N}$ such that $L_j \in V'_t$ is the smallest index $z \in \mathbb{N}$ such that $L_z = K$. Thus, there is some large enough t^* so that the algorithm outputs the index z for all $t \geq t^*$. Moreover, the set \mathcal{X}' and the number n_0 satisfy the conditions of [Lemma 8.1](#), thus the algorithm identifies with exact exponential rates in the statistical setting. This concludes the proof. \square

8.3 Proof of [Proposition 3.10](#) (Identification of Collections of Finite Languages)

In this section, we give the proof of [Proposition 3.10](#).

Proof of [Proposition 3.10](#). Recall Gold's algorithm [[Gol67](#)] that identifies in the limit for such collections \mathcal{L} : at any step $n \in \mathbb{N}$ let S_n be the set of elements the adversary has presented so far. Output $\min \{j \in \mathbb{N} : S_n \subseteq L_j\}$. Consider any valid distribution \mathcal{P} with respect to \mathcal{L} . Then, \mathcal{P} is supported on some language $K \in \mathcal{L}$. Let $\{x_{i_1}, \dots, x_{i_k}\} := K$ and p_{i_j} be the mass of element $x_{i_j}, j \in [k]$. For every $n \in \mathbb{N}$, let \mathcal{E}_n be the event that the n i.i.d. draws from \mathcal{P} contain the set $\{x_{i_1}, \dots, x_{i_k}\}$, i.e.,

$$\{x_{i_1}, \dots, x_{i_k}\} = \{X_1, \dots, X_n\}.$$

Notice that under \mathcal{E}_n , the algorithm identifies correctly. Then, for the complement of this event, we have that

$$\begin{aligned} \Pr_{X_1, \dots, X_n \sim \mathcal{P}^n} [\exists j \in [k] : x_{i_j} \notin \{X_1, \dots, X_n\}] &\leq \sum_{j \in [k]} \Pr_{X_1, \dots, X_n \sim \mathcal{P}^n} [x_{i_j} \notin \{X_1, \dots, X_n\}] \quad (\text{by a union bound}) \\ &= \sum_{j \in [k]} (1 - p_{i_j})^n \quad (\text{since we have i.i.d. draws}) \\ &\leq \sum_{j \in [k]} e^{-p_{i_j} \cdot n} \quad (\text{using that } 1 - z \leq e^{-z} \text{ for all } z \in \mathbb{R}) \\ &\leq k \cdot e^{-\min_{j \in [k]} p_{i_j} \cdot n}. \end{aligned}$$

This concludes the proof. \square

8.4 Proof of Theorem 3.11 (Identification from Positive and Negative Examples)

We now move on to the task of language identification with both positive and negative examples. The main difference between this setting and binary classification is that the objective function is different. In particular, in our setting, the learner is required to *identify* the target language, whereas in the classification setting the learner is required to output a function that labels most of the elements of the domain according to some target labeling function. Thus, it is clear that the identification task is more challenging than the classification task. Sticking to the notation we used before, we have a countable set of languages $\mathcal{L} = \{L_1, L_2, \dots\}$, where each $L \in \mathcal{L}$ is also countable and $\cup_{L \in \mathcal{L}} L \subseteq \mathcal{X}$, for some countable domain \mathcal{X} . Recall the notion of valid distribution in this setting [Ang88]: a distribution \mathcal{P} is valid with respect to \mathcal{L} if and only if $\text{supp}(\mathcal{P}) \subseteq \mathcal{X} \times \{0, 1\}$ and there exists some $K \in \mathcal{L}$ such that for all $x \in K$ we have $\mathcal{P}[(x, 1)] > 0, \mathcal{P}[(x, 0)] = 0$ and for all $x \notin K$ we have $\mathcal{P}[(x, 0)] > 0, \mathcal{P}[(x, 1)] = 0$ (see Definition 15).

Next, recall that for any $n \in \mathbb{N}$ and set of labeled examples $S_n = (x_1, y_1), \dots, (x_n, y_n) \in (\mathcal{X} \times \{0, 1\})^n$ the error of the learner $\{h_n: (\mathcal{X} \times \{0, 1\})^n \rightarrow \mathbb{N}\}_{n \in \mathbb{N}}$ for this task is

$$\text{er}(h_n(S_n)) = \mathbb{1} \left\{ L_{h_n(S_n)} \neq K \right\}. \quad (11)$$

Notice that, under this definition, $\mathbb{E}[\text{er}(h_n)] = \Pr[L_{h_n} \neq K]$, i.e., the probability that h_n fails to identify the correct language after it sees n examples from the data-generating distribution.

Our proof proceeds in two parts. First, we show that for all countable collections of languages that are non-trivial for identification (Definition 13) exponential rate is the best possible for identification with positive and negative examples. The approach is essentially identical to Lemma 5.1 and the lower bound from Bousquet et al. [BHM+21]. Then, we show that all countable collections of languages are learnable at exponential rates with positive and negative examples.

The formal statement regarding the exponential rates lower bound follows.

Lemma 8.2. *Let \mathcal{L} be a non-trivial collection of languages. Then, for any learning algorithm $\mathcal{A} = \{h_n\}_{n \in \mathbb{N}}$ there exists a valid distribution \mathcal{P} such that $\mathbb{E}[\text{er}(h_n)] \geq C \cdot e^{-c \cdot n}$, for infinitely many $n \in \mathbb{N}$.*

Proof. Since \mathcal{L} is non-trivial, there exist $L, L' \in \mathcal{L}$ and $x \in \mathcal{X}$ such that $L \neq L'$ and $x \in L, x \in L'$. Let $\mathcal{P}_L, \mathcal{P}_{L'}$ be valid distributions for L, L' that place at least $1/2$ mass on $(x, 1)$ and they spread the remaining mass arbitrarily as follows: half of the remaining mass of \mathcal{P}_L (respectively $\mathcal{P}_{L'}$) is spread arbitrarily on all the elements of L (respectively L') with label 1, and the other half on all the elements of $K \setminus L$ (respectively $K \setminus L'$) with label 0. Notice that since $L \neq L'$ at least one of them has at least one more element other than x . For any $n \in \mathbb{N}$, under both distributions, with probability at least 2^{-n} , the algorithm will only see the element $(x, 1)$ appearing in the sample. Let \mathcal{E}_n be that event and condition on it. Notice that

$$\Pr \left[L_{h_n((x,1), \dots, (x,1))} = L \mid \mathcal{E}_n \right] + \Pr \left[L_{h_n((x,1), \dots, (x,1))} = L' \mid \mathcal{E}_n \right] \leq 1,$$

where the probability is with respect to the randomness of the learning algorithm. Thus, we have that $\Pr \left[L_{h_n((x,1), \dots, (x,1))} \neq L \mid \mathcal{E}_n \right] \geq 1/2$ or $\Pr \left[L_{h_n((x,1), \dots, (x,1))} \neq L' \mid \mathcal{E}_n \right] \geq 1/2$. By the pigeonhole principle, for at least one of L, L' , the previous inequality holds for infinitely many $n \in \mathbb{N}$. Assume

without loss of generality it holds for L and denote by \hat{N} the set of $n \in \mathbb{N}$ for which it holds. Then, for all $n \in \hat{N}$ we have that

$$\begin{aligned}
\mathbb{E}_{(X_1, Y_1), \dots, (X_n, Y_n) \sim \mathcal{P}_L^n} [\text{er}(h_n((X_1, Y_1), \dots, (X_n, Y_n)))] &= \Pr_{(X_1, Y_1), \dots, (X_n, Y_n) \sim \mathcal{P}_L^n} [L_{h_n}((X_1, Y_1), \dots, (X_n, Y_n)) \neq L] \\
&\geq \Pr_{(X_1, Y_1), \dots, (X_n, Y_n) \sim \mathcal{P}_L^n} [L_{h_n}((X_1, Y_1), \dots, (X_n, Y_n)) \neq L \mid \mathcal{E}_n] \cdot \\
&\quad \Pr_{(X_1, Y_1), \dots, (X_n, Y_n) \sim \mathcal{P}_L^n} [\mathcal{E}_n] \\
&\geq \frac{1}{2^n} \cdot \Pr [L_{h_n}((x, 1), \dots, (x, 1)) \neq L \mid \mathcal{E}_n] \\
&\quad \text{(by the definition of } \mathcal{E}_n \text{)} \\
&\geq \frac{1}{2^{n+1}}, \quad \text{(due to the assumption on } L \text{)}
\end{aligned}$$

which concludes the proof. \square

We now move on to establishing the upper bound. Following the approach from the setting with only positive examples, we show that an infinite draw of i.i.d. samples from any valid distribution is a “complete presentation” of the target language K , i.e., all the elements of K appear with label 1 and all elements outside of K appear with label 0. This follows immediately from [Proposition 5.2](#).

The next step towards proving [Theorem 3.11](#) is to show if \mathcal{A} is an algorithm that identifies the target language in the limit in the adversarial (online) setting of Gold [\[Gol67\]](#) with positive and negative examples, then \mathcal{A} is a consistent algorithm in the statistical setting. This implies that for any valid distribution \mathcal{P} , there is some number $t^* := t^*(\mathcal{A}, \mathcal{P}) \in \mathbb{N}$ such that, when we draw t^* many i.i.d. samples from \mathcal{P} , it will identify the target language with probability at least $6/7$. We denote the time of the last mistake of algorithm $\mathcal{A} = \{h_n\}_{n \in \mathbb{N}}$ on a labeled sequence of examples $(x_1, y_1), (x_2, y_2), \dots$ by $T_{\mathcal{A}}(x_1, y_1, x_2, y_2, \dots)$, i.e.,

$$T_{\mathcal{A}}(x_1, y_1, x_2, y_2, \dots) = \inf \left\{ n_0 \in \mathbb{N} : L_{h_n}(x_1, y_1, \dots, x_n, y_n) = K, \forall n > n_0 \right\}.$$

The next result formalizes the claim. Its proof is almost identical to [Proposition 5.3](#), but we present it for completeness.

Proposition 8.3. *Fix any family of languages \mathcal{L} over a countable domain. For any algorithm $\mathcal{A} = \{h_n\}_{n \in \mathbb{N}}$ that identifies \mathcal{L} in the limit with positive and negative examples in the online setting and any valid distribution \mathcal{P} ([Definition 15](#)), there exists a number t^* such that*

$$\Pr_{\{(X_i, Y_i)\}_{i \in \mathbb{N}} \sim \mathcal{P}^\infty} [T_{\mathcal{A}}(X_1, Y_1, X_2, Y_2, \dots) \leq t^*] \geq \frac{6}{7}.$$

Proof. Let $(X_1, Y_1), (X_2, Y_2), \dots$, be a countable i.i.d. sample from \mathcal{P} . From [Proposition 5.2](#) we get that this sample is a valid input to \mathcal{A} since, with probability one, all elements of K appear with label 1 and all elements of $\mathcal{X} \setminus K$ appear with label 0. Consider the execution of \mathcal{A} on prefixes of the sequence and denote by $T_{\mathcal{A}} := T_{\mathcal{A}}(X_1, Y_1, X_2, Y_2, \dots)$ the time it made its last mistake. We have that $\Pr_{\{(X_i, Y_i)\}_{i \in \mathbb{N}} \sim \mathcal{P}^\infty} [T_{\mathcal{A}} \in \mathbb{N}] = 1$. Thus,

$$\lim_{t \rightarrow \infty} \Pr_{\{(X_i, Y_i)\}_{i \in \mathbb{N}} \sim \mathcal{P}^\infty} [T_{\mathcal{A}}(X_1, Y_1, X_2, Y_2, \dots) \geq t] = 0.$$

Thus, as required, there exists some $t^* \in \mathbb{N}$ such that

$$\Pr_{\{(X_i, Y_i)\}_{i \in \mathbb{N}} \sim \mathcal{P}^\infty} [T_{\mathcal{A}}(X_1, Y_1, X_2, Y_2, \dots) \geq t^*] \leq \frac{1}{7}.$$

□

The problem is that this time t^* depends on the algorithm \mathcal{A} and the unknown distribution \mathcal{P} . Suppose we knew a number t^* so that for all $t \geq t^*$

$$\mathbb{E}_{S \sim \mathcal{P}^t} [\text{er}(h_t(S)) > 0] = \Pr_{S \sim \mathcal{P}^t} [L_{h_t(S)} \neq K] < \frac{1}{4}.$$

Then, we could design an identification algorithm $\{h_n\}_{n \in \mathbb{N}}$ with exponential rates as follows. First, we break up the data into batches, each of length t^* . Second, we run the identification algorithm from the online setting separately for each batch. Finally, we aggregate these algorithms by taking a majority vote. Now, by the definition of t^* and Hoeffding's inequality, the probability that more than one-third of the classifiers have not identified the language is exponentially small.

There are two issues with this approach:

- In our language setting, it can be the case that two identification algorithms are correct but output different indices for the true language. This was also an issue for the positive examples case and we can resolve it in a similar way using [Lemma 5.4](#).
- The second issue is that t^* depends on the distribution \mathcal{P} . In the previous case of positive examples, the solution was to *guess* the number t^* using some very slowly increasing function $g(n)$. This was the reason why we did not manage to get exponential rates. Interestingly, in the setting where we have access to positive and negative examples, we can *estimate* t^* from samples, as we see below.

The main lemma behind this result is the following and corresponds to an adaptation of Lemma 4.4 from Bousquet et al. [[BHM+21](#)] with a different loss function. This lemma will allow us to get exactly exponential rates, compared to the rates obtained in the positive examples regime.

Lemma 8.4 (Estimation of Stopping Time). *Let S_n be n i.i.d. examples observed from \mathcal{P} which is valid with respect to some language $K \in \mathcal{L}$. There exists universally measurable $t_n = t_n(S_n)$, whose definition does not depend on \mathcal{P} , so that the following holds. Given t^* such that*

$$\Pr_{S_{t^*}} [L_{h_{t^*}(S_{t^*})} \neq K] < \frac{1}{8},$$

there exist $C, c > 0$ independent of n (but depending on \mathcal{P}, t^) so that*

$$\Pr[t_n \in \mathcal{T}] \geq 1 - Ce^{-cn}$$

where

$$\mathcal{T} = \left\{ 1 \leq t \leq t^* : \Pr_{S_t} [L_{h_t(S_t)} \neq K] < \frac{3}{8} \right\}.$$

Given the above lemma, we are now ready to complete the proof of [Theorem 3.11](#).

Proof of Theorem 3.11. First, the exponential rate lower bound follows immediately from Lemma 8.2.

The output of our identification algorithm h_n is the majority of the $h_{t_n}^i$ for $i \in I_n = \{1, 2, \dots, \lfloor n/(2t_n) \rfloor\}$, after we apply to them the post-processing computation described in Lemma 5.4 to map them to the same index of K . Let $z \in \mathbb{N}$ be the smallest number for which $L_z = K$. It remains to show that

$$\mathbb{E}_{S_n \sim \mathcal{P}^n} [\text{er}(h_n)] = \Pr_{S_n \sim \mathcal{P}^n} [L_{h_n(S_n)} \neq K] \leq Ce^{-cn},$$

for some constants $C, c > 0$ depending on \mathcal{P} .

Let us consider our predictors $\{h_{t_n}^i\}_{i \in I_n}$ that are obtained by running the identification algorithm on independent parts of the dataset of size t_n . Since these predictors might be outputting different indices (descriptions) of the same language, we find the smallest indexed language the output of each classifier can be mapped to. By Lemma 5.4, for n sufficiently large, all the indices from the outputs of the classifiers $h_{t_n}^i, i \in I_n$, that correspond to K will be mapped to z .

Let us fix $t \in \mathcal{T}$, where \mathcal{T} is defined in Lemma 8.4, and consider the predictors $\{h_t^i\}$ for $i \in I = \{1, 2, \dots, \lfloor n/(2t) \rfloor\}$. We also denote by $\{\hat{h}_t^i\}$ the output of predictor $\{h_t^i\}$ after applying the post-processing result from Lemma 8.4. For n sufficiently large, standard concentration bounds give that

$$\Pr \left[\frac{1}{|I_n|} \sum_{i \in I_n} \mathbb{1}\{\hat{h}_t^i \neq z\} > \frac{7}{16} \right] < e^{-\lfloor n/2t^* \rfloor / 128}.$$

This means that, except on an event of exponentially small probability, we have that \hat{h}_t^i outputs index z . Recall that $L_z = K$.

Now we employ our estimation t_n in the above calculation. In particular,

$$\Pr [\hat{h}_{t_n}^i \neq z \text{ for at least half of } i \in I_n] \leq p_1 + p_2,$$

where

$$p_1 := \Pr[t_n \notin \mathcal{T}] < Ce^{-cn},$$

and

$$p_2 := \Pr [\exists t \in \mathcal{T}: h_t^i \text{ does not predict } K \text{ for at least half indices}] \leq t^* e^{-\lfloor n/2t^* \rfloor / 128}.$$

This gives the desired result. \square

We conclude this section with the proof of Lemma 8.4.

Proof of Lemma 8.4. We split the training set S_n into two sets and we further split the sets into batches. The idea is to use the first set to train multiple independent instances of the online learning algorithm and the second set to estimate its identification error.

More concretely, let \mathcal{A} the algorithm that identifies in the limit. For each batch size $1 \leq t \leq \lfloor n/2 \rfloor$ and batch index $1 \leq i \leq \lfloor n/(2t) \rfloor$, we let

$$I_t^i = h_t \left(X_{(i-1)t+1}, Y_{(i-1)t+1}, X_{it}, Y_{it} \right)$$

be the index of the output of the learning algorithm that is trained on batch i of a subset of the dataset that has size t . This is a mapping from the training samples to indices of the predicted language.

For every fixed t , the outputs $\{I_t^i\}_{i \leq \lfloor n/2t \rfloor}$ are trained on different parts of the first half of the training set and they are independent of the whole second half of the training set. This means that we can view every $\{I_t^i\}_{i \leq \lfloor n/2t \rfloor}$ as an independent draw of the distribution of $h_t(\cdot)$. To estimate the identification error of $h_t(\cdot)$, we will make use of the second half of the training set. We define

$$\hat{e}_t = \frac{1}{\lfloor n/2t \rfloor} \sum_{i=1}^{\lfloor n/2t \rfloor} \mathbb{1} \left\{ \mathbb{1} \left\{ X_s \in L_{I_t^i} \right\} \neq Y_s \text{ for some } \frac{n}{2} \leq s \leq n \right\}.$$

We underline that this can be computed using just membership calls to the predicted languages. Now observe that, almost surely,

$$\hat{e}_t \leq e_t = \frac{1}{\lfloor n/2t \rfloor} \sum_{i=1}^{\lfloor n/2t \rfloor} \mathbb{1} \left\{ \Pr_{(X,Y) \sim \mathcal{P}} \left[\mathbb{1} \left\{ X \in L_{I_t^i} \right\} \neq Y \right] > 0 \right\} = \frac{1}{\lfloor n/2t \rfloor} \sum_{i=1}^{\lfloor n/2t \rfloor} \mathbb{1} \left\{ L_{I_t^i} \neq K \right\}.$$

We define $\hat{t}_n = \inf\{t \leq \lfloor n/2 \rfloor : \hat{e}_t < 1/4\}$, where we assume that $\inf \emptyset = \infty$.

We now want to bound the probability that $\hat{t}_n > t^*$. Using Hoeffding's inequality we get that

$$\Pr[\hat{t}_n > t^*] \leq \Pr\left[\hat{e}_{t^*} \geq \frac{1}{4}\right] \leq \Pr\left[e_{t^*} \geq \frac{1}{4}\right] = \Pr\left[e_{t^*} - \frac{1}{8} \geq \frac{1}{8}\right] = \Pr\left[e_{t^*} - \mathbb{E}[e_{t^*}] \geq \frac{1}{8}\right] \leq e^{-\lfloor n/2t^* \rfloor / 32}.$$

This implies that $\hat{t}_n \leq t^*$ except for an event with exponentially small probability.

Moreover, there is some $\varepsilon > 0$ such that for all $1 \leq t \leq t^*$ with

$$\Pr_{S_t \sim \mathcal{P}^t} \left[\Pr_{(X,Y) \sim \mathcal{P}} \left[\mathbb{1} \left\{ X \in L_{h_t(S_t)} \right\} \neq Y \right] > 0 \right] > \frac{3}{8},$$

we have that $\Pr_{S_t \sim \mathcal{P}^t} \left[\Pr_{(X,Y) \sim \mathcal{P}} \left[\mathbb{1} \left\{ X \in L_{h_t(S_t)} \right\} \neq Y \right] > \varepsilon \right] > 1/4 + 1/16$ (this holds by continuity).

Now fix some $1 \leq t \leq t^*$ such that

$$\Pr_{S_t \sim \mathcal{P}^t} \left[\Pr_{(X,Y) \sim \mathcal{P}} \left[\mathbb{1} \left\{ X \in L_{h_t(S_t)} \right\} \neq Y \right] > 0 \right] > \frac{3}{8}$$

(if it exists). Then, using Hoeffding's inequality again we get that

$$\Pr \left[\frac{1}{\lfloor n/2t \rfloor} \sum_{i=1}^{\lfloor n/2t \rfloor} \mathbb{1} \left\{ \Pr_{(X,Y) \sim \mathcal{P}} \left[\mathbb{1} \left\{ X \in L_{I_t^i} \right\} \neq Y \right] > \varepsilon \right\} < \frac{1}{4} \right] \leq e^{\lfloor n/2t^* \rfloor / 128}.$$

For any language L such that $\Pr_{(X,Y) \sim \mathcal{P}} [\mathbb{1} \{X \in L\} \neq Y] > \varepsilon$, then

$$\Pr [\mathbb{1} \{X_s \in L\} \neq Y_s \text{ for some } n/2 \leq s \leq n] \geq 1 - (1 - \varepsilon)^{n/2}.$$

As we mentioned before, $\{I_t^i\}_{i \leq \lfloor n/2t \rfloor}$ are independent of $(X_s, Y_s)_{s > n/2}$. Thus, applying a union bound we get that the probability that all I_t^i that have $\Pr_{(X,Y) \sim \mathcal{P}} [\mathbb{1} \{X \in L_{I_t^i}\} \neq Y] > \varepsilon$ make at least one error on the second half of the training set is

$$\begin{aligned} \Pr \left[\mathbb{1} \left\{ \Pr_{(X,Y) \sim \mathcal{P}} \left[\mathbb{1} \left\{ X \in L_{I_t^i} \right\} \neq Y \right] > \varepsilon \right\} \leq \mathbb{1} \left\{ \mathbb{1} \left\{ X_s \in L_{I_t^i} \right\} \neq Y_s \text{ for some } n/2 < s \leq n \right\} \text{ for all } i \in [\lfloor n/2t \rfloor] \right] \\ \geq 1 - \left\lfloor \frac{n}{2t} \right\rfloor (1 - \varepsilon)^{n/2}. \end{aligned}$$

Thus, we get that

$$\Pr[\widehat{t}_n = t] \leq \Pr\left[\widehat{e}_t < \frac{1}{4}\right] \leq \left\lfloor \frac{n}{2} \right\rfloor (1 - \varepsilon)^{n/2} + e^{-\lfloor \frac{n}{2t^*} \rfloor / 128}.$$

Using the previous estimates and applying a union bound, we get that

$$\Pr[\widehat{t}_n \notin \mathcal{T}] \leq e^{-\lfloor n/2t^* \rfloor / 32} + t^* \left\lfloor \frac{n}{2} \right\rfloor (1 - \varepsilon)^{n/2} + t^* e^{-\lfloor n/2t^* \rfloor / 128} \leq Ce^{-cn},$$

for some constants $C, c > 0$. Note that $C = C(\mathcal{P}, t^*)$ and $c = c(\mathcal{P}, t^*)$. □

Acknowledgments

We thank Ahmad Beirami and Manolis Zampetakis for helpful discussions and references after the original draft. We thank Dylan McKay for discussions and references about the theory of computation during the preparation of this paper, and specifically for informing us about [Lemmas 6.2](#) and [6.3](#). We thank Kyriakos Lotidis for useful discussions about the lower bound of Angluin [\[Ang88\]](#). We also thank Yuan Deng, Sid Mitra, Ansong Ni, Argyris Oikonomou, Xizhi Tan, and Manolis Zampetakis for their feedback on a draft of this paper. Alkis Kalavasis was supported by the Institute for Foundations of Data Science at Yale. Grigoris Velegkas was supported by the AI Institute for Learning-Enabled Optimization at Scale (TILOS).

References

- [AB09] S. Arora and B. Barak. *Computational Complexity: A Modern Approach*. Cambridge University Press, 2009. ISBN: 9781139477369. URL: <https://books.google.com/books?id=nGvI7c0u00QC> (cit. on p. 29).
- [AB17] Martin Arjovsky and Leon Bottou. “Towards Principled Methods for Training Generative Adversarial Networks”. In: *International Conference on Learning Representations*. 2017. URL: https://openreview.net/forum?id=Hk4_qw5xe (cit. on pp. 1, 3, 5).
- [AB91] Leonard M Adleman and Manuel Blum. “Inductive Inference and Unsolvability”. In: *The Journal of Symbolic Logic* 56.3 (1991), pp. 891–900. DOI: [10.2307/2275058](https://doi.org/10.2307/2275058) (cit. on p. 5).
- [ABL+22] Noga Alon, Mark Bun, Roi Livni, Maryanthe Malliaris, and Shay Moran. “Private and Online Learnability Are Equivalent”. In: *J. ACM* 69.4 (Aug. 2022). ISSN: 0004-5411. DOI: [10.1145/3526074](https://doi.org/10.1145/3526074). URL: <https://doi.org/10.1145/3526074> (cit. on p. 18).
- [AGR13] Joseph Anderson, Navin Goyal, and Luis Rademacher. “Efficient Learning of Simplices”. In: *Proceedings of the 26th Annual Conference on Learning Theory*. Ed. by Shai Shalev-Shwartz and Ingo Steinwart. Vol. 30. Proceedings of Machine Learning Research. Princeton, NJ, USA: PMLR, June 2013, pp. 1020–1045. URL: <https://proceedings.mlr.press/v30/Anderson13.html> (cit. on p. 19).
- [AHK+24] Idan Attias, Steve Hanneke, Alkis Kalavasis, Amin Karbasi, and Grigoris Velegkas. “Universal Rates for Regression: Separations between Cut-Off and Absolute Loss”. In: *Proceedings of Thirty Seventh Conference on Learning Theory*. Ed. by Shipra Agrawal and Aaron Roth. Vol. 247. Proceedings of Machine Learning Research. PMLR, June 2024, pp. 359–405. URL: <https://proceedings.mlr.press/v247/attias24a.html> (cit. on pp. 9, 16).
- [AL24] Zeyuan Allen-Zhu and Yuanzhi Li. *Physics of Language Models: Part 1, Learning Hierarchical Language Structures*. 2024. arXiv: [2305.13673](https://arxiv.org/abs/2305.13673) [cs.CL]. URL: <https://arxiv.org/abs/2305.13673> (cit. on p. 17).
- [AL98] András Antos and Gábor Lugosi. “Strong Minimax Lower Bounds for Learning”. In: *Machine Learning* 30.1 (1998), pp. 31–56. DOI: [10.1023/A:1007454427662](https://doi.org/10.1023/A:1007454427662). URL: <https://doi.org/10.1023/A:1007454427662> (cit. on pp. 4, 24).
- [AML+24] Sumukh K Aithal, Pratyush Maini, Zachary C. Lipton, and J. Zico Kolter. *Understanding Hallucinations in Diffusion Models through Mode Interpolation*. 2024. arXiv: [2406.09358](https://arxiv.org/abs/2406.09358) [cs.LG]. URL: <https://arxiv.org/abs/2406.09358> (cit. on p. 16).
- [Ang79] Dana Angluin. “Finding Patterns Common to a Set of Strings (Extended Abstract)”. In: *Proceedings of the Eleventh Annual ACM Symposium on Theory of Computing*. STOC ’79. Atlanta, Georgia, USA: Association for Computing Machinery, 1979, pp. 130–141. ISBN: 9781450374385. DOI: [10.1145/800135.804406](https://doi.org/10.1145/800135.804406). URL: <https://doi.org/10.1145/800135.804406> (cit. on pp. 1, 3, 5–7, 10, 14, 17, 20, 21, 30, 96).

- [Ang80] Dana Angluin. “Inductive Inference of Formal Languages From Positive Data”. In: *Information and Control* 45.2 (1980), pp. 117–135. ISSN: 0019-9958. DOI: [https://doi.org/10.1016/S0019-9958\(80\)90285-5](https://doi.org/10.1016/S0019-9958(80)90285-5). URL: <https://www.sciencedirect.com/science/article/pii/S0019995880902855> (cit. on pp. 3, 6, 7, 9, 14, 18–21, 28, 30, 33, 90).
- [Ang82] Dana Angluin. “Inference of Reversible Languages”. In: *J. ACM* 29.3 (July 1982), pp. 741–765. ISSN: 0004-5411. DOI: [10.1145/322326.322334](https://doi.org/10.1145/322326.322334). URL: <https://doi.org/10.1145/322326.322334> (cit. on p. 17).
- [Ang88] Dana Angluin. *Identifying Languages From Stochastic Examples*. Yale University. Department of Computer Science, 1988. URL: <http://www.cs.yale.edu/publications/techreports/tr614.pdf> (cit. on pp. 1–4, 8, 10, 14, 16–18, 24, 26, 33, 41, 63, 69, 74).
- [AOS+16] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. *Concrete Problems in AI Safety*. 2016. arXiv: [1606.06565](https://arxiv.org/abs/1606.06565) [cs.AI]. URL: <https://arxiv.org/abs/1606.06565> (cit. on p. 1).
- [AP23] Konstantinos Andriopoulos and Johan Pouwelse. *Augmenting LLMs with Knowledge: A Survey on Hallucination Prevention*. 2023. arXiv: [2309.16459](https://arxiv.org/abs/2309.16459) [cs.CL]. URL: <https://arxiv.org/abs/2309.16459> (cit. on p. 1).
- [AS83] Dana Angluin and Carl H. Smith. “Inductive Inference: Theory and Methods”. In: *ACM Comput. Surv.* 15.3 (Sept. 1983), pp. 237–269. ISSN: 0360-0300. DOI: [10.1145/356914.356918](https://doi.org/10.1145/356914.356918). URL: <https://doi.org/10.1145/356914.356918> (cit. on pp. 5, 17).
- [AWK+24] Ekin Akyürek, Bailin Wang, Yoon Kim, and Jacob Andreas. “In-Context Language Learning: Architectures and Algorithms”. In: *Forty-first International Conference on Machine Learning*. 2024. URL: <https://openreview.net/forum?id=3Z9CRr5srL> (cit. on p. 17).
- [BAG20] Satwik Bhattamishra, Kabir Ahuja, and Navin Goyal. “On the Ability and Limitations of Transformers to Recognize Formal Languages”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Ed. by Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu. Online: Association for Computational Linguistics, Nov. 2020, pp. 7096–7116. DOI: [10.18653/v1/2020.emnlp-main.576](https://aclanthology.org/2020.emnlp-main.576). URL: <https://aclanthology.org/2020.emnlp-main.576> (cit. on p. 17).
- [Bah14] Dzmitry Bahdanau. “Neural machine translation by jointly learning to align and translate”. In: *arXiv preprint arXiv:1409.0473* (2014) (cit. on p. 1).
- [BB75] Lenore Blum and Manuel Blum. “Toward a Mathematical Theory of Inductive Inference”. In: *Information and Control* 28.2 (1975), pp. 125–155. ISSN: 0019-9958. DOI: [https://doi.org/10.1016/S0019-9958\(75\)90261-2](https://doi.org/10.1016/S0019-9958(75)90261-2). URL: <https://www.sciencedirect.com/science/article/pii/S0019995875902612> (cit. on p. 5).
- [BCE+23] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. *Sparks of Artificial General Intelligence: Early experiments with GPT-4*. 2023. arXiv: [2303.12712](https://arxiv.org/abs/2303.12712) [cs.CL]. URL: <https://arxiv.org/abs/2303.12712> (cit. on pp. 1, 6).

- [BCP+90] Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. “A statistical approach to machine translation”. In: *Comput. Linguist.* 16.2 (June 1990), pp. 79–85. ISSN: 0891-2017 (cit. on p. 6).
- [BDd+92] Peter F. Brown, Vincent J. Della Pietra, Peter V. deSouza, Jenifer C. Lai, and Robert L. Mercer. “Class-Based n -gram Models of Natural Language”. In: *Computational Linguistics* 18.4 (1992), pp. 467–480. URL: <https://aclanthology.org/J92-4003> (cit. on p. 1).
- [BDV00] Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. “A Neural Probabilistic Language Model”. In: *Advances in Neural Information Processing Systems*. Ed. by T. Leen, T. Dietterich, and V. Tresp. Vol. 13. MIT Press, 2000. URL: https://proceedings.neurips.cc/paper_files/paper/2000/file/728f206c2a01bf572b5940d7d9a8fa4c-Paper.pdf (cit. on p. 1).
- [Ber86] Robert C Berwick. “Learning From Positive-Only Examples: The Subset Principle and Three Case Studies”. In: *Machine learning: An artificial intelligence approach 2* (1986), pp. 625–645 (cit. on p. 18).
- [BGH+23] Mark Bun, Marco Gaboardi, Max Hopkins, Russell Impagliazzo, Rex Lei, Toniann Pitassi, Satchit Sivakumar, and Jessica Sorrell. “Stability Is Stable: Connections between Replicability, Privacy, and Adaptive Generalization”. In: *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*. STOC 2023. Orlando, FL, USA: Association for Computing Machinery, 2023, pp. 520–527. ISBN: 9781450399135. DOI: 10.1145/3564246.3585246. URL: <https://doi.org/10.1145/3564246.3585246> (cit. on p. 18).
- [BHM+21] Olivier Bousquet, Steve Hanneke, Shay Moran, Ramon van Handel, and Amir Yehudayoff. “A Theory of Universal Learning”. In: *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*. STOC 2021. Virtual, Italy: Association for Computing Machinery, 2021, pp. 532–541. ISBN: 9781450380539. DOI: 10.1145/3406325.3451087. URL: <https://doi.org/10.1145/3406325.3451087> (cit. on pp. 1, 4, 9, 14, 16–18, 24–26, 34, 51, 69, 71, 96–98).
- [BHM+23] Olivier Bousquet, Steve Hanneke, Shay Moran, Jonathan Shafer, and Ilya Tolstikhin. “Fine-Grained Distribution-Dependent Learning Curves”. In: *Proceedings of Thirty Sixth Conference on Learning Theory*. Ed. by Gergely Neu and Lorenzo Rosasco. Vol. 195. Proceedings of Machine Learning Research. PMLR, July 2023, pp. 5890–5924. URL: <https://proceedings.mlr.press/v195/bousquet23a.html> (cit. on p. 16).
- [Bid23] Joseph R Biden. “Executive Order on the Safe, Secure, and Trustworthy Development And Use of Artificial Intelligence””. In: (2023). URL: <https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/> (cit. on p. 1).

- [BJM83] Lalit R. Bahl, Frederick Jelinek, and Robert L. Mercer. “A Maximum Likelihood Approach to Continuous Speech Recognition”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-5.2 (1983), pp. 179–190. DOI: [10.1109/TPAMI.1983.4767370](https://doi.org/10.1109/TPAMI.1983.4767370) (cit. on p. 6).
- [Bor23] Ali Borji. *A Categorical Archive of ChatGPT Failures*. 2023. arXiv: [2302.03494](https://arxiv.org/abs/2302.03494) [cs.CL]. URL: <https://arxiv.org/abs/2302.03494> (cit. on p. 1).
- [Bre07] Joan Bresnan. “Is Syntactic Knowledge Probabilistic? Experiments With the English Dative Alternation”. In: *Linguistics in Search of its Evidential Base*. Ed. by Sam Featherston and Wolfgang Sternefeld. Berlin, New York: De Gruyter Mouton, 2007, pp. 75–96. ISBN: 9783110198621. DOI: [doi:10.1515/9783110198621.75](https://doi.org/10.1515/9783110198621.75). URL: <https://doi.org/10.1515/9783110198621.75> (cit. on p. 1).
- [BZW+19] David Bau, Jun-Yan Zhu, Jonas Wulff, William Peebles, Hendrik Strobelt, Bolei Zhou, and Antonio Torralba. “Seeing What a GAN Cannot Generate”. In: *Proceedings of the International Conference Computer Vision (ICCV)*. 2019 (cit. on p. 1).
- [Cla14] Alexander Clark. “Distributional Learning as a Theory of Language Acquisition”. In: *Proceedings of the 5th Workshop on Cognitive Aspects of Computational Language Learning (CogACLL)*. Ed. by Alessandro Lenci, Muntsa Padró, Thierry Poibeau, and Aline Villavicencio. Gothenburg, Sweden: Association for Computational Linguistics, Apr. 2014, p. 29. DOI: [10.3115/v1/W14-0506](https://doi.org/10.3115/v1/W14-0506). URL: <https://aclanthology.org/W14-0506> (cit. on p. 1).
- [CMY23] Zachary Chase, Shay Moran, and Amir Yehudayoff. “Stability and Replicability in Learning”. In: *2023 IEEE 64th Annual Symposium on Foundations of Computer Science (FOCS)*. 2023, pp. 2430–2439. DOI: [10.1109/FOCS57990.2023.00148](https://doi.org/10.1109/FOCS57990.2023.00148) (cit. on p. 18).
- [Daw82] A Philip Dawid. “The Well-Calibrated Bayesian”. In: *Journal of the American Statistical Association* 77.379 (1982), pp. 605–610. DOI: [10.1080/01621459.1982.10477856](https://doi.org/10.1080/01621459.1982.10477856). eprint: <https://www.tandfonline.com/doi/pdf/10.1080/01621459.1982.10477856>. URL: <https://www.tandfonline.com/doi/abs/10.1080/01621459.1982.10477856> (cit. on p. 16).
- [DDS15] Anindya De, Ilias Diakonikolas, and Rocco A. Servedio. “Learning From Satisfying Assignments”. In: *Proceedings of the 2015 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*. 2015, pp. 478–497. DOI: [10.1137/1.9781611973730.33](https://doi.org/10.1137/1.9781611973730.33). eprint: <https://epubs.siam.org/doi/pdf/10.1137/1.9781611973730.33>. URL: <https://epubs.siam.org/doi/abs/10.1137/1.9781611973730.33> (cit. on p. 19).
- [Den98] François Denis. “PAC Learning from Positive Statistical Queries”. In: *Algorithmic Learning Theory*. Ed. by Michael M. Richter, Carl H. Smith, Rolf Wiehagen, and Thomas Zeugmann. Berlin, Heidelberg: Springer Berlin Heidelberg, 1998, pp. 112–126. ISBN: 978-3-540-49730-1 (cit. on p. 18).
- [DGL05] François Denis, Rémi Gilleron, and Fabien Letouzey. “Learning From Positive and Unlabeled Examples”. In: *Theoretical Computer Science* 348.1 (2005). Algorithmic Learning Theory (ALT 2000), pp. 70–83. ISSN: 0304-3975. DOI: <https://doi.org/10.1016/j.tcs.2005.09.007>. URL: <https://www.sciencedirect.com/science/article/pii/S0304397505005256> (cit. on p. 18).

- [DGT+18] Constantinos Daskalakis, Themis Gouleakis, Christos Tzamos, and Manolis Zampetakis. “Efficient Statistics, in High Dimensions, from Truncated Samples”. In: *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*. 2018, pp. 639–649. DOI: [10.1109/FOCS.2018.00067](https://doi.org/10.1109/FOCS.2018.00067) (cit. on p. 19).
- [DGT+19] Constantinos Daskalakis, Themis Gouleakis, Christos Tzamos, and Manolis Zampetakis. “Computationally and Statistically Efficient Truncated Regression”. In: *Proceedings of the Thirty-Second Conference on Learning Theory*. Ed. by Alina Beygelzimer and Daniel Hsu. Vol. 99. Proceedings of Machine Learning Research. PMLR, June 2019, pp. 955–960. URL: <https://proceedings.mlr.press/v99/daskalakis19a.html> (cit. on p. 19).
- [DKP+24] Ilias Diakonikolas, Daniel M. Kane, Thanasis Pittas, and Nikos Zarifis. “Statistical Query Lower Bounds for Learning Truncated Gaussians”. In: *Proceedings of Thirty Seventh Conference on Learning Theory*. Ed. by Shipra Agrawal and Aaron Roth. Vol. 247. Proceedings of Machine Learning Research. PMLR, 30 Jun–03 Jul 2024, pp. 1336–1363. URL: <https://proceedings.mlr.press/v247/diakonikolas24b.html> (cit. on p. 19).
- [DKT+21] Constantinos Daskalakis, Vasilis Kontonis, Christos Tzamos, and Emmanouil Zampetakis. “A Statistical Taylor Theorem and Extrapolation of Truncated Densities”. In: *Proceedings of Thirty Fourth Conference on Learning Theory*. Ed. by Mikhail Belkin and Samory Kpotufe. Vol. 134. Proceedings of Machine Learning Research. PMLR, Aug. 2021, pp. 1395–1398. URL: <https://proceedings.mlr.press/v134/daskalakis21a.html> (cit. on p. 19).
- [DLN+24] Anindya De, Huan Li, Shivam Nadimpalli, and Rocco A. Servedio. “Detecting Low-Degree Truncation”. In: *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*. STOC 2024. Vancouver, BC, Canada: Association for Computing Machinery, 2024, pp. 1027–1038. ISBN: 9798400703836. DOI: [10.1145/3618260.3649633](https://doi.org/10.1145/3618260.3649633). URL: <https://doi.org/10.1145/3618260.3649633> (cit. on p. 19).
- [DNS23] Anindya De, Shivam Nadimpalli, and Rocco A. Servedio. “Testing Convex Truncation”. In: *Proceedings of the 2023 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*. 2023, pp. 4050–4082. DOI: [10.1137/1.9781611977554.ch155](https://doi.org/10.1137/1.9781611977554.ch155). eprint: <https://epubs.siam.org/doi/pdf/10.1137/1.9781611977554.ch155>. URL: <https://epubs.siam.org/doi/abs/10.1137/1.9781611977554.ch155> (cit. on p. 19).
- [DP09] D.P. Dubhashi and A. Panconesi. *Concentration of Measure for the Analysis of Randomized Algorithms*. Cambridge University Press, 2009. ISBN: 9781139480994. URL: <https://books.google.com/books?id=UUohAwAAQBAJ> (cit. on p. 40).
- [EEG+24] Benjamin L. Edelman, Ezra Edelman, Surbhi Goel, Eran Malach, and Nikolaos Tsilivis. *The Evolution of Statistical Induction Heads: In-Context Learning Markov Chains*. 2024. arXiv: [2402.11004](https://arxiv.org/abs/2402.11004) [cs.LG]. URL: <https://arxiv.org/abs/2402.11004> (cit. on p. 17).

- [EGZ20] Javid Ebrahimi, Dhruv Gelda, and Wei Zhang. “How Can Self-Attention Networks Recognize Dyck-n Languages?” In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Ed. by Trevor Cohn, Yulan He, and Yang Liu. Online: Association for Computational Linguistics, Nov. 2020, pp. 4301–4306. DOI: [10.18653/v1/2020.findings-emnlp.384](https://doi.org/10.18653/v1/2020.findings-emnlp.384). URL: <https://aclanthology.org/2020.findings-emnlp.384> (cit. on p. 17).
- [Eld11] Ronen Eldan. “A Polynomial Number of Random Points Does Not Determine the Volume of a Convex Body”. In: *Discrete & Computational Geometry* 46.1 (2011), pp. 29–47. DOI: [10.1007/s00454-011-9328-x](https://doi.org/10.1007/s00454-011-9328-x). URL: <https://doi.org/10.1007/s00454-011-9328-x> (cit. on p. 19).
- [Elm90] Jeffrey L. Elman. “Finding Structure in Time”. In: *Cognitive Science* 14.2 (1990), pp. 179–211. DOI: https://doi.org/10.1207/s15516709cog1402_1. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1207/s15516709cog1402_1. URL: https://onlinelibrary.wiley.com/doi/abs/10.1207/s15516709cog1402_1 (cit. on p. 17).
- [FJK96] A. Frieze, M. Jerrum, and R. Kannan. “Learning Linear Transformations”. In: *Proceedings of 37th Conference on Foundations of Computer Science*. 1996, pp. 359–368. DOI: [10.1109/SFCS.1996.548495](https://doi.org/10.1109/SFCS.1996.548495) (cit. on p. 19).
- [FKT20] Dimitris Fotakis, Alkis Kalavasis, and Christos Tzamos. “Efficient parameter estimation of truncated boolean product distributions”. In: *Conference on learning theory*. PMLR. 2020, pp. 1586–1600. URL: <https://doi.org/10.1007/s00453-022-00961-9> (cit. on p. 19).
- [FSW+24] Shangbin Feng, Weijia Shi, Yike Wang, Wenxuan Ding, Vidhisha Balachandran, and Yulia Tsvetkov. “Don’t Hallucinate, Abstain: Identifying LLM Knowledge Gaps via Multi-LLM Collaboration”. In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Lun-Wei Ku, Andre Martins, and Vivek Srikumar. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 14664–14690. URL: <https://aclanthology.org/2024.acl-long.786> (cit. on p. 1).
- [GLL+24] Xinyan Guan, Yanjiang Liu, Hongyu Lin, Yaojie Lu, Ben He, Xianpei Han, and Le Sun. “Mitigating Large Language Model Hallucinations via Autonomous Knowledge Graph-Based Retrofitting”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 38. 16. 2024, pp. 18126–18134. URL: <https://doi.org/10.1609/aaai.v38i16.29770> (cit. on p. 16).
- [Gol16] Yoav Goldberg. “A Primer on Neural Network Models for Natural Language Processing”. In: *Journal of Artificial Intelligence Research* 57 (2016), pp. 345–420. DOI: <https://doi.org/10.1613/jair.4992> (cit. on p. 1).
- [Gol67] E. Mark Gold. “Language Identification in the Limit”. In: *Information and Control* 10.5 (1967), pp. 447–474. ISSN: 0019-9958. DOI: [https://doi.org/10.1016/S0019-9958\(67\)91165-5](https://doi.org/10.1016/S0019-9958(67)91165-5). URL: <https://www.sciencedirect.com/science/article/pii/S0019995867911655> (cit. on pp. 1, 3, 6–9, 12, 14, 17–21, 24, 28, 30, 33, 36, 60, 68, 70, 93, 96, 97).

- [GPM+20] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. “Generative Adversarial networks”. In: *Commun. ACM* 63.11 (Oct. 2020), pp. 139–144. ISSN: 0001-0782. DOI: [10.1145/3422622](https://doi.org/10.1145/3422622). URL: <https://doi.org/10.1145/3422622> (cit. on pp. 1, 3, 5).
- [GS01] F.A. Gers and E. Schmidhuber. “LSTM Recurrent Networks Learn Simple Context-Free and Context-Sensitive Languages”. In: *IEEE Transactions on Neural Networks* 12.6 (2001), pp. 1333–1340. DOI: [10.1109/72.963769](https://doi.org/10.1109/72.963769) (cit. on p. 17).
- [GZA+23] Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Harkirat Singh Behl, Xin Wang, Sébastien Bubeck, Ronen Eldan, Adam Tauman Kalai, Yin Tat Lee, and Yuanzhi Li. *Textbooks Are All You Need*. 2023. arXiv: [2306.11644](https://arxiv.org/abs/2306.11644) [cs.CL]. URL: <https://arxiv.org/abs/2306.11644> (cit. on p. 1).
- [Hah20] Michael Hahn. “Theoretical Limitations of Self-Attention in Neural Sequence Models”. In: *Transactions of the Association for Computational Linguistics* 8 (Jan. 2020), pp. 156–171. ISSN: 2307-387X. DOI: [10.1162/tac1_a_00306](https://doi.org/10.1162/tac1_a_00306). eprint: https://direct.mit.edu/tac1/article-pdf/doi/10.1162/tac1_a_00306/1923102/tac1_a_00306.pdf. URL: https://doi.org/10.1162/tac1_a_00306 (cit. on p. 17).
- [HBD+20] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. “The Curious Case of Neural Text Degeneration”. In: *International Conference on Learning Representations*. 2020. URL: <https://openreview.net/forum?id=rygGQyrFvH> (cit. on p. 27).
- [HCS+22] Dan Hendrycks, Nicholas Carlini, John Schulman, and Jacob Steinhardt. *Unsolved Problems in ML Safety*. 2022. arXiv: [2109.13916](https://arxiv.org/abs/2109.13916) [cs.LG]. URL: <https://arxiv.org/abs/2109.13916> (cit. on p. 1).
- [HG23] Michael Hahn and Navin Goyal. *A Theory of Emergent In-Context Learning as Implicit Structure Induction*. 2023. arXiv: [2303.07971](https://arxiv.org/abs/2303.07971) [cs.CL]. URL: <https://arxiv.org/abs/2303.07971> (cit. on p. 17).
- [HHG+20] John Hewitt, Michael Hahn, Surya Ganguli, Percy Liang, and Christopher D. Manning. “RNNs Can Generate Bounded Hierarchical Languages With Optimal Memory”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Ed. by Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu. Online: Association for Computational Linguistics, Nov. 2020, pp. 1978–2010. DOI: [10.18653/v1/2020.emnlp-main.156](https://doi.org/10.18653/v1/2020.emnlp-main.156). URL: <https://aclanthology.org/2020.emnlp-main.156> (cit. on p. 17).
- [HKK+18] Steve Hanneke, Adam Tauman Kalai, Gautam Kamath, and Christos Tzamos. “Actively Avoiding Nonsense in Generative Models”. In: *Proceedings of the 31st Conference On Learning Theory*. Ed. by Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet. Vol. 75. Proceedings of Machine Learning Research. PMLR, June 2018, pp. 209–227. URL: <https://proceedings.mlr.press/v75/hanneke18a.html> (cit. on p. 1).

- [HKM+22] Steve Hanneke, Amin Karbasi, Shay Moran, and Grigoris Velezgas. “Universal Rates for Interactive Learning”. In: *Advances in Neural Information Processing Systems*. Ed. by Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho. 2022. URL: <https://openreview.net/forum?id=dTKMy00PTJ> (cit. on p. 16).
- [HMZ23] Steve Hanneke, Shay Moran, and Qian Zhang. “Universal Rates for Multiclass Learning”. In: *Proceedings of Thirty Sixth Conference on Learning Theory*. Ed. by Gergely Neu and Lorenzo Rosasco. Vol. 195. Proceedings of Machine Learning Research. PMLR, July 2023, pp. 5615–5681. URL: <https://proceedings.mlr.press/v195/hanneke23a.html> (cit. on pp. 16, 26).
- [HS97] Sepp Hochreiter and Jürgen Schmidhuber. “Long Short-Term Memory”. In: *Neural Computation* 9.8 (Nov. 1997), pp. 1735–1780. ISSN: 0899-7667. DOI: 10.1162/neco.1997.9.8.1735. eprint: <https://direct.mit.edu/neco/article-pdf/9/8/1735/813796/neco.1997.9.8.1735.pdf>. URL: <https://doi.org/10.1162/neco.1997.9.8.1735> (cit. on p. 1).
- [HYM+23] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. *A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions*. 2023. arXiv: 2311.05232 [cs.CL]. URL: <https://arxiv.org/abs/2311.05232> (cit. on p. 1).
- [JLF+23] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. “Survey of Hallucination in Natural Language Generation”. In: *ACM Comput. Surv.* 55.12 (Mar. 2023). ISSN: 0360-0300. DOI: 10.1145/3571730. URL: <https://doi.org/10.1145/3571730> (cit. on p. 1).
- [JSR+24] Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Léo Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. *Mixtral of Experts*. 2024. arXiv: 2401.04088 [cs.LG]. URL: <https://arxiv.org/abs/2401.04088> (cit. on p. 29).
- [KKM+23] Alkis Kalavasis, Amin Karbasi, Shay Moran, and Grigoris Velezgas. “Statistical Indistinguishability of Learning Algorithms”. In: *Proceedings of the 40th International Conference on Machine Learning*. Ed. by Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett. Vol. 202. Proceedings of Machine Learning Research. PMLR, July 2023, pp. 15586–15622. URL: <https://proceedings.mlr.press/v202/kalavasis23a.html> (cit. on p. 18).
- [KM24] Jon Kleinberg and Sendhil Mullainathan. “Language Generation in the Limit”. In: *Advances in Neural Information Processing Systems*. Vol. 37. 2024. URL: <https://arxiv.org/abs/2404.06757> (cit. on pp. 1–4, 6–8, 11, 12, 14–17, 20, 22–24, 26, 28, 30, 32, 47, 51–55, 57, 59, 67, 91, 92).

- [Kni24] Will Knight. *Inside the Creation of DBRX, the World’s Most Powerful Open Source AI Model* | WIRED. Mar. 2024. URL: <https://www.wired.com/story/dbrx-inside-the-creation-of-the-worlds-most-powerful-open-source-ai-model/> (cit. on p. 29).
- [KTZ19] Vasilis Kontonis, Christos Tzamos, and Manolis Zampetakis. “Efficient Truncated Statistics with Unknown Truncation”. In: *2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)* (2019), pp. 1578–1595. URL: <https://api.semanticscholar.org/CorpusID:199442432> (cit. on p. 19).
- [KV24] Adam Tauman Kalai and Santosh S. Vempala. “Calibrated Language Models Must Hallucinate”. In: *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*. STOC 2024. Vancouver, BC, Canada: Association for Computing Machinery, 2024, pp. 160–171. ISBN: 9798400703836. DOI: [10.1145/3618260.3649777](https://doi.org/10.1145/3618260.3649777). URL: <https://doi.org/10.1145/3618260.3649777> (cit. on pp. 1, 2, 14, 16).
- [KVK22] Alkis Kalavasis, Grigoris Veleghas, and Amin Karbasi. “Multiclass Learnability Beyond the PAC Framework: Universal Rates and Partial Concept Classes”. In: *Advances in Neural Information Processing Systems*. Ed. by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh. Vol. 35. Curran Associates, Inc., 2022, pp. 20809–20822. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/82f0dae85424eb743017c90380e7ab9b-Paper-Conference.pdf (cit. on pp. 16, 26).
- [KWT+24] Katie Kang, Eric Wallace, Claire Tomlin, Aviral Kumar, and Sergey Levine. *Unfamiliar Finetuning Examples Control How Language Models Hallucinate*. 2024. arXiv: [2403.05612](https://arxiv.org/abs/2403.05612) [cs.LG]. URL: <https://arxiv.org/abs/2403.05612> (cit. on p. 1).
- [LAG+23] Bingbin Liu, Jordan T. Ash, Surbhi Goel, Akshay Krishnamurthy, and Cyril Zhang. “Transformers Learn Shortcuts to Automata”. In: *The Eleventh International Conference on Learning Representations*. 2023. URL: <https://openreview.net/forum?id=De4FYqjFueZ> (cit. on p. 17).
- [LBH15] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. “Deep Learning”. In: *Nature* 521.7553 (2015), pp. 436–444. DOI: [10.1038/nature14539](https://doi.org/10.1038/nature14539). URL: <https://doi.org/10.1038/nature14539> (cit. on p. 1).
- [Lit88] Nick Littlestone. “Learning Quickly When Irrelevant Attributes Abound: A New Linear-Threshold Algorithm”. In: *Machine Learning* 2.4 (1988), pp. 285–318. DOI: [10.1007/BF00116827](https://doi.org/10.1007/BF00116827). URL: <https://doi.org/10.1007/BF00116827> (cit. on pp. 9, 17, 18, 96, 97).
- [LMZ24] Jane H. Lee, Anay Mehrotra, and Manolis Zampetakis. “Efficient Statistics With Unknown Truncation, Polynomial Time Algorithms, Beyond Gaussians”. In: *2024 IEEE 65th Annual Symposium on Foundations of Computer Science (FOCS)*. 2024. URL: <https://arxiv.org/abs/2410.01656> (cit. on p. 19).
- [LRT24] Jiaxun Li, Vinod Raman, and Ambuj Tewari. *Generation Through the Lens of Learning Theory*. 2024. arXiv: [2410.13714](https://arxiv.org/abs/2410.13714) [cs.LG]. URL: <https://arxiv.org/abs/2410.13714> (cit. on p. 17).

- [Lu23] Zhou Lu. *Non-uniform Online Learning: Towards Understanding Induction*. 2023. arXiv: 2312.00170 [cs.LG]. URL: <https://arxiv.org/abs/2312.00170> (cit. on pp. 17, 18).
- [Man53] Benoit Mandelbrot. “An Informational Theory of the Statistical Structure of Language”. In: *Communication theory* 84 (1953), pp. 486–502. URL: <http://pdodds.w3.uvm.edu/research/papers/others/1953/mandelbrot1953a.pdf> (cit. on p. 1).
- [Mer19] William Merrill. “Sequential Neural Networks as Automata”. In: *Proceedings of the Workshop on Deep Learning and Formal Languages: Building Bridges*. Ed. by Jason Eisner, Matthias Gallé, Jeffrey Heinz, Ariadna Quattoni, and Guillaume Rabusseau. Florence: Association for Computational Linguistics, Aug. 2019, pp. 1–13. DOI: 10.18653/v1/W19-3901. URL: <https://aclanthology.org/W19-3901> (cit. on p. 17).
- [MIB+24] Kyle Mahowald, Anna A. Ivanova, Idan A. Blank, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. “Dissociating Language and Thought in Large Language Models”. In: *Trends in Cognitive Sciences* 28.6 (2024), pp. 517–540. ISSN: 1364-6613. DOI: <https://doi.org/10.1016/j.tics.2024.01.011>. URL: <https://www.sciencedirect.com/science/article/pii/S1364661324000275> (cit. on p. 1).
- [MKB+10] Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernocký, and Sanjeev Khudanpur. “Recurrent neural network based language model”. In: *Interspeech*. Vol. 2. 3. Makuhari. 2010, pp. 1045–1048. URL: https://www.fit.vut.cz/research/group/speech/public/publi/2010/mikolov_interspeech2010_IS100722.pdf (cit. on p. 1).
- [MM23] Maryanthe Malliaris and Shay Moran. *The Unstable Formula Theorem Revisited via Algorithms*. 2023. arXiv: 2212.05050 [math.LO]. URL: <https://arxiv.org/abs/2212.05050> (cit. on pp. 12, 18).
- [MS23] William Merrill and Ashish Sabharwal. “The Parallelism Tradeoff: Limitations of Log-Precision Transformers”. In: *Transactions of the Association for Computational Linguistics* 11 (2023), pp. 531–545. DOI: 10.1162/tac1_a_00562. URL: <https://aclanthology.org/2023.tac1-1.31> (cit. on p. 17).
- [MSS23] Shay Moran, Hilla Scheffler, and Jonathan Shafer. “The Bayesian Stability Zoo”. In: *Advances in Neural Information Processing Systems*. Ed. by A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine. Vol. 36. Curran Associates, Inc., 2023, pp. 61725–61746. URL: https://proceedings.neurips.cc/paper_files/paper/2023/file/c2586b71fd150fb56952e253a9c551cc-Paper-Conference.pdf (cit. on p. 18).
- [Nat87] Balaubramaniam Kausik Natarajan. “On Learning Boolean Functions”. In: *Proceedings of the Nineteenth Annual ACM Symposium on Theory of Computing*. STOC ’87. New York, New York, USA: Association for Computing Machinery, 1987, pp. 296–304. ISBN: 0897912217. DOI: 10.1145/28395.28427. URL: <https://doi.org/10.1145/28395.28427> (cit. on p. 18).
- [OAA+24] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel

Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rameev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O’Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian

- Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. *GPT-4 Technical Report*. 2024. arXiv: 2303.08774 [cs.CL]. URL: <https://arxiv.org/abs/2303.08774> (cit. on p. 6).
- [Pit89] Leonard Pitt. “Probabilistic Inductive Inference”. In: *J. ACM* 36.2 (Apr. 1989), pp. 383–433. ISSN: 0004-5411. DOI: [10.1145/62044.62053](https://doi.org/10.1145/62044.62053). URL: <https://doi.org/10.1145/62044.62053> (cit. on p. 17).
- [Ple21] Orestis Plevrakis. “Learning from Censored and Dependent Data: The case of Linear Dynamics”. In: *Proceedings of Thirty Fourth Conference on Learning Theory*. Ed. by Mikhail Belkin and Samory Kpotufe. Vol. 134. Proceedings of Machine Learning Research. PMLR, Aug. 2021, pp. 3771–3787. URL: <https://proceedings.mlr.press/v134/plevrakis21a.html> (cit. on p. 19).
- [PNP24] Binghui Peng, Srini Narayanan, and Christos Papadimitriou. *On Limitations of the Transformer Architecture*. 2024. arXiv: 2402.08164 [stat.ML]. URL: <https://arxiv.org/abs/2402.08164> (cit. on p. 16).
- [RG09] Luis Rademacher and Navin Goyal. “Learning Convex Bodies is Hard”. In: *COLT*. 2009. URL: <http://www.cs.mcgill.ca/~colt2009/papers/030.pdf#page=1> (cit. on p. 19).
- [RHW86] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. “Learning Internal Representations by Error Propagation”. In: *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1: Foundations*. Cambridge, MA, USA: MIT Press, 1986, pp. 318–362. ISBN: 026268053X. DOI: <https://ieeexplore.ieee.org/document/6302929> (cit. on p. 1).
- [SAN96] Jenny R Saffran, Richard N Aslin, and Elissa L Newport. “Statistical Learning by 8-Month-Old Infants”. In: *Science* 274.5294 (1996), pp. 1926–1928. DOI: <https://doi.org/10.1126/science.274.5294.1926> (cit. on p. 1).
- [Sat23] Adam Satariano. “E.U. Agrees on Landmark Artificial Intelligence Rules”. In: *The New York Times* 8 (2023). URL: <https://www.nytimes.com/2023/12/08/technology/eu-ai-act-regulation.html> (cit. on p. 1).
- [Sch97] Dale Schuurmans. “Characterizing Rational versus Exponential Learning Curves”. In: *Journal of Computer and System Sciences* 55.1 (1997), pp. 140–160. ISSN: 0022-0000. DOI: <https://doi.org/10.1006/jcss.1997.1505>. URL: <https://www.sciencedirect.com/science/article/pii/S0022000097915051> (cit. on pp. 4, 24).
- [SDD+23] Timo Schick, Jane Dwivedi-Yu, Roberto Dessi, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. “Toolformer: Language Models Can Teach Themselves to Use Tools”. In: *Advances in Neural Information Processing Systems*. Ed. by A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine. Vol. 36. Curran Associates, Inc., 2023, pp. 68539–68551. URL: https://proceedings.neurips.cc/paper_files/paper/2023/file/d842425e4bf79ba039352da0f6Paper-Conference.pdf (cit. on p. 29).

- [Sha51a] Claude E. Shannon. “Prediction and Entropy of Printed English”. In: *The Bell System Technical Journal* 30.1 (1951), pp. 50–64. DOI: [10.1002/j.1538-7305.1951.tb01366.x](https://doi.org/10.1002/j.1538-7305.1951.tb01366.x) (cit. on p. 1).
- [Sha51b] Claude E Shannon. “The Redundancy of English”. In: *Cybernetics; Transactions of the 7th Conference, New York: Josiah Macy, Jr. Foundation*. 1951, pp. 248–272. URL: <https://jontalle.web.engr.illinois.edu/uploads/537.F18/Papers/Shannon50b.pdf> (cit. on p. 1).
- [Shi90] Takeshi Shinohara. “Inductive Inference From Positive Data Is Powerful”. In: *Proceedings of the Third Annual Workshop on Computational Learning Theory*. COLT ’90. Rochester, New York, USA: Morgan Kaufmann Publishers Inc., 1990, pp. 97–110. ISBN: 1558601465. URL: https://api.lib.kyushu-u.ac.jp/opac_download_md/3127/rifis-tr-20.pdf (cit. on p. 18).
- [SHT23] Clayton Sanford, Daniel J Hsu, and Matus Telgarsky. “Representational Strengths and Limitations of Transformers”. In: *Advances in Neural Information Processing Systems*. Ed. by A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine. Vol. 36. Curran Associates, Inc., 2023, pp. 36677–36707. URL: https://proceedings.neurips.cc/paper_files/paper/2023/file/73bf692447f174984f30499ec9b20e04-Paper-Conference.pdf (cit. on p. 17).
- [Sip12] Michael Sipser. *Introduction to the Theory of Computation*. Introduction to the Theory of Computation. Cengage Learning, 2012. ISBN: 9781133187813. URL: <https://books.google.com/books?id=4J1ZMAEACAAJ> (cit. on pp. 5, 14, 20, 27).
- [SK23] Adam Satariano and Cecilia Kang. “How Nations Are Losing a Global Race to Tackle A.I.’s Harms.” In: *The New York Times (Digital Edition)* (2023), NA–NA. URL: <https://www.nytimes.com/2023/12/06/technology/ai-regulation-policies.html> (cit. on p. 1).
- [Soa99] R.I. Soare. *Recursively Enumerable Sets and Degrees: A Study of Computable Functions and Computably Generated Sets*. Perspectives in Mathematical Logic. Springer Berlin Heidelberg, 1999. ISBN: 9783540152996. URL: <https://books.google.com/books?id=9I7P100LU5gC> (cit. on pp. 5, 14).
- [Sol64] R.J. Solomonoff. “A Formal Theory of Inductive Inference. Part I”. In: *Information and Control* 7.1 (1964), pp. 1–22. ISSN: 0019-9958. DOI: [https://doi.org/10.1016/S0019-9958\(64\)90223-2](https://doi.org/10.1016/S0019-9958(64)90223-2). URL: <https://www.sciencedirect.com/science/article/pii/S0019995864902232> (cit. on p. 17).
- [SSA18] Konstantin Shmelkov, Cordelia Schmid, and Karteek Alahari. “How Good Is My GAN?” In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 213–229. URL: https://openaccess.thecvf.com/content_ECCV_2018/html/Konstantin_Shmelkov_How_good_is_ECCV_2018_paper.html (cit. on p. 1).
- [SVL14] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. *Sequence to Sequence Learning with Neural Networks*. 2014. arXiv: 1409.3215 [cs.CL]. URL: <https://arxiv.org/abs/1409.3215> (cit. on p. 1).

- [Tea24] The Mosaic Research Team. *Introducing DBRX: A New State-of-the-Art Open LLM* | Databricks Blog. <https://www.databricks.com/blog/introducing-dbrx-new-state-art-open-llm>. Mar. 2024. URL: <https://www.databricks.com/blog/introducing-dbrx-new-state-art-open-llm> (cit. on p. 29).
- [TLI+23] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. *LLaMA: Open and Efficient Foundation Language Models*. 2023. arXiv: [2302.13971](https://arxiv.org/abs/2302.13971) [cs.CL]. URL: <https://arxiv.org/abs/2302.13971> (cit. on pp. 1, 6).
- [Tur50] Alan M. Turing. “Computing Machinery and Intelligence”. In: *Mind* LIX.236 (1950), pp. 433–460. DOI: [10.1093/MIND/LIX.236.433](https://doi.org/10.1093/mind/LIX.236.433). URL: <https://doi.org/10.1093/mind/LIX.236.433> (cit. on p. 1).
- [Val84] Leslie G Valiant. “A Theory of the Learnable”. In: *Commun. ACM* 27.11 (Nov. 1984), pp. 1134–1142. ISSN: 0001-0782. DOI: [10.1145/1968.1972](https://doi.org/10.1145/1968.1972). URL: <https://doi.org/10.1145/1968.1972> (cit. on p. 18).
- [Vap13] Vladimir Vapnik. *The Nature of Statistical Learning Theory*. Springer science & business media, 2013. URL: <https://doi.org/10.1007/978-1-4757-3264-1> (cit. on p. 18).
- [VL23] Tom Viering and Marco Loog. “The Shape of Learning Curves: A Review”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45.6 (2023), pp. 7799–7819. DOI: [10.1109/TPAMI.2022.3220744](https://doi.org/10.1109/TPAMI.2022.3220744) (cit. on p. 24).
- [VSP+17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. “Attention is All you Need”. In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Vol. 30. Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf (cit. on p. 1).
- [WM23] Karen Weise and Cade Metz. “When A.I. Chatbots Hallucinate”. In: *The New York Times* 9 (2023), pp. 610–23. URL: <https://www.nytimes.com/2023/05/01/business/ai-chatbots-hallucination.html> (cit. on p. 1).
- [WWS+22] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter brian, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. “Chain-of-Thought Prompting Elicits Reasoning in Large Language Models”. In: *Advances in Neural Information Processing Systems*. Ed. by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh. Vol. 35. Curran Associates, Inc., 2022, pp. 24824–24837. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf (cit. on p. 1).
- [XJK24] Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. *Hallucination is Inevitable: An Innate Limitation of Large Language Models*. 2024. arXiv: [2401.11817](https://arxiv.org/abs/2401.11817) [cs.CL]. URL: <https://arxiv.org/abs/2401.11817> (cit. on p. 16).

- [XRL+22] Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. “An Explanation of In-context Learning as Implicit Bayesian Inference”. In: *International Conference on Learning Representations*. 2022. URL: <https://openreview.net/forum?id=RdJVFCHjUMI> (cit. on p. 17).
- [YPP+21] Shunyu Yao, Binghui Peng, Christos Papadimitriou, and Karthik Narasimhan. “Self-Attention Networks Can Process Bounded Hierarchical Languages”. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Ed. by Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli. Online: Association for Computational Linguistics, Aug. 2021, pp. 3770–3785. DOI: [10.18653/v1/2021.acl-long.292](https://doi.org/10.18653/v1/2021.acl-long.292). URL: <https://aclanthology.org/2021.acl-long.292> (cit. on p. 17).
- [ZL95] Thomas Zeugmann and Steffen Lange. “A Guided Tour Across the Boundaries of Learning Recursive Languages”. In: *Algorithmic Learning for Knowledge-Based Systems: GOSLER Final Report*. Ed. by Klaus P. Jantke and Steffen Lange. Berlin, Heidelberg: Springer Berlin Heidelberg, 1995, pp. 190–258. ISBN: 978-3-540-44737-5. DOI: [10.1007/3-540-60217-8_12](https://doi.org/10.1007/3-540-60217-8_12). URL: https://doi.org/10.1007/3-540-60217-8_12 (cit. on p. 18).
- [ZLC+23] Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. *Siren’s Song in the AI Ocean: A Survey on Hallucination in Large Language Models*. 2023. arXiv: [2309.01219](https://arxiv.org/abs/2309.01219) [cs.CL]. URL: <https://arxiv.org/abs/2309.01219> (cit. on p. 1).

A Further Discussion on Decidability of $\text{MOP}(\cdot)$

In this section, we present a generating algorithm $\mathcal{G} = (\mathcal{G}_n)$ for which MOP is undecidable. The corresponding generating algorithm is static in the sense that $\mathcal{G}_1 = \mathcal{G}_2 = \dots$. For each t , \mathcal{G}_t is the following randomized Turing machine.

Input: None

Setup: The internal random tape contains symbols from $\{0, 1, 2\}$

Description:

1. Read bits $r = r_1 r_2 \dots$ from the random tape until the first 2 is found on the tape
2. Let $b = 1$ if r is of the form $\langle M \rangle w$ where $\langle M \rangle$ is a valid representation of a Turing machine and w is any (possibly) empty string
3. **If** $r = 1$ **then:** Execute M on input w and **return** $\langle M \rangle w1$ if M halts
4. **Else:** **return** 0

Proposition A.1. $\text{MOP}(\cdot)$ is undecidable for the above Turing machine.

Proof. The proof is a simple reduction to the halting problem, which is well-known to be undecidable. To see this, observe that to decide whether M halts on input w , it suffices to check if $\langle M \rangle w1$ is in the support of the above machine. \square

B Results With Subset Oracles

In this section, we design algorithms that have access to a subset oracle that, given indices i and j , answers whether “ $L_i \subseteq L_j$?” We give two algorithms (1) an algorithm that identifies in the limit without requiring tell-tale oracles and (2) a best-of-both-words algorithm that generates consistently and achieves breadth whenever possible.

B.1 Identification in the Limit Without Tell-Tale Oracle via Subset Oracles

Angluin [Ang80] showed that a collection of (recursive) languages \mathcal{L} is identifiable in the limit if and only if each $L_i \in \mathcal{L}$ has a finite “tell-tale” set T_i that, roughly speaking, enables one to eliminate L_i if it is not the target. If one has access to an oracle that, given an index i , outputs the tell-tale T_i , then one can identify \mathcal{L} using an algorithm by Angluin [Ang80]. Our next result shows that, if one has access to queries of the form “ $L_i \subseteq L_j$?”, then one can identify any identifiable language collection without access to a tell-tale oracle.

Theorem B.1. *Let S be an oracle that, given indices i and j , outputs Yes if $L_i \subseteq L_j$ and outputs No otherwise. Fix any identifiable collection of languages $\mathcal{L} = \{L_1, L_2, \dots\}$. There is an algorithm \mathcal{A} that given, an enumeration of a target language $K \in \mathcal{L}$ (for any K) and access to S , identifies K in the limit.*

Importantly, \mathcal{A} does not need to be provided the tell-tale of K or any other language in \mathcal{L} .

Interestingly, the algorithm in [Theorem B.1](#) is the same as an algorithm proposed by Kleinberg and Mullainathan [\[KM24\]](#): the algorithm defines a certain notion of critical languages and selects the last critical language, say L_j . Naturally, since we want to identify the language, instead of outputting an element of L_j the algorithm will guess the index. This algorithm identifies K in the sense that after some finite time t^* , it outputs an index z such that $L_z = K$. The specific index z outputted, however, may change infinitely often. This can also be avoided by using the post-processing routine in [Lemma 5.4](#).

Proof. The algorithm to identify K is simple:

For $t \in \{1, 2, \dots\}$ **do**:

1. Observe element x_t and let S_t be the set of all elements observed so far
2. Construct a *version space* V_t consisting of all languages in $L_{\leq t}$ consistent with S_t , i.e.,

$$V_t: \{L_j: 1 \leq j \leq t, L_j \supseteq S_t\}.$$

Define an language $L_i \in V_t$ to be critical if L_i is the smallest-index language in V_t or L_i is a subset of all languages preceding it in V_t , i.e., if $L_i \subseteq L_j$ for all $1 \leq j < i$

3. Construct the set $C_t \subseteq V_t$ of all critical languages
4. **return** i where L_i is the largest-indexed language in the set of critical languages C_t

Let i be the first index such that $K = L_i$. The above algorithm outputs an index z with $L_z = K$ when K is the last element of the set of critical languages C_t . This condition is implied by the following two conditions.

(A) K is in the set of critical languages C_t .

(B) All the languages L_j with $j > i$ that are included in C_t satisfy $L_j = K$.

Result (4.3) of Kleinberg and Mullainathan [\[KM24\]](#) shows that there is a finite time t_a after which Condition (A) holds. We will show that there is also a finite time t_b after which Condition (B) holds. This shows that, for any $t \geq \max\{t_a, t_b\}$, \mathcal{A} outputs an index $z(t)$ such $L_{z(t)} = K$. As mentioned before one can convert this into an algorithm that outputs a fixed index (after some finite time) using the post-processing routine in [Lemma 5.4](#).

Condition (B) holds after a finite time. Since \mathcal{L} is identifiable, it must satisfy Angluin’s tell-tale criteria ([Definition 10](#)) and, hence, $K = L_i$ has a finite tell-tale set T_i . (Recall that T_i is not known to us; our proof will not need this.) Fix any $j > i$ and any time $t \geq t_a$ (after which K is guaranteed to be a critical language). If L_j is a critical language, then by the definition of critical languages and the fact that K is critical (P1) $L_j \subseteq K$ and (P2) $L_j \in V_t$. Further, if $L_j \subsetneq K$, then L_j cannot contain the tell-tale T_i of K (by the properties of tell-tales; [Definition 10](#)). Therefore, if S_t (the set of samples seen until step t) contains T_i , then L_j cannot be in the version space as otherwise it would need to contain T_i . It follows that, if $S_t \supseteq T_i$ and $t \geq t_a$, then either P2 will be violated or P1 will imply that $L_j = K$. Finally, since T_i is finite and, hence, there is a finite time t'_b when all elements of T_i have been observed and, the result follows by letting $t_b := \max\{t_a, t'_b\}$.

□

B.2 Best-Of-Both Worlds: Generating With Breadth When Possible

Consider the variant of the algorithm in the previous section which instead of outputting the index of the last critical language outputs an unseen sample from the language. This is precisely the algorithm of Kleinberg and Mullainathan [KM24].²⁰ Kleinberg and Mullainathan [KM24] showed that this algorithm consistently generates in the limit. An immediate corollary of the identification result in the last section is that this algorithm also achieves breadth for any identifiable collection \mathcal{L} .

Corollary B.2. *Let S be an oracle that, given indices i and j , outputs Yes if $L_i \subseteq L_j$ and outputs No otherwise. Fix any collection of countably many languages $\mathcal{L} = \{L_1, L_2, \dots\}$.*

There is a generating algorithm \mathcal{G} that given, an enumeration of a target language $K \in \mathcal{L}$ (for any K) and access to S , consistently generates from K in the limit. Moreover, whenever \mathcal{L} is identifiable, this algorithm generates consistently with breadth in the limit.

Importantly, \mathcal{G} does not need to be provided the tell-tale of K or any other language in \mathcal{L} .

Finally, we recall that for any non-identifiable collection \mathcal{L} , generation with breadth is impossible whenever a MOP(\mathcal{G}) can be implemented.

C Further Results for Consistent Generation With Approximate Breadth

Result in the Limit

In this section, we study another notion of generation with approximate breadth. Informally, requires that the generating algorithm is consistent and puts zero mass only on *finitely* many points of the target language K . The formal definition is as follows.

Definition 21 (Generation with Approximate Breadth). *A generating algorithm $\mathcal{G} = (\mathcal{G}_n)$ is said to generate with approximate breadth for a collection $\mathcal{L} = \{L_1, L_2, \dots\}$ if, for any $K \in \mathcal{L}$ and enumeration x_1, x_2, \dots of K , there is an $n_0 \geq 1$, such after seeing $n \geq n_0$ elements x_1, \dots, x_n , $\text{supp}(\mathcal{G}_n) \subseteq K$ and $|K \setminus \text{supp}(\mathcal{G}_n)|$ is finite.*

Observe that the above is a weakening of generation with breadth – since the generating algorithm \mathcal{G} is allowed to miss elements in the target language infinitely often. This weakening of generation with breadth turns out to be incomparable to the notion of unambiguous generation studied in Section 3.3. To see this, observe that (1) on the one hand, \mathcal{G} can satisfy the above definition while generating with breadth from a language L that is a strict subset of K and (2) on the other hand, unambiguous generation allows the generator to generate samples outside of K infinitely often, which is barred by the above definition.

Our main result in this section is as follows.

²⁰Note that since we only output an element from the last critical language and not its index, we do not need to perform the post-processing used in the previous section, so this is Kleinberg and Mullainathan [KM24]’s algorithm.

Theorem C.1 (Impossibility of Approximate Generation in the Limit). *For every non-identifiable collection of countably many languages \mathcal{L} , no generating algorithm stable in the limit, for which $\text{MOP}(\cdot)$ (Definitions 5 and 6) is decidable, can generate from \mathcal{L} in the limit with approximate breadth according to Definition 21.*

The proof of Theorem C.1, like the proof of Theorem 3.5 is also by a contradiction to the non-identifiability of \mathcal{L} . The difference is that in this proof we also need to identify the finitely many elements of K “missed” by \mathcal{G} .

Proof. Recall that Gold [Gol67] showed that for any collection of countably many languages, there is always an algorithm I_{PN} that, given a positive and negative enumeration of the target, identifies it in the limit. Now, toward a contradiction, suppose there is a stable generator for which $\text{MOP}(\mathcal{G})$ is decidable and it has the property described in Definition 21. We claim that using \mathcal{G} and I_{PN} , we can construct an identifier for \mathcal{L} (from positive examples), which contradicts the fact that \mathcal{L} is non-identifiable.

Fix any target language K , its enumeration s_1, s_2, \dots , and the (unknown) constant t^* , such that after t^* many iterations the algorithm achieves generation according to Definition 21. We claim that the following algorithm identifies \mathcal{L} .

Input: Access to a generator \mathcal{G} for \mathcal{L} that (1) that, in the limit, becomes consistent and satisfies $|K \setminus \text{supp}(\mathcal{G})| < \infty$ and (2) for which $\text{MOP}(\mathcal{G})$ is decidable, and access to the algorithm I_{PN} that identifies \mathcal{L} in the limit from a positive and negative enumeration of the target language.

Description:

1. **For each** $t \in \mathbb{N}$ **do:**

- (a) Observe the t -th sample s_t and let S_t be the set of samples seen so far
- (b) Train the generator \mathcal{G} from scratch over the t samples in S_t
- (c) For each $1 \leq i \leq t$, label the i -th string x_i in the domain as $y_i = \text{MOP}(\mathcal{G})(x_i)^a$
- (d) For each $1 \leq i \leq t$, if $x_i \in S_t$ and $y_i = 0$, set $y_i = 1$ # to correct elements missed by \mathcal{G}
- (e) Train I_{PN} from scratch on samples x_1, \dots, x_t with labels y_1, \dots, y_t
- (f) **output** the index guessed by I_{PN} and go to the next iteration

^aHere, $\text{MOP}(\mathcal{G})(x)$ is the answer to the membership oracle problem for \mathcal{G} given input x .

Since $\text{MOP}(\mathcal{G})$ is decidable, the above algorithm can be implemented using a Turing machine. We claim that the above algorithm identifies the target language K after a finite number of iterations. To formalize this, fix any enumeration s_1, s_2, \dots of the target language. Since after a finite time t^* , \mathcal{G} becomes consistent, stabilizes, and satisfies $|K \setminus \text{supp}(\mathcal{G})| < \infty$, after iteration t^* , for any string x , $\text{MOP}(\mathcal{G})(x)$ matches $\mathbb{1}\{x \in K\}$ except for the finitely many elements of $M = K \setminus \text{supp}(\mathcal{G}_t)$. Note that since \mathcal{G} 's support stabilizes, M is independent of the iteration $t \geq t^*$. Let t' be the time when all elements of M appear in the enumeration s_1, s_2, \dots . Observe that in all iterations $t \geq \max\{t^*, t'\}$, the labels in Step 2 (d) correct all the mismatches between $\text{MOP}(\mathcal{G})(x)$ matches $\mathbb{1}\{x \in K\}$. Finally, since I_{PN} identifies in the limit, there is a finite t_{PN} such that I_{PN} identifies K once it is given labels for the first $t \geq t_{\text{PN}}$ examples in the domain. Combining this with the previous information

implies that that I_{PN} and, hence, our algorithm identifies K after $\max \{t^*, t', t_{\text{PN}}\} < \infty$ iterations. This gives the desired contradiction, proving [Theorem C.1](#). Note that the above identification algorithm does not need to know any of t^* , t' , and t_{PN} . \square

Result in the Statistical Setting

Our approach to get this result follows the same high-level idea with [Theorem 3.6](#). The error of a generating algorithm in this setting, based on [Definition 21](#) is

$$\text{er}(\mathcal{G}_n) = \mathbb{1} \{ \text{supp}(\mathcal{G}_n) \not\subseteq K \text{ or } |K \setminus \text{supp}(\mathcal{G}_n)| = \infty \} \quad (12)$$

The formal statement is below.

Theorem C.2 (Impossibility of Approximate Generation). *For every non-identifiable collection of countably many languages \mathcal{L} , no stable generating algorithm, for which $\text{MOP}(\cdot)$ ([Definitions 5 and 6](#)) is decidable, can generate from \mathcal{L} with approximate breadth according to [Definition 21](#), at any rate.*

Using the tools we developed for the setting of unambiguous generation, we can show the following result which transforms a learner that works in the statistical setting into a learner that works in the online setting. We start with the following lemma.

Lemma C.3. *Let $R: \mathbb{N} \rightarrow \mathbb{R}_{\geq 0}$ be a rate function, i.e., $\lim_{n \rightarrow \infty} R(n) = 0$, let \mathcal{L} be a language collection, and $(\mathcal{G}_n: \mathcal{X}^n \rightarrow \mathfrak{G})_{n \in \mathbb{N}}$ be generating algorithm for which $\text{MOP}(\cdot)$ is decidable and which satisfies the following two properties:*

- $(\mathcal{G}_n)_{n \in \mathbb{N}}$ is a stable generator ([Definition 7](#)), and
- for its approximate generation error $\text{er}(\cdot)$ ([Equation \(12\)](#)) it holds that, for every valid distribution \mathcal{P} with respect to \mathcal{L} there exist $c, C > 0$ such that $\mathbb{E}_{X_1, \dots, X_n \sim \mathcal{P}^n} [\text{er}(\mathcal{G}_n(X_1, \dots, X_n))] \leq C \cdot R(c \cdot n)$.

Then, for every valid distribution \mathcal{P} with respect to \mathcal{L} it holds that

$$\Pr_{\{X_i\}_{i \in \mathbb{N}} \sim \mathcal{P}^\infty} [\exists n^* \in \mathbb{N}: \forall n \geq n^* \text{ it holds that } \text{er}(\mathcal{G}_n(X_1, \dots, X_n)) = 0] = 1.$$

Proof of [Lemma C.3](#). Assume towards contradiction that there exists some valid \mathcal{P} with respect to \mathcal{L} so that

$$\Pr_{\{X_i\}_{i \in \mathbb{N}} \sim \mathcal{P}^\infty} [\exists n^* \in \mathbb{N}: \forall n \geq n^* \text{ it holds that } \{\text{er}(\mathcal{G}_n(X_1, \dots, X_n)) = 0\}] = c' < 1.$$

Let us also denote $c'' := 1 - c'$. Notice that $c'' > 0$. Since \mathcal{P} is a valid distribution with respect to \mathcal{L} , it is supported over some $K \in \mathcal{L}$, so we have that, with probability 1, an infinite i.i.d. draw from \mathcal{P} is an enumeration of K (see [Proposition 5.2](#)). Let us call this event \mathcal{E}_1 .

Moreover, since \mathcal{G}_n is a stable generator (in an online sense), under the event \mathcal{E}_1 (i.e., when the samples from \mathcal{P} form an enumeration of K), there exists some smallest number $t^* := t^*(X_1, \dots) \in \mathbb{N}$ such that for all $n \geq t^*$

$$\text{supp}(\mathcal{G}_n(X_1, \dots, X_n)) = \text{supp}(\mathcal{G}_{n+1}(X_1, \dots, X_{n+1})).$$

Now, t^* depends on the specific enumeration drawn and, hence, the distribution \mathcal{P} induces a distribution over t^* . Further, note that with probability 1, $t^* < \infty$. Hence, $\Pr_{\{X_i\}_{i \in \mathbb{N}} \sim \mathcal{P}^\infty} [t^*(X_1, \dots) > n]$ approaches 0 as $n \rightarrow \infty$. In particular, there is some number $n_1 \in \mathbb{N}$ such that for all $n \geq n_1$

$$\Pr_{\{X_i\}_{i \in \mathbb{N}} \sim \mathcal{P}^\infty} [t^*(X_1, \dots) > n] \leq \frac{c''}{3}.$$

Moreover, since the generator achieves rate $R(\cdot)$ and $\lim_{n \rightarrow \infty} R(n) = 0$, it holds that

$$\lim_{n \rightarrow \infty} \Pr_{X_1, \dots, X_n \sim \mathcal{P}^n} [\text{er}(\mathcal{G}_n(X_1, \dots, X_n)) \neq 0] = 0.$$

Thus, there is some $n_2 \in \mathbb{N}$ such that, for all $n \geq n_2$

$$\Pr_{X_1, \dots, X_n \sim \mathcal{P}^n} [\text{er}(\mathcal{G}_n(X_1, \dots, X_n)) \neq 0] \leq \frac{c''}{3}.$$

Let $n_3 := \max\{n_1, n_2\}$. Hence, taking a union bound, we see that with probability at least $1 - 2c''/3$ over the draw of $\{X_i\}_{i \in \mathbb{N}}$ it holds that

- $\text{er}(\mathcal{G}_{n_3}(X_1, \dots, X_{n_3})) = 0$, and
- $\text{supp}(\mathcal{G}_n(X_1, \dots, X_n)) = \text{supp}(\mathcal{G}_{n_3}(X_1, \dots, X_{n_3}))$, for all $n \geq n_3$.

By the definition of $\text{er}(\cdot)$, for any $n, n' \in \mathbb{N}$, samples x_{i_1}, \dots, x_{i_n} and $x_{j_1}, \dots, x_{j_{n'}}$ it holds that

$$\begin{aligned} \text{supp}(\mathcal{G}_n(x_{i_1}, \dots, x_{i_n})) &= \text{supp}(\mathcal{G}_{n'}(x_{j_1}, \dots, x_{j_{n'}})) \implies \\ \text{er}(\mathcal{G}_n(x_{i_1}, \dots, x_{i_n})) &= \text{er}(\mathcal{G}_{n'}(x_{j_1}, \dots, x_{j_{n'}})) \end{aligned}$$

These two conditions immediately imply that, with probability at least $1 - 2c''/3 > c'$, for all $n \geq n_3$ it holds that

- $\text{er}(\mathcal{G}_n(X_1, \dots, X_n)) = 0$, and
- $\text{supp}(\mathcal{G}_{n+1}(X_1, \dots, X_{n+1})) = \text{supp}(\mathcal{G}_n(X_1, \dots, X_n))$.

Hence,

$$\Pr_{\{X_i\}_{i \in \mathbb{N}} \sim \mathcal{P}^\infty} [\exists n^* \in \mathbb{N}: \forall n \geq n^* \text{ it holds that } \{\text{er}(\mathcal{G}_n(X_1, \dots, X_n)) = 0\}] > c',$$

which gives the desired contradiction. This concludes the proof. \square

Using the previous result, we derive the next statement regarding the conversion of a statistical learner to an online one.

Lemma C.4. *Let $R: \mathbb{N} \rightarrow \mathbb{R}_{\geq 0}$ be a rate function, i.e., $\lim_{n \rightarrow \infty} R(n) = 0$, let \mathcal{L} be a language collection, and $(\mathcal{G}_n: \mathcal{X}^n \rightarrow \mathfrak{G})_{n \in \mathbb{N}}$ be generating algorithm for which MOP is decidable and satisfies the following two properties:*

- $(\mathcal{G}_n)_{n \in \mathbb{N}}$ is a stable generator ([Definition 7](#)), and
- for its approximate generation error ([Equation \(12\)](#)) it holds that, for every valid distribution \mathcal{P} with respect to \mathcal{L} there exist $c, C > 0$ such that $E_{X_1, \dots, X_n \sim \mathcal{P}^n} [\text{er}(\mathcal{G}_n(X_1, \dots, X_n))] \leq C \cdot R(c \cdot n)$.

Then, there is a randomized generating algorithm $(g'_n: \mathcal{X}^n \xrightarrow{r} \mathfrak{G})_{n \in \mathbb{N}}$ for which, for any target language $K \in \mathcal{L}$ and every enumeration σ of K , it holds that

- $(g'_n)_{n \in \mathbb{N}}$ is a stable generator (Definition 7), and

-

$$\Pr [\exists n^* \in \mathbb{N}: \forall n \geq n^* \text{ it holds that } \text{er}(g'_n(\sigma_1, \dots, \sigma_n)) = 0] = 1,$$

where the probability is with respect to the randomness of the algorithm.

The proof of Lemma C.4 is identical to the proof of Lemma 7.4, since the only property of the error function that is needed is that once the algorithm has stabilized then its error also stabilizes. For completeness, we give the details below.

Proof of Lemma C.4. Let $K \in \mathcal{L}$ be any target language and σ be any enumeration of K . Let \mathcal{P}_σ be the distribution defined in Definition 20. We know that, by definition, \mathcal{P}_σ is valid with respect to \mathcal{L} , since it is supported on K . Let $(g'_n)_{n \in \mathbb{N}}$ be a generator which, for every $n \in \mathbb{N}$, runs g_n on \mathcal{P}_σ . In order to draw samples from \mathcal{P}_σ the generator g'_n uses its internal randomness and the process described in Proposition 7.2. Since \mathcal{P}_σ is a valid distribution with respect to \mathcal{L} , Lemma C.4 gives us that

$$\Pr_{\{X_i\}_{i \in \mathbb{N}} \sim \mathcal{P}_\sigma^\infty} [\exists n^* \in \mathbb{N}: \forall n \geq n^* \text{ it holds that } \{\text{er}(g_n(X_1, \dots, X_n)) = 0\}] = 1.$$

Hence, this implies that

$$\Pr [\exists n^* \in \mathbb{N}: \forall n \geq n^* \text{ it holds that } \{\text{er}(g'_n(\sigma_1, \dots, \sigma_n)) = 0\}] = 1,$$

where the probability is taken with respect to the internal randomness of the algorithm. Moreover, since $(g_n)_{n \in \mathbb{N}}$ is a stable generator it also holds that $(g'_n)_{n \in \mathbb{N}}$ is a stable generator. This concludes the proof. \square

The proof of Theorem C.2 follows as a corollary of the previous result (Lemma C.4) and Theorem C.1.

Proof of Theorem C.2. Let \mathcal{L} be a countable collection of languages. Assume that such a stable generating algorithm exists. Then, using the construction from Lemma C.4 to get a stable generator that generates missing only finitely many elements in the limit, for every target language $K \in \mathcal{L}$ and every enumeration σ of K , with probability 1. This contradicts the impossibility result from Theorem C.1. \square

D Further Comparison With Online Learning

In this section, we provide some comparisons between the Gold-Angluin (GA) model [Gol67; Ang79] and the online game of Bousquet et al. [BHM+21] (that extends the standard online learning model of Littlestone [Lit88]), which at first sight share a lot of similarities. However, there are important differences between the two models, which we believe are worth highlighting.

Let us first recall the setting of Bousquet et al. [BHM+21], appropriately rephrased to the context of our work. There is a domain \mathcal{X} , a collection of languages $\mathcal{L} \subseteq \{0,1\}^{\mathcal{X}}$, and two players, the learner and the adversary, play a game over an infinite sequence of discrete rounds. In every round $t \in \mathbb{N}$, the adversary presents an example $x_t \in \mathcal{X}$ to the learner and the learner guesses its label \hat{y}_t . Subsequently, the adversary reveals the true label y_t to the learner. We say that the learner makes a mistake if $y_t \neq \hat{y}_t$. This can be thought of as the learner trying to guess whether the example belongs to the target language. The adversary has to satisfy the following constraint. For every round $t \in \mathbb{N}$ there must be some $L_t \in \mathcal{L}$ such that $\mathbb{1}\{x_\tau \in L_t\} = y_\tau, \forall \tau \leq t$. The goal of the learner is to make finitely many mispredictions and the goal of the adversary is to force infinitely many such mispredictions.

- In the GA model, the goal of the learner is to identify the true language in a finite number of steps. In the online game of Bousquet et al. [BHM+21], the goal of the learner is to make a finite number of mistakes in its predictions. Moreover, in the GA model, the learner observes only positive examples while in the standard online setting, the adversary can provide both positive and negative examples.
- The set of languages identifiable in the limit in Gold’s model is characterized by Angluin’s criterion (Definition 10). The collections that are online learnable with a finite number of mistakes is characterized by the absence of infinite Littlestone trees [BHM+21]. If there is a uniform bound on the number of mistakes, then this corresponds to the standard finiteness of the Littlestone dimension [Lit88].
- In the GA model, the adversary fixes the true language in advance. In (realizable) online learning, the only thing that matters is *consistency* of the hypothesis class on the given examples, *i.e.*, for every sequence of examples, along with their labels, there should exist some hypothesis in the hypothesis class that perfectly labels the given sequence (which can change as the online game progresses, see also Section 3 of Bousquet et al. [BHM+21]). This subtle difference leads to starkly different learning landscapes.
- In the GA model, the adversary must include all elements of K in the enumeration (the domain is countable). In Littlestone’s online setting, there is no such restriction (the feature space can even be uncountable).
- Another crucial difference is that the algorithm does not receive feedback about its guess in the GA model. This is in contrast to the standard online setting where, at each round, the learner gets feedback about its prediction. However, it is important to stress that the incentives of the adversaries in Gold’s model and in the model of Bousquet et al. [BHM+21] are different. In the GA model, the goal is not to maximize the number of mistakes that the learner does, but to essentially not allow the learner to identify the true target language.

To further show separations between the online setting of Gold [Gol67] and the online game of Bousquet et al. [BHM+21], we will need two important definitions coming from the work of Bousquet et al. [BHM+21].

Definition 22 (Littlestone Tree [BHM+21]). A Littlestone tree for \mathcal{L} is a complete binary tree of depth $d \leq \infty$ whose internal nodes are labeled by \mathcal{X} , and whose two edges connecting a node to its children are labeled 0 and 1, such that every finite path emanating from the root is consistent with a language $L \in \mathcal{L}$.

More precisely, a Littlestone tree is a collection

$$\{x_{\mathbf{u}} : 0 \leq k < d, \mathbf{u} \in \{0, 1\}^k\} \subseteq \mathcal{X}$$

such that for every $\mathbf{y} \in \{0, 1\}^d$ and $n < d$, there exists $L \in \mathcal{L}$ so that $\mathbb{1}\{x_{\mathbf{y}_{\leq k}} \in L\} = y_{k+1}$ for $0 \leq k \leq n$. We say that \mathcal{L} has an infinite Littlestone tree if there is a Littlestone tree for \mathcal{L} of depth $d = \infty$.

For instance, thresholds over \mathbb{N} do not have an infinite Littlestone tree (yet, they have Littlestone trees of arbitrary length). On the other side, thresholds over the reals have an infinite Littlestone tree. Given this definition, we can show a separation between the online setting of Bousquet et al. [BHM+21] and the GA model.

First, we note that a language collection is not online learnable in the online setting of Bousquet et al. [BHM+21] if and only if it has an infinite Littlestone tree. We will show that there exists a countable language collection \mathcal{L} over a countable domain that has an infinite Littlestone tree but it is identifiable in the limit with positive examples.

Example 3 (Infinite Littlestone Tree but Identifiable with Positive Examples). Consider a countable domain \mathcal{X} and fix an enumeration $x_{\emptyset}, x_0, x_1, x_{00}, x_{01}, x_{10}, x_{11}, \dots$ of its elements. We use this enumeration to create the nodes of a Littlestone tree, *i.e.* the root consists of x_{\emptyset} , its left, right child is x_0, x_1 , respectively etc. Then, for each level $d \in \mathbb{N}$ and any path $y \in \{0, 1\}^d$ we create a *finite* language L_y with index y , whose elements are the elements of \mathcal{X} that appear on the path with label 1. We consider the language collection $\mathcal{L} = \{L_y : y \in \{0, 1\}^d, d \in \mathbb{N}\}$. This means that for any finite level $d < \infty$, we add all the 2^d languages in the collection (where for any path y , L_y contains the elements x that are labeled with 1 in the path). Hence, \mathcal{L} contains all the finite prefixes of the paths of the infinite Littlestone tree. The language collection \mathcal{L} is countably infinite since the collection of all finite paths of the binary tree admits an enumeration. By construction, this class induces an infinite Littlestone tree and hence is not online learnable in the game of Bousquet et al. [BHM+21]. However, since it is a countable collection of finite languages, it is identifiable in the limit with positive examples in the GA model (see also Section 3.4.3 and 8.3).

E Borel-Cantelli Lemmas

In this section, we present two well-known results due to Borel and Cantelli which are useful for our derivations.

Lemma E.1 (First Borel-Cantelli Lemma). Let $\{\mathcal{E}_n\}_{n \in \mathbb{N}}$ be a sequence of events. If

$$\sum_{n \in \mathbb{N}} \Pr[\mathcal{E}_n] < \infty,$$

then the probability that infinitely many of them occur is 0, that is

$$\Pr \left[\limsup_{n \rightarrow \infty} \mathcal{E}_n \right] = 0.$$

The previous result has a *partial* converse, which we state below.

Lemma E.2 (Second Borel-Cantelli Lemma). *Let $\{\mathcal{E}_n\}_{n \in \mathbb{N}}$ be a sequence of independent events. If*

$$\sum_{n \in \mathbb{N}} \Pr[\mathcal{E}_n] = \infty,$$

then the probability that infinitely many of them occur is 1, that is

$$\Pr \left[\limsup_{n \rightarrow \infty} \mathcal{E}_n \right] = 1.$$

Notice that, unlike the first Borel-Cantelli lemma, the second one requires that the events are *independent*.