NON-UNIFORM LANGUAGE IDENTIFICATION IN THE LIMIT

Anonymous authors
Paper under double-blind review

ABSTRACT

We study language identification in the limit and ask when the number of examples needed for identification can be bounded independently of their order of presentation. The learner must identify every target in the limit and be correct whenever it has observed a target-dependent number of distinct positive examples. In learning theory, allowing the bound to depend on the target is known as a *non-uniform* guarantee. We characterize which countable collections admit non-uniform identification through a variant of Angluin's condition for identification in the limit, and similarly characterize the stronger requirement of *uniform* identification, where the number of examples is the same for each target. We extend these characterizations to variants studied in classical and recent work, asking whether they can overcome the obstacles to non-uniform identification. Two variants relax what counts as a valid output: (1) accepting a language that differs from the target on finitely many strings, and (2) allowing the learner to output a list of up to k guesses, one of which must be correct. Two more variants give the learner additional information: (1) negative examples (*i.e.*, strings outside the target), or (2) auxiliary labels, called *colors,* that depend on the target and are attached either to each example or to each of its prefixes, as in Charikar, Kleinberg, and Pabbaraju (2026a;b). For some collections that are identifiable in the limit, non-uniform identification fails even with finitely many disagreements, lists of k guesses, negative examples, or one of m colors on each example, for any finite k and m. In contrast, coloring each prefix of each example with one of just two colors, by a rule that does not depend on the collection, makes every countable collection of infinite languages non-uniformly identifiable.

1 INTRODUCTION

How can a machine learn the rules of a language from examples? Gold (1967) formalized this question as language identification in the limit. He argued that, since the rules of English cannot be written down explicitly, a machine designed to learn the structure of English "will have to learn its rules from implicit information." In the model we study, a language is a set of finite strings over a fixed finite alphabet. The learner knows a countable collection $\mathcal{L}$ of candidate languages and observes examples $x_1, x_2, \ldots$ from an unknown target $L \in \mathcal{L}$: they may arrive in any order and may be repeated, but every string in L eventually appears. After each example, the learner guesses a language from $\mathcal{L}$; it is never told which strings lie outside L or whether its guess is correct. It is said to identify L in the limit if all its guesses beyond some finite time are correct, *i.e.*, they equal L.

Gold (1967) showed that even the collection of regular languages cannot be identified in this model. Angluin (1980) subsequently characterized the collections that are identifiable. Identification in the limit only requires the learner to be correct eventually. It says nothing about how many examples the learner needs, and even when identification is possible, this number need not be bounded in terms of the target alone. The learner may have to wait for an example that rules out a competing candidate, and arbitrarily many distinct examples can arrive first. For instance, consider the two languages consisting of all natural numbers and of the even numbers, encoded as strings. A learner can guess the even numbers until it sees an odd number, after which it guesses all natural numbers forever; this identifies either target in the limit. Yet an adversary can present arbitrarily many distinct even numbers before revealing an odd number. Such a sequence is consistent with both targets, so no learner can ensure it outputs the correct target after any fixed number of distinct examples.

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

We ask when enough distinct examples suffice for identification without relying on the identity of these examples. We require a learner that identifies every target in $\mathcal{L}$ in the limit and, for each target $L \in \mathcal{L}$, is correct whenever it has observed at least $d^*(L)$ distinct examples, where the finite bound $d^*(L)$ must hold for every order of presentation of L. Following standard terminology in learning theory, we call this requirement *non-uniform identification*. Requiring one bound for the whole collection gives the stronger requirement of *uniform identification*. By definition, uniform identification implies non-uniform identification, which implies identification in the limit.

Why non-uniform guarantees? Non-uniform and uniform identification both count the number of distinct examples the learner has seen, not the total number of examples. This is necessary and standard (Kleinberg & Mullainathan, 2024; Li et al., 2025), as otherwise, if two languages in $\mathcal{L}$ share a string, an adversary can repeat it arbitrarily often. Data-constrained scaling laws for language models make a similar distinction: repeated training data is worth less than fresh data (Muennighoff et al., 2023). Counting distinct examples lets us ask, for exact identification in Gold's model, *whether enough distinct examples always suffice or whether particular informative examples must appear early.*

Non-uniform identification poses this question in the worst case: since any finite subset of the target can appear first, the bound must hold even when the first examples delay the evidence that distinguishes the target. Gold's criterion also requires success for every enumeration, but it lets the learner wait for this evidence, such as an odd number in the opening example, however late it arrives.

As in Gold's model, the amount of evidence needed may depend on the target: languages that are harder to distinguish within a collection may need more, and a common bound may not exist. Such non-uniform guarantees have a long history in learning theory, including PAC learning (Benedek & Itai, 1994; Shalev-Shwartz & Ben-David, 2014), online learning (Lu, 2024; Hanneke et al., 2025c), and language generation (Li et al., 2025; Charikar & Pabbaraju, 2025a). In PAC learning, for example, finite unions of intervals on the real line admit no common sample bound, yet a union of m intervals can be learned from a number of samples that grows with m (Benedek & Itai, 1994). Non-uniform identification brings this kind of guarantee to learning from positive examples.

Identification and generation. Language generation in the limit, introduced by Kleinberg & Mullainathan (2024), is a useful comparison. Here, the learner sees the same examples but instead of guessing the identity of the target language, it tries to guess a fresh string from it and succeeds if eventually every output belongs to the target language. Despite this close relationship, Kleinberg & Mullainathan (2024) showed that generation is possible for every countable collection of infinite languages, while identification is impossible for many such collections. This result sparked a burst of work on generation (e.g., Li et al., 2025; Charikar & Pabbaraju, 2025a; Kalavasis et al., 2025; 2026; Mehrotra et al., 2026; Raman & Raman, 2025; Bai et al., 2026; Peale et al., 2025; Charikar & Pabbaraju, 2025b; Hanneke et al., 2025a), in which non-uniform guarantees are a recurring theme. In particular, Li et al. (2025) and Charikar & Pabbaraju (2025a) independently showed that every such collection admits non-uniform generation, *i.e.*, a generator that always outputs an unseen string in the target after a target-dependent number of distinct examples. Identification is in stark contrast: the opening example shows that non-uniform identification can fail even when identification in the limit is possible, as Kleinberg & Mullainathan (2024) noted for a similar pair of languages. In that example, a generator can always output an unseen even number, whereas an identifier must distinguish two targets sharing arbitrarily many distinct examples.

Identification has a much longer history than generation. Since the work of Gold (1967) and Angluin (1980), many variants have been studied (see Angluin & Smith, 1983; Jain et al., 1999; Lange et al., 2008), including identification with anomalies or bounded "mind changes" (Case & Lynes, 1982; Case & Smith, 1983), teams of learners (Smith, 1982; Jain & Sharma, 1996), probabilistic learners (Pitt, 1989), stochastic examples (Angluin, 1988), and noisy data (Stephan, 1997). Recent work, motivated in part by generation, revisits identification by relaxing the requirement on the output (Charikar et al., 2025; Kalavasis et al., 2025; 2026) or by giving the learner auxiliary information, such as computational traces or annotations (Peng et al., 2026; Charikar et al., 2026b).

Surprisingly, despite this long history, non-uniform identification has received little attention. Indeed, to our knowledge, no characterization of it is known. This leaves two basic questions open.

- **Q1:** First, what distinguishes non-uniformly identifiable collections from identifiable ones?
- **Q2:** Second, can relaxing the output requirements or providing auxiliary information, which make identification in the limit much more tractable, do the same for non-uniform identification?

1.1 Our Contributions

Our first result characterizes non-uniform identification and helps answer **Q1**.

Informal Theorem 1 (Non-uniform identification; see Theorem 3.1)**.** *A countable collection $\mathcal{L}$ is non-uniformly identifiable if and only if (i) every two distinct languages in $\mathcal{L}$ have finite intersection, and (ii) for each $L \in \mathcal{L}$, the proper subsets of L that belong to $\mathcal{L}$ have bounded size.*

To see how Informal Theorem 1 answers **Q1**, compare it with Angluin (1980)'s characterization of identification in the limit: Condition (ii) strengthens Angluin's tell-tale condition (see Theorem 2.1), while condition (i) is a new requirement. Condition (i) is necessary because a non-uniform identification requires the learner to commit to a target L after $d^*(L)$ distinct examples. If two languages share infinitely many strings, as in the opening example, an adversary can fool the learner: it presents more strings from their intersection than the non-uniform bound for either language allows and only then reveals a string that separates them, so the learner must commit to one language before this point and, hence, make a mistake in at least one case. Further, we also show that $\mathcal{L}$ is uniformly identifiable exactly when its *intersection dimension* $I(\mathcal{L}) = \sup_{L \neq L' \in \mathcal{L}} |L \cap L'|$ is finite (Theorem A.1).

Our next result gives an alternative, more combinatorial characterization of non-uniform identification. It follows a common pattern in learning theory: once the uniform version of a problem is characterized, the non-uniform version is typically characterized by unions of uniformly learnable classes. For example, a hypothesis class $\mathcal{H}$ is non-uniformly PAC learnable if and only if $\mathcal{H} = \bigcup_{n \geq 1} \mathcal{H}_n$ for classes $\mathcal{H}_n$ of finite VC dimension (Benedek & Itai, 1994). Similar characterizations are known for online learning and for language generation in the limit (Lu, 2024; Li et al., 2025), but, surprisingly, not for identification in the limit.

Informal Theorem 2 (Union form; see Theorem 3.4)**.** *A countable collection $\mathcal{L}$ is non-uniformly identifiable if and only if $\mathcal{L} = \bigcup_{n \geq 1} \mathcal{C}_n$ for uniformly identifiable collections $\mathcal{C}_1 \subseteq \mathcal{C}_2 \subseteq \cdots$, where each $\mathcal{C}_n$ contains every language in $\mathcal{L}$ that is a subset of a language in $\mathcal{C}_n$.*

Compared with PAC learning, the collections must be nested, $\mathcal{C}_1 \subseteq \mathcal{C}_2 \subseteq \cdots$, as in non-uniform generation's characterization (Li et al., 2025). Unlike generation, each $\mathcal{C}_n$ must also be closed under sublanguages in $\mathcal{L}$. Both requirements are necessary (see Remark A.2), and the second adds to the differences between generation and identification in the limit. The above thus gives a second answer to **Q1:** non-uniformly identifiable collections are exactly those built in this way from uniformly identifiable ones. This puts non-uniform identification in a similar form as non-uniform PAC and online learning, with the *intersection dimension* $I(\mathcal{C}_n)$ playing the role of the VC dimension of $\mathcal{H}_n$.

Relaxed Outputs and Auxiliary Information. Our remaining results address **Q2**: whether relaxing the output requirements or providing auxiliary information in the input makes non-uniform identification possible for more collections. For simplicity, here we consider collections with only infinite languages (see Section 3 for the general statements). For these collections, our characterization for non-uniform identification can be simplified: we only require Condition (i) in Informal Theorem 1; as Condition (i) implies Condition (ii). We begin with variants that relax the output requirements.

A. Relaxing the output. First, we allow the learner to output $k \geq 2$ guesses for the target instead of just one, and the learner succeeds as long one of the guesses is correct (Charikar et al., 2025).

Informal Theorem 3 (k-list; see Theorem 3.5)**.** *A countable collection of infinite languages is non-uniformly k-list identifiable exactly when every $k+1$ distinct languages have finite intersection.*

Charikar et al. (2025) introduced this relaxation and showed that a collection is k-list identifiable in the limit if and only if it is a union of k collections that are each identifiable in the limit. Our results show that the analogous decomposition fails for non-uniform identification: indeed, if A, B, and C are disjoint infinite sets, then $\{A \cup B, B \cup C, C \cup A\}$ is non-uniformly 2-list identifiable (by Informal Theorem 3), yet any split into two collections places two languages with infinite intersection in the same collection, violating Condition (i) of Informal Theorem 1.

Regarding **Q2**, lists do help: the opening example, for instance, is non-uniformly 2-list identifiable, since the learner can always output both languages. They do not remove the obstacle behind condition (i) of Informal Theorem 1, however: languages that share infinitely many strings cannot be told apart after any fixed number of distinct examples. Lists only move this obstacle from pairs of languages to groups of $k+1$: if $k+1$ languages share infinitely many strings, non-uniform k-list identification fails, even when the collection is identifiable in the limit, as for $\{E\} \cup \{E \cup \{n\} : n \text{ odd}\}$, where E is the set of even numbers.

Our second relaxation, *approximate identification*, allows one guess but relaxes the notion of correctness: a guess is correct if it differs from the target on finitely many strings (Case & Lynes, 1982).

Informal Theorem 4 (Approximate identification; see Theorem 3.7)**.** *A countable collection of infinite languages is non-uniformly approximately identifiable if and only if every two languages in it that differ on infinitely many strings have finite intersection.*

The uniform versions of Informal Theorems 3 and 4 replace these finiteness conditions with a common finite bound on the same intersections (Theorems A.8 and A.9). Regarding **Q2**, approximate identification helps only with languages that differ on finitely many strings, since a guess among such languages is correct for each of them. It does not help with languages that share infinitely many strings and also differ on infinitely many. In particular, unlike lists, it does not handle the opening example, whose two languages share every even number and differ on every odd one.

B. Auxiliary information. In several parts of the LLM training pipeline, like supervised finetuning (SFT), LLMs might be trained not only on positive examples, but also on some auxiliary information about the process that generated them, *e.g.*, the chain-of-thought (CoT). Motivated by this, Peng et al. (2026) studied identification in the limit with computational traces, and Charikar et al. (2026a;b) proposed a further abstraction where strings appear with additional information is the form of *colors*; we study two instantiations of this setting, one where each string comes with a single (target-dependent) color (called terminal colorings), and one where each prefix of the string is given a color (called trace colorings).

Informal Theorem 5 (Colors; see Theorems 3.9 and 3.10)**.** *Let $\mathcal{L}$ be a countable collection of infinite languages. (i) If each example is colored with one of $m \geq 2$ colors, rules permitting non-uniform identification of $\mathcal{L}$ exist if and only if every $m+1$ distinct languages in $\mathcal{L}$ have finite intersection. (ii) If each prefix of each example is colored with one of two colors every countable $\mathcal{L}$ permits non-uniform identification.*

In the previous result, the condition in (i) coincides with that for m-list identification in Informal Theorem 3, so these coloring face the same obstructions as lists. Our results for uniform identification with colors (Propositions A.5 and 3.11) show that, with either kind of coloring, some collections of infinite languages cannot be uniformly identified with any finite number of colors. Colors on prefixes, however, behave very differently, and they give the only positive answer to **Q2** among the variants we study. Perhaps surprisingly, coloring prefixes of examples with just two colors suffices to make every countable collection of infinite languages non-uniformly identifiable, including collections that are not identifiable in the limit without colors.

Lastly, we consider negative examples, *i.e.*, strings outside the target, presented along with the positive ones. This setting goes back to Gold (1967) and is also motivated by post-training methods that learn from incorrect outputs (Zhu et al., 2025). Interestingly, while Gold showed that identification with this type of information becomes possible for *all* countable collections, our result shows that the set of non-uniformly identifiable languages *shrinks* compared to the positive-only version. The result is postponed to Section 3 (see Theorems A.3 and 3.12).

	Standard Setting	Relaxing the Output		Auxiliary Information	
		Approximate	k-list	Negative Examples	Colors
Non-Uniform	Thm. 3.1	Thm. 3.7	Thm. 3.5	Thm. 3.12	Thm. 3.8
Uniform	Thm. A.1	Thm. A.8	Thm. A.9	Thm. A.3	Thm. 3.8

Technical Overview. We now explain some of the aspects of our techniques. Recall that the main obstacle to language identification in the limit is overgeneralization: an observer who sees only positives examples from L may not distinguish it from its supersets. Angluin's condition characterizes exactly when such overgeneralization can be avoided, and it relies on the existence of "tell-tales:" for every $L \in \mathcal{L}$ there is a finite $T_L \subset L$ such that all (proper) subsets of L do not contain it. Crucially, these tell-tales might depend on the identities of particular elements. At a high level, to establish non-uniform guarantees, we need to pinpoint conditions that allow for tell-tale sets which do not depend on the identities on the elements, just their cardinality. It turns out that to achieve that we need to control the sizes of subsets of each language.

As a warm-up, consider the collection $\{\mathbb{N}\} \cup \{\{1,2\},\{3,4,5\},\{6,7,8,9\},\dots\}$. A tell-tale for $\mathbb{N}$ is $T = \{1,3\}$, since all subsets of $\mathbb{N}$ are eliminated after seeing T. However, in non-uniform

identification, we cannot afford to wait until we see $\{1, 3\}$. In fact, no non-uniform guarantee is possible since for any claimed bound $d^*(\mathbb{N})$ on how many examples an identifier needs to see to identify $\mathbb{N}$, the adversary can find another language L with at least $d^*(\mathbb{N})$ elements, enumerate those, and create ambiguity between $\mathbb{N}$ and L. In general, subsets of unbounded size lead to these types of issues. In contrast, if all subsets have bounded size, seeing sufficiently many examples simultaneously rules out all these subsets (acting as a tell-tale). For languages that have infinite size, we additionally need to control the sizes of intersections of pairs of languages; as otherwise, like the example in the introduction, if a pair of languages has infinite intersection, then the adversary can keep presenting new examples from the intersection. These two ideas together are sufficient to give a characterization of non-uniform identification. Both of them can be naturally adapted to obtain a characterization on uniform identification by requiring that the sizes of the subsets are globally bounded across different languages as well. In general, this high-level principle of converting non-uniform results to uniform results is applied to the other settings we consider.

Perhaps our most technically involved results are on colorings, where we use auxiliary colors to reduce intersection sizes to make collections non-uniformly or uniformly identifiable. With trace colorings this is not too difficult since we get many colors for long strings (Theorem 3.9), and for terminal colorings this is possible by directly coloring strings differently in the languages which share them (provided there are not too many languages which share the string) (Theorem 3.10). However, uniform colorings are much more challenging. For example, upgrading the conditions on classes that make them non-uniformly identifiable with terminal colorings by requiring a global bound on intersection sizes still does not prevent unbounded blowup of pairwise intersections if there are many such different intersections (even though this technique works for many of our other settings). We show this with a carefully constructed example: representing languages as points in the plane and strings as low-degree polynomials passing through them, we can ensure not too many languages share any given string (not many low-degree polynomials can share a fixed set of points), yet pairwise intersections of languages are large: no matter the coloring, a large aggregate number of polynomials and pairs of points creates unbounded pairwise intersections (see Proposition 3.11).

Lastly, we highlight that for the rest of the settings we consider, natural strategies to obtain the results fail. For instance, Charikar et al. (2025) showed that a collection is k-list identifiable in the limit if and only if it can be written as a union of k identifiable collections. As we explained earlier, the same qualitative result does not go through for non-uniform identification, hence we need to carefully reason about properties of k-wise intersections. Moreover, for non-uniform approximate identification a natural approach is to consider the collection $\widetilde{\mathcal{L}}$ where all sets languages that differ on finitely many elements are replaced by a "representative", and ask that $\widetilde{\mathcal{L}}$ is non-uniformly identifiable. However, we show in Appendix A.4 that this is not sufficient to get a tight characterization.

1.2 Further Related Work

Non-uniform learning. The non-uniform learning results discussed above share a common structure: a class is learnable if and only if it is a countable union of classes that each admit a uniform guarantee. For probably approximately correct (PAC) learning, these are classes of finite VC dimension (Benedek & Itai, 1994); for online learning, they are classes of finite Littlestone dimension (Lu, 2024). Hanneke et al. (2025b) studies a complementary PAC model in which the learning rate may depend on the distribution of inputs, but the same rate must hold for every target concept. Their results also sharpen target-dependent PAC bounds.

Language identification and generation. Classical identification and its variants form a broad literature; see Jain et al. (1999). Recent work studies list identification (Charikar et al., 2025), computational traces (Peng et al., 2026), and color annotations that do not depend on an underlying machine (Charikar et al., 2026b). Charikar et al. (2026a) show that two terminal colors suffice for ordinary identification of every countable collection of infinite languages. Our characterization shows that no fixed finite terminal-color palette suffices for non-uniform identification of every such collection. The generation literature also studies how much of the target language a generator can produce and whether its set of possible outputs eventually stops changing (Kalavasis et al., 2025; 2026), as well as tolerance to incorrect or missing examples (Mehrotra et al., 2026).

2 Preliminaries

We now introduce notation and formally define the identification problems we study. Let Σ denote a finite alphabet (for instance, Σ can be $\{0, 1\}$ or the English alphabet). A language L is a subset of Σ^*, where Σ^* is the set of finite strings over Σ. We assume throughout that languages are nonempty.

We write $\mathcal{X} = \Sigma^*$ and call it the example space. A *language class*, or collection, $\mathcal{L}$ is a nonempty countable set of languages. Without loss of generality, we fix an enumeration $\mathcal{L} = \{L_1, L_2, \dots\}$ without repetitions, so that each language has a unique index. For languages L and L', we write $|L|$ for the cardinality of L and $L \triangle L' = (L \setminus L') \cup (L' \setminus L)$ for their symmetric difference.

An identification algorithm, or *identifier*, M maps each finite sequence of strings in $\mathcal{X}$ to the index of a language in $\mathcal{L}$. Language identification is modeled as a game between an identifier M and an adversary. The adversary is all-knowing: it has unbounded computational power and knows M. First, the adversary selects a target language $L \in \mathcal{L}$ and an *enumeration* $S = (x_1, x_2, \dots)$ of L: an infinite sequence of strings in L in which every string of L appears at least once. We write $S_t = \{x_1, \dots, x_t\}$ for the set of distinct examples observed by round t. In each round t, the adversary shows x_t, and the identifier outputs the index $M(x_1, \dots, x_t)$. The identifier's goal is to output the index of the target in all but finitely many rounds.

Definition 1 (Identification in the limit Gold (1967))**.** *Fix a language class $\mathcal{L}$ and an identifier M. We say M identifies $L \in \mathcal{L}$ in the limit if, for all enumerations S of L, there exists a finite round $d^*(L, S)$ such that M outputs the index of L for all rounds $t \geq d^*(L, S)$. We say M identifies the entire class $\mathcal{L}$ in the limit if it identifies L in the limit for all $L \in \mathcal{L}$.*

Gold (1967) introduced the problem of language identification in the limit and provided several necessary or sufficient conditions for a collection $\mathcal{L}$ to be identifiable. Subsequently, Angluin (1980) characterized the classes that are identifiable in the limit. Roughly speaking, each language must be distinguishable from its proper sublanguages in $\mathcal{L}$ by a finite set of its strings.

Theorem 2.1 (Angluin, 1980)**.** *A language class $\mathcal{L}$ is identifiable in the limit if and only if every $L \in \mathcal{L}$ has a* tell-tale*, i.e., a finite set $T \subseteq L$ such that $T \nsubseteq L'$ for every $L' \in \mathcal{L}$ with $L' \subsetneq L$.*

One limitation of language identification in the limit is that we have no guarantees on the sample complexity of M. More precisely, the identifier may take arbitrary many rounds until it succeeds, depending on both the target language and enumeration. A natural strengthening of identification in the limit then requires *uniformity* of the sample complexity of the identifier over enumerations.

Definition 2 (Non-uniform identification in the limit)**.** *We say identifier M non-uniformly identifies $L \in \mathcal{L}$ in the limit if M identifies L in the limit and there exists a positive integer threshold $d^*(L)$ such that, for every enumeration S of L, M outputs the index of L in every round t with $|S_t| \geq d^*(L)$. We say M non-uniformly identifies $\mathcal{L}$ in the limit if it non-uniformly identifies every $L \in \mathcal{L}$.*

As discussed in the introduction, non-uniform guarantees have a long history in learning theory, including PAC learning (Benedek & Itai, 1994; Shalev-Shwartz & Ben-David, 2014) and online learning (Lu, 2024; Hanneke et al., 2025c). Such guarantees were studied for language generation in the limit by Li et al. (2025); Charikar & Pabbaraju (2025a). We can further strengthen this to uniform identification in the limit.

Definition 3 (Uniform identification in the limit)**.** *We say identifier M uniformly identifies $\mathcal{L}$ in the limit if it identifies every language in $\mathcal{L}$ in the limit and there exists a common positive integer threshold d^* such that, for every target $L \in \mathcal{L}$ and every enumeration of L, M outputs the index of L in every round t with $|S_t| \geq d^*$.*

3 Our Results

3.1 Characterization of Identification in the Limit

Our first result, whose proof appears in Appendix A.1, characterizes non-uniform identification.

Theorem 3.1 (Characterization of Non-Uniform Identifiability)**.** *Language class $\mathcal{L}$ is non-uniformly identifiable in the limit if and only if both of the following hold:*

1. *For all distinct $L, L' \in \mathcal{L}$, $|L \cap L'| < \infty$.*
2. *For each $L \in \mathcal{L}$, there exists $d \in \mathbb{N}$ such that $|L'| < d$ for all $L' \in \mathcal{L}$ with $L' \subsetneq L$.*

Recall that Angluin's characterization (Theorem 2.1) requires each $L \in \mathcal{L}$ to have a finite tell-tale. This helps the identifier avoid getting stuck on supersets of the target: since it sees no negative examples, the identifier never sees evidence against a superset of the target. Condition (2) strengthens this requirement: it bounds the sizes, not the number, of proper sublanguages of L. Equivalently, every sufficiently large finite set of strings of L is a tell-tale. Condition (1) is a new requirement. A learner that only needs to be correct eventually can wait for a string that separates two languages, but a non-uniform guarantee forces it to commit after finitely many distinct examples, which is impossible if two languages share infinitely many strings.

Example 3.2. To see why Condition (2) is needed even when every language has a tell-tale, consider the collection $\{\mathbb{N}\} \cup \{\{1,2\}, \{3,4,5\}, \{6,7,8,9\}, \dots\}$. This collection is identifiable in the limit as $\{1,3\}$ is a tell-tale for $\mathbb{N}$ and each finite language is its own tell-tale. However, it violates Condition (2) of Theorem 3.1. Indeed, given any claimed non-uniform bound $d^*(\mathbb{N})$, the adversary can always choose a finite language with more than $d^*(\mathbb{N})$ elements and enumerate it so the identifier cannot separate this language from $\mathbb{N}$ in only $d^*(\mathbb{N})$ distinct examples.

Our next result, Theorem 3.3, characterizes uniform identification through a combinatorial dimension, analogous to the Littlestone dimension for online learning (Littlestone, 1988) and the closure dimension for language generation (Li et al., 2025). Its proof appears in Appendix A.1.

Definition 4. *Define the* intersection dimension *of a collection $\mathcal{L}$ as $I(\mathcal{L}) \coloneqq \sup_{L \neq L' \in \mathcal{L}} |L \cap L'|$.*

Theorem 3.3. *Language class $\mathcal{L}$ is uniformly identifiable in the limit if and only if $I(\mathcal{L}) < \infty$.*

We can reformulate this characterization in the style of Theorem 3.1, as a bound on the sizes of intersections of distinct languages in $\mathcal{L}$ (Theorem A.1). The combinatorial, intersection dimension is useful as it gives a second characterization of non-uniform identification:

Theorem 3.4. *Language class $\mathcal{L}$ is non-uniformly identifiable in the limit if and only if there exists a non-decreasing sequence of classes $\mathcal{C}_1 \subseteq \mathcal{C}_2 \subseteq \dots$ with $\mathcal{L} = \bigcup_{n=1}^{\infty} \mathcal{C}_n$ such that*

$$I(\mathcal{C}_n) < \infty \qquad \text{and} \qquad \text{if } L \in \mathcal{C}_{i+1} \setminus \mathcal{C}_i, \text{ then } L \not\subset L' \text{ for all } L' \in \mathcal{C}_i\,.$$

This perspective places online learning, language generation, and language identification in a common framework. In each, a class is non-uniformly learnable exactly when it is a union of uniformly learnable classes, but the union needs more structure as we move from online learning to generation and then to identification. For online learning, as for PAC learning (Benedek & Itai, 1994), any countable union suffices (Lu, 2024). For generation, the elements of the union must be nested (Li et al., 2025). For identification, Theorem 3.4 additionally requires the second condition above. These requirements are necessary (see Remark A.2). The proof of Theorem 3.4 appears in Appendix A.1.

3.2 k-list Identification in the Limit

In k-list identification, the identifier outputs a list of k indices in each round, rather than a single index, and succeeds if, after some finite round, every output list contains the target's index (Charikar et al., 2025). We consider two variants: in *stable* k-list identification, the list must eventually stop changing, while in *unstable* k-list identification, it may keep changing as long as it contains the target. The unstable variant has roots in team learning (Smith, 1982; Jain & Sharma, 1996), and Charikar et al. (2025) characterized the collections are k-list identifiable (with instability). We first characterize non-uniform k-list identification without stability (see Appendix A.5 for a proof):

Theorem 3.5. *A countable language class $\mathcal{L}$ is non-uniformly k-list identifiable in the limit (without stability) if and only if both of the following conditions hold:*

1. *Every intersection of $k+1$ distinct languages is finite.*
2. *For every k distinct languages $L_1, \dots, L_k$, $\sup_{L \in \mathcal{L}: L \subseteq L_1 \cap \cdots \cap L_k, L \neq L_i} |L| < \infty$.*

We show the following separation between stable and unstable non-uniform k-list identification:

Proposition 3.6. *Consider the language class $\mathcal{L} = \{\mathbb{N}, A_1, A_2, \dots\}$ where $A_n = \{1, \dots, n\}$. Then $\mathcal{L}$ is non-uniformly 2-list identifiable in the limit without stability, but not non-uniformly k-list identifiable with stability for any finite k.*

A tight characterization of stable non-uniform k-list identification remains open. In the uniform case, both variants share a characterization (Theorem A.9). This is because an identifier could use knowledge of the target-independent uniform bound. The proof of this result is in Appendix A.5.

3.3 Approximate Identification in the Limit

Our second relaxation is approximate identification in the limit, which allows one guess but deems guesses that differ from the target on finitely many strings as correct. This relaxation goes back to identification with anomalies (Case & Lynes, 1982; Case & Smith, 1983); see also Jain et al. (1999).

Formally, M approximately identifies $L \in \mathcal{L}$ in the limit if, after some finite round, every output of M has finite symmetric difference with L. As with lists, we say M is *stable* if its outputs must eventually stop changing and *unstable* if they may keep changing. Non-uniform and uniform approximate identification add bounds on the number of distinct examples, as in Definitions 2 and 3.

Unlike k-list identification (Proposition 3.6), stable and unstable approximate identification are equivalent, since an unstable identifier becomes stable if it outputs a fixed representative of its original guess under the equivalence relation of finite symmetric difference (see Appendix A.4).

Our next result characterizes non-uniform approximate identification and generalizes Theorem 3.1:

Theorem 3.7. *Language class $\mathcal{L}$ is non-uniformly approximately identifiable in the limit (with or without stability) if and only if both of the following hold:*

1. *For each $L \in \mathcal{L}$, there is $d \in \mathbb{N}$ with $|L'| < d$ for all $L' \in \mathcal{L}$ with $L' \subset L$ and $|L \setminus L'| = \infty$.*
2. *For all $L, L' \in \mathcal{L}$, either $|L \cap L'| < \infty$ or $|L \triangle L'| < \infty$.*

We present the analogous uniform characterization in Theorem A.8; see Appendix A.4 for a proof.

3.4 Characterization of Identification in the Limit with Colors

Next, we turn to variants that give the learner auxiliary information (see Section 1.1 for motivation). Inspired by chain-of-thought training, Peng et al. (2026) studied identification from examples annotated with the computational traces that produced them, and Charikar et al. (2026a;b) abstracted such annotations as *colors*. We adopt the latter, more general setting.

Definition 5. *Consider a set P, which we call the* palette*. Given a language class $\mathcal{L}$, a coloring scheme is a collection of coloring functions $\{c_L\}_{L \in \mathcal{L}}$ where $c_L : L \to P^*$. In other words, coloring functions are language-dependent functions mapping strings to finite sequences of colors from P.*

The identifier knows the coloring scheme and, in each round t, observes the example x_t together with its colors $c_L(x_t)$ under the target L. Our next result, whose proof appears in Appendix A.3, characterizes non-uniform identification in the limit with colors:

Theorem 3.8. *For $L, L' \in \mathcal{L}$, let $B(L, L', c) \coloneqq \{x \in L \cap L' : c_{L'}(x) = c_L(x)\}$. Language class $\mathcal{L}$ is non-uniformly identifiable in the limit with coloring functions $\{c_L\}_{L \in \mathcal{L}}$ if and only if:*

1. *For all $L \in \mathcal{L}$, some $d \in \mathbb{N}$ satisfies $|B(L, L', c) \cap L'| < d$ for all $L' \in \mathcal{L}$ with $L' \subset L$.*
2. *For all distinct $L, L' \in \mathcal{L}$, $|B(L, L', c) \cap L \cap L'| < \infty$.*

We present a characterization of the uniform setting in Appendix A.3.

Trace colorings. A *trace coloring* assigns a color to every prefix of a string. Formally, each $L \in \mathcal{L}$ has a *trace function* tr_L that maps every finite prefix of a string in L to a color in P, and its trace coloring function is $c_L(x_1 x_2 \cdots x_n) = (\mathrm{tr}_L(\varepsilon), \mathrm{tr}_L(x_{\le 1}), \ldots, \mathrm{tr}_L(x_{\le n}))$. Charikar et al. (2026b) introduced these colorings and showed that any countable collection of languages over alphabet Σ can be identified in the limit with $|\Sigma| + 1$ trace colors, but did not study non-uniform identification.

Equipped with Theorem 3.8, we can ask when a coloring scheme exists that makes a given class non-uniformly or uniformly identifiable. Our next result, whose proof appears in Appendix A.3, shows that trace colors greatly expand the set of language classes which are non-uniformly identifiable in the limit. In fact, we are able to ensure that the coloring function has an additional desirable property called *collection independence*: concretely, collection independence requires a set of coloring functions $\{c_L\}_{L \subset \Sigma^*}$ such that their restrictions to (suitably defined) collections $\mathcal{L}$ result in a (non)-uniformly identifiable collection. In other words, we require each coloring function c_L (for $L \subset \Sigma^*$) to be independent of the collection $\mathcal{L}$.

Theorem 3.9. *There is a collection independent trace coloring scheme using two colors that makes every countable collection of languages non-uniformly identifiable in the limit with color traces.*

This is the only variant we study in which every class that is identifiable in the limit becomes non-uniformly identifiable; in fact, so does every countable class. Uniform identification, however, can still fail: some class is not uniformly identifiable with any trace coloring scheme over a finite palette, even one that depends on the class (Proposition A.5). It is an interesting open question to identify collections that become uniformly identifiable by some trace coloring scheme with a finite palette.

Terminal colorings. Another type of coloring function of recent interest is a *terminal coloring*, which assigns a single color to each string, *i.e.*, $c_L : L \to P$. Our next result, proved in Appendix A.3, shows that they are much weaker than trace colorings, even though Charikar et al. (2026a) show that every countable collection becomes identifiable with just a two-color collection-independent terminal coloring scheme.

Theorem 3.10. *For $k \geq 2$, there exists a collection dependent terminal coloring scheme using k colors that makes a countable collection $\mathcal{L}$ non-uniformly identifiable in the limit with colors if and only if every intersection of $k+1$ distinct languages in $\mathcal{L}$ is finite.*

This scheme must depend on the collection: for every collection independent terminal coloring scheme, some collection that satisfies this condition is not non-uniformly identifiable (Proposition A.6). As with trace colorings, uniform identification can still fail, even if we enforce a global bound on the intersection size from Theorem 3.10:

Proposition 3.11. *There exists a language class in which every three distinct languages share at most one string, but which is not uniformly identifiable in the limit with any terminal coloring scheme using any finite number of colors.*

The idea is that although every three languages share at most one string, any finite palette forces some pair of languages to share many strings on which their colors agree. See Appendix 3.4 for the construction and proof.

3.5 Characterization of Identification in the Limit with Negative Examples

Finally, we study identification with negative examples (see Section 1.1 for motivation): the identifier sees strings both in and outside the target, each labeled with whether it belongs to the target. This is Gold's model of learning from an informant (Gold, 1967). Formally, the adversary selects a target $L \in \mathcal{L}$ and an enumeration $S = ((x_1, b_1), (x_2, b_2), \dots)$ in which every string of Σ^* appears at least once, with $b_t = 1$ if $x_t \in L$ and $b_t = 0$ otherwise. Then, in round t, the identifier sees (x_t, b_t) and outputs an index. The identifier's goal is to correctly guess the index of the target language, and in particular to succeed in the limit (for all but a finite number of rounds). Identification in the limit and its non-uniform and uniform versions are defined as in Definitions 1 to 3.

Negative examples greatly enlarge the set of identifiable classes: indeed, simply outputting the first language consistent with the labeled examples identifies every countable class in the limit (Gold, 1967). This rule works as it cannot get stuck on a superset S of the target, since a negative example eventually rules S out. We next show that for non-uniform identification, negative examples surprisingly have the *opposite* effect: fewer classes are non-uniformly identifiable with them than without.

Theorem 3.12. *Language class $\mathcal{L}$ is non-uniformly identifiable in the limit with negative examples if and only if for all distinct $L, L' \in \mathcal{L}$, $\left|(L \cap L') \cup (L^C \cap L'^C)\right| < \infty$.*

Intuitively, the set of identifiable collections shrinks because the thresholds also count negative examples, the adversary can delay distinguishing examples by presenting strings that lie in both or in neither of two candidate languages. For example, the class $\mathcal{L} = \{\{1\}, \{2\}\}$ over the domain $\mathbb{N}$ is non-uniformly identifiable in the standard setting, but not non-uniformly identifiable with negative examples, since the adversary can present arbitrarily many strings outside both languages. See Theorem A.3 for a characterization of uniform identification in the limit with negative examples, which requires a uniform bound on the size of these sets. Both proofs appear in Appendix A.2.

4 Conclusion

We study non-uniform identification in the limit, which, perhaps surprisingly, has received little attention despite the long line of work on identification in the limit and its many variants (Gold, 1967; Angluin, 1980; Jain et al., 1999; Lange et al., 2008; Charikar et al., 2025; 2026b). We characterize the collections that are non-uniformly identifiable (Theorem 3.1) and give a second, union-form characterization (Theorem 3.4). This union-form characterization has a similar form to non-uniform characterizations of PAC and online learning. It, however, has some crucial differences from the characterization of non-uniform generation (Li et al., 2025) (Section 1.1), adding to the differences between these two seemingly similar problems. As one might expect, our characterization is more restrictive than Angluin's characterization of identification in the limit (Angluin, 1980). We then ask whether classic and recent variants of identification close this gap (**Q2**). They are of two types: those that relax the output, by allowing a list of k guesses (Charikar et al., 2025) or a guess that differs from the target on finitely many strings (Case & Lynes, 1982), and those that give the learner auxiliary information: colors on examples (Charikar et al., 2026a;b) or negative examples (Gold, 1967). We characterize non-uniform identification in each variant (Theorems 3.5, 3.7, 3.8 and 3.12). Most of these variants do not close the gap, and negative examples even shrink the class of non-uniformly identifiable collections. One exception is colors on prefixes of examples, which is motivated by auxiliary information provided by chain-of-thought traces during post-training (Peng et al., 2026; Charikar et al., 2026b). These do close the gap: surprisingly, with just two colors, every countable collection of infinite languages non-uniformly identifiable (Theorem 3.9).

AI USE STATEMENT

Generative AI tools assisted with copyediting the manuscript and formatting it. The authors reviewed the AI-assisted material and take responsibility for the paper's content.

REFERENCES

Dana Angluin. Inductive inference of formal languages from positive data. *Information and control*, 45(2):117–135, 1980.

Dana Angluin. Identifying languages from stochastic examples. Technical report, Yale University, Department of Computer Science, 1988.

Dana Angluin and Carl H. Smith. Inductive inference: Theory and methods. *ACM Computing Surveys*, 15(3):237–269, 1983. doi: 10.1145/356914.356918.

Yannan Bai, Debmalya Panigrahi, and Ian Zhang. Language generation in the limit: Noise, loss, and feedback. In *Proceedings of the 2026 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2026. doi: 10.1137/1.9781611978971.31. URL https://arxiv.org/abs/2507.15319.

Gyora M. Benedek and Alon Itai. Nonuniform learnability. *Journal of Computer and System Sciences*, 48(2):311–323, 1994. doi: 10.1016/S0022-0000(05)80005-4. URL https://www.sciencedirect.com/science/article/pii/S0022000005800054.

John Case and Christopher Lynes. Machine inductive inference and language identification. In *Automata, Languages and Programming (ICALP 1982)*, volume 140 of *Lecture Notes in Computer Science*, pp. 107–115. Springer, 1982. doi: 10.1007/BFb0012761.

John Case and Carl Smith. Comparison of identification criteria for machine inductive inference. *Theoretical Computer Science*, 25(2):193–220, 1983.

Moses Charikar and Chirag Pabbaraju. Exploring facets of language generation in the limit. In *Proceedings of Thirty Eighth Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pp. 854–887, 2025a. URL https://proceedings.mlr.press/v291/charikar25a.html.

Moses Charikar and Chirag Pabbaraju. Pareto-optimal non-uniform language generation. *arXiv preprint arXiv:2510.02795*, 2025b.

Moses Charikar, Chirag Pabbaraju, and Ambuj Tewari. A characterization of list language identification in the limit. *arXiv preprint arXiv:2511.04103*, 2025.

Moses Charikar, Jon Kleinberg, and Chirag Pabbaraju. Globally consistent coloring schemes for language identification. *arXiv preprint arXiv:2607.11606*, 2026a.

Moses Charikar, Jon Kleinberg, and Chirag Pabbaraju. Language identification with succinct machine-independent traces. In Steve Hanneke and Tor Lattimore (eds.), *Proceedings of Thirty Ninth Conference on Learning Theory*, volume 336 of *Proceedings of Machine Learning Research*, pp. 1054–1074. PMLR, 29 Jun–03 Jul 2026b. URL https://proceedings.mlr.press/v336/charikar26b.html.

E Mark Gold. Language identification in the limit. *Information and control*, 10(5):447–474, 1967.

Steve Hanneke, Amin Karbasi, Anay Mehrotra, and Grigoris Velegkas. On union-closedness of language generation. In *Advances in Neural Information Processing Systems*, volume 38, 2025a. URL https://arxiv.org/abs/2506.18642.

Steve Hanneke, Shay Moran, and Maximilian Thiessen. Marginal-nonuniform PAC learnability. In *Advances in Neural Information Processing Systems*, volume 38, 2025b. URL https://openreview.net/forum?id=aoVCFtox89.

Steve Hanneke, Amirreza Shaeiri, and Hongao Wang. Non-uniform multiclass learning with bandit feedback. In *Advances in Neural Information Processing Systems*, volume 38, 2025c. doi: 10.52202/085713-4962. URL https://papers.nips.cc/paper_files/paper/2025/hash/da3a6154db2eda4ca55c7d2280c5a9ae-Abstract-Conference.html.

Sanjay Jain and Arun Sharma. Computational limits on team identification of languages. *Information and Computation*, 130(1):19–60, 1996.

Sanjay Jain, Daniel N. Osherson, James S. Royer, and Arun Sharma. *Systems that learn: an introduction to learning theory*. MIT Press, 2 edition, 1999. URL https://mitpress.mit.edu/9780262100779/systems-that-learn/.

Alkis Kalavasis, Anay Mehrotra, and Grigoris Velegkas. On the limits of language generation: Trade-offs between hallucination and mode collapse. In *Proceedings of the 57th Annual ACM Symposium on Theory of Computing*, 2025. URL https://arxiv.org/abs/2411.09642.

Alkis Kalavasis, Anay Mehrotra, and Grigoris Velegkas. On characterizations for language generation: Interplay of hallucinations, breadth, and stability. In Matus Telgarsky and Jonathan Ullman (eds.), *Proceedings of The 37th International Conference on Algorithmic Learning Theory*, volume 313 of *Proceedings of Machine Learning Research*, pp. 1–49. PMLR, 2026. URL https://proceedings.mlr.press/v313/kalavasis26a.html.

Jon Kleinberg and Sendhil Mullainathan. Language generation in the limit. *Advances in Neural Information Processing Systems*, 37:66058–66079, 2024.

Steffen Lange, Thomas Zeugmann, and Sandra Zilles. Learning indexed families of recursive languages from positive data: A survey. *Theoretical Computer Science*, 397(1–3):194–232, 2008. doi: 10.1016/j.tcs.2008.02.030.

Jiaxun Li, Vinod Raman, and Ambuj Tewari. Generation through the lens of learning theory. In Nika Haghtalab and Ankur Moitra (eds.), *Proceedings of Thirty Eighth Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pp. 4740–4776. PMLR, 2025. URL https://proceedings.mlr.press/v291/raman25a.html.

Nick Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine Learning*, 2(4):285–318, 1988. doi: 10.1007/BF00116827.

Zhou Lu. When is inductive inference possible? In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://arxiv.org/abs/2312.00170v2.

Anay Mehrotra, Grigoris Velegkas, Xifan Yu, and Felix Zhou. Language generation with infinite contamination. In *Proceedings of Thirty Ninth Conference on Learning Theory*, volume 336 of *Proceedings of Machine Learning Research*, pp. 5053–5112, 2026. URL https://proceedings.mlr.press/v336/mehrotra26a.html.

Niklas Muennighoff, Alexander M. Rush, Boaz Barak, Teven Le Scao, Aleksandra Piktus, Nouamane Tazi, Sampo Pyysalo, Thomas Wolf, and Colin Raffel. Scaling data-constrained language models. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Charlotte Peale, Vinod Raman, and Omer Reingold. Representative language generation. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 48518–48541, 2025. URL https://proceedings.mlr.press/v267/peale25a.html.

Binghui Peng, Amin Saberi, and Grigoris Velegkas. Language identification in the limit with computational trace. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=1OAGf7ntSE.

Leonard Pitt. Probabilistic inductive inference. *Journal of the ACM*, 36(2):383–433, 1989.

Ananth Raman and Vinod Raman. Generation from noisy examples. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 51079–51093, 2025. URL https://arxiv.org/abs/2501.04179.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014. URL https://www.cs.huji.ac.il/~shais/UnderstandingMachineLearning.

Carl H. Smith. The power of pluralism for automatic program synthesis. *Journal of the ACM*, 29 (4):1144–1165, 1982.

Frank Stephan. Noisy inference and oracles. *Theoretical Computer Science*, 185(1):129–157, 1997. doi: 10.1016/S0304-3975(97)00018-2.

Takashi Tabe and Thomas Zeugmann. Two variations of inductive inference of languages from positive data. Technical Report RIFIS-TR-CS-105, Research Institute of Fundamental Information Science, Kyushu University, 1995.

Xinyu Zhu, Mengzhou Xia, Zhepei Wei, Wei-Lin Chen, Danqi Chen, and Yu Meng. The surprising effectiveness of negative reinforcement in LLM reasoning. In *Advances in Neural Information Processing Systems*, volume 38, pp. 126546–126573, 2025. URL https://proceedings.nips.cc/paper_files/paper/2025/hash/b8026fe01f5f8f7a80a1a9bf01d08cad-Abstract-Conference.html.

A Full Results and Proofs

A.1 Full Results and Proofs for Identification in the Limit

In this section, we present full proofs for our results characterizing non-uniform and uniform identification in the standard setting. We begin with our characterization of non-uniform identification.

Proof of Theorem 3.1 (Characterization of non-uniform identifiability). We first prove sufficiency and then necessity.

Part 1 (Sufficiency): Suppose both conditions (1) and (2) hold. Consider the following algorithm: the identifier in round t, outputs the index of the lowest-indexed language L_j such that $S_t \subseteq L_j$ and $S_t \nsubseteq L'$ for all $L' \in \mathcal{L}$ with $L' \subsetneq L_j$, where S_t is the set of examples seen so far. Such an L_j exists for valid S_t by Condition (2); on other inputs, output index 1. In other words, the identifier outputs the first language consistent with the examples seen so far for which it is sure the target is not one of its subsets. Now fix target language L_ℓ. We claim this algorithm identifies L_ℓ in a uniform number of distinct examples independent of its enumeration. Consider any L_i with $i < \ell$. If $L_\ell \not\subset L_i$, we must see an example from $L_\ell \setminus L_i$ after at most $s_i := |L_\ell \cap L_i| + 1$ distinct examples and our identifier will rule out L_i. If $L_\ell \subset L_i$, the identifier will never output L_i since $S_t \subseteq L_\ell$. Set $s_i = 0$ in this case and interpret an empty maximum as 0. This shows that no language appearing before L_ℓ is output by the identifier after seeing $\max_{i<\ell} s_i$ distinct examples. Finally, we must ensure the identifier rules out all subsets of L_ℓ in a uniform number of distinct examples (so that the identifier will output ℓ after ruling out all previous indices). Let d be the bound in Condition (2) for L_ℓ. This is true because Condition (2) ensures that no matter which $d+1$ distinct elements we see, they cannot be consistent with any of the subsets of L_ℓ. Hence, after at most $d^*(L_\ell) = \max(d+1, \max_{i<\ell} s_i)$ distinct examples, the identifier will only output ℓ. If L_ℓ is finite and this bound exceeds $|L_\ell|$, then once $S_t = L_\ell$, L_ℓ is eligible and every earlier language is inconsistent or ineligible, so it is identified in the limit. Therefore, $\mathcal{L}$ is non-uniformly identifiable in the limit.

Part 2 (Necessity): We first show necessity of Condition (2). Suppose for the sake of contradiction that Condition (2) were violated for language L but an identifier claims to identify L with uniform bound $d^*(L)$. The adversary can select some $L' \subset L$ such that $|L'| > d^*(L)$ and output $d^*(L)$ examples from L'. Then the identifier must output L in all future rounds along any continuation within L', since each finite prefix can also be extended to an enumeration of L. However, the adversary could then continue this sequence to an enumeration of L' and force the identifier to fail to identify L'. Next, we similarly show necessity of Condition (1). Suppose there exist two languages $L, L' \in \mathcal{L}$ with infinite intersection. Suppose for the sake of contradiction that $\mathcal{L}$ were non-uniformly identifiable. Let $d' = \max(d^*(L), d^*(L'))$. Consider a sequence S consisting of d' distinct examples such that $S \subseteq L \cap L'$. This sequence can be continued to an enumeration of L, which means the identifier must output the index of L in all rounds after d' once it sees S since it claims to non-uniformly identify L in the limit. However, this contradicts the identifier non-uniformly identifying L', since S can also be continued to an enumeration of L' and so the identifier should output the index of L' in all rounds after d'. Therefore, if $\mathcal{L}$ does not have finite pairwise intersections, then it is not non-uniformly identifiable in the limit. □

Next, we state and prove the analogous characterization for uniform identifiability.

Theorem A.1 (Characterization of Uniform Identifiability)**.** *Language class $\mathcal{L}$ is uniformly identifiable in the limit if and only if there exists some $d \in \mathbb{N}$ such that for all distinct $L, L' \in \mathcal{L}$, $|L \cap L'| < d$.*

Proof of Theorem A.1 (Characterization of uniform identifiability). The proof is an easy modification of Theorem 3.1.

Part 1 (Sufficiency): In our proof of the non-uniform characterization, notice that we got non-uniform bounds $d^*(L_\ell) = \max(d+1, \max_{i<\ell} s_i)$ for each $L_\ell \in \mathcal{L}$. Under the stated bound, every proper subset $L' \subsetneq L$ has $|L'| = |L \cap L'| < d$. If $|L \cap L'| < d$ for all distinct $L, L' \in \mathcal{L}$, we see

that $d^*(L_\ell) = \max(d+1, \max_{i<\ell} s_i) \leq \max(d+1, d+1) = d+1$. This is uniform over all target languages, so $\mathcal{L}$ is uniformly identifiable in the limit.

Part 2 (Necessity): Suppose an identifier claimed to uniformly identify language class $\mathcal{L}$ after seeing d^* distinct examples, but there exists no uniform bound on the intersection sizes of distinct languages in $\mathcal{L}$. Let distinct $L, L' \in \mathcal{L}$ such that $|L \cap L'| > d^*$. Consider a sequence S consisting of $d' > d^*$ distinct examples such that $S \subseteq L \cap L'$. This sequence can be continued to an enumeration of L, which means the identifier must output the index of L in all rounds after d' once it sees S since it claims to uniformly identify L in the limit with bound $d^* < d'$. However, this contradicts the identifier uniformly identifying L' with bound d^*, since S can also be continued to an enumeration of L' and so the identifier should output the index of L' in all rounds after d'. Therefore, if $\mathcal{L}$ does not have uniformly bounded pairwise intersections, then it is not uniformly identifiable in the limit. □

Finally, we prove equivalences between these characterizations and our union form, first in the non-uniform case:

Proof of Theorem 3.4 (Union form characterization of non-uniform identifiability). We show that this union form characterization is equivalent to the characterization in Theorem 3.1.

Part 1 (Sufficiency): Suppose $\mathcal{L} = \bigcup_{n=1}^{\infty} \mathcal{C}_n$ and the conditions in Theorem 3.4 are satisfied. Given any distinct $L, L' \in \mathcal{L}$ we must have $L, L' \in \mathcal{C}_i$ for some i, and since $I(\mathcal{C}_i) < \infty$ their intersection must be finite. This shows that Condition (1) in Theorem 3.1 is satisfied. Next, consider any L. Let $L \in \mathcal{C}_i$ for some i. Then by Condition (2) in Theorem 3.4 and the fact that our sequence of classes is non-decreasing, all subsets of L must be contained in $\mathcal{C}_i$. Then, since $I(\mathcal{C}_i) < \infty$, these proper subsets must have bounded size (otherwise intersections of L with its proper subsets would have arbitrarily large size and $I(\mathcal{C}_i)$ would not be finite). This shows that Condition (2) in Theorem 3.1 is satisfied and finishes the sufficiency proof.

Part 2 (Necessity): Suppose all distinct languages in $\mathcal{L}$ intersect finitely and each language has bounded proper subsets. Fix an ordering of $\mathcal{L}$. Here, "subsets" means members of $\mathcal{L}$. Let $\mathcal{C}_1$ consist of L_1 together with all subsets of L_1. Then let $\mathcal{C}_i$ consist of $\mathcal{C}_{i-1}$ together with the lowest indexed language in $\mathcal{L}$ not already added to a class, along with all subsets of that language. Once all languages have been added, keep the sequence constant. It is easy to verify that $\mathcal{C}_{i-1} \subseteq \mathcal{C}_i$ and $\mathcal{L} = \bigcup_{n=1}^{\infty} \mathcal{C}_n$. Furthermore, $I(\mathcal{C}_i) < \infty$ for all i: each $\mathcal{C}_i$ consists of subsets of finitely many selected languages, whose pairwise intersections are finite by Condition (1), while Condition (2) bounds their proper subsets. Finally, the second condition in Theorem 3.4 is satisfied since all subsets of L are added as soon as L is added (and so we never add a language that's a subset of a language we've seen in an earlier class). □

Remark A.2 (Nesting and the second condition in Theorem 3.4 are necessary)**.** Let A be an infinite set of strings and let $x \neq y$ be strings outside A. The collection $\{A \cup \{x\}, A \cup \{y\}\}$ is the union of two uniformly identifiable singletons, each closed under sublanguages, but it is not non-uniformly identifiable because its languages share A. Hence the classes $\mathcal{C}_n$ must be nested. The collection consisting of $\mathbb{N}$ and disjoint finite sets $B_1, B_2, \ldots \subseteq \mathbb{N}$ with $|B_n| = n$ is the nested union of the uniformly identifiable collections $\mathcal{C}_n = \{\mathbb{N}, B_1, \ldots, B_n\}$, but it is not non-uniformly identifiable because $\mathbb{N}$ has arbitrarily large proper sublanguages in the collection. Here the second condition fails, since $B_{n+1} \in \mathcal{C}_{n+1} \setminus \mathcal{C}_n$ is a subset of $\mathbb{N} \in \mathcal{C}_n$; hence this condition cannot be dropped.

We next show that the intersection-dimension characterization of uniform identification in Theorem 3.3 is equivalent to the pairwise-intersection characterization in Theorem A.1.

Proof of Theorem 3.3 (Union form characterization of uniform identifiability). We show that the union form characterization is equivalent to the characterization in Theorem A.1

Part 1 (Sufficiency) Suppose $I(\mathcal{L}) < \infty$. This implies for every distinct $L, L' \in \mathcal{L}$, we must have $|L \cap L'| \leq I(\mathcal{L})$, therefore satisfying the condition in Theorem A.1 for any $d \in \mathbb{N}$ with $d > I(\mathcal{L})$.

Part 2 (Necessity) Suppose there exists $d \in \mathbb{N}$ such that $|L \cap L'| < d$ for all distinct $L, L' \in \mathcal{L}$. By definition, we see that $I(\mathcal{L}) \leq d < \infty$, which is the condition in Theorem 3.3. □

A.2 Full Results and Proofs for Identification in the Limit with Negative Examples

In this section, we present full results and proofs for identification in the limit with negative examples. Again, we begin with a proof of our characterization of non-uniform identification:

Proof of Theorem 3.12 (Characterization of non-uniform identifiability with negative examples). We first prove sufficiency and then necessity.

Part 1 (Sufficiency): Suppose the condition holds. Consider the following algorithm: the identifier in round t, outputs the index of the lowest-indexed language L_j that is consistent with the labels of the examples seen so far. Now fix target language L_ℓ. We claim this algorithm identifies L_ℓ in a uniform number of distinct examples independent of its enumeration. Consider any L_i with $i < \ell$. Since $s_i := \left|(L_\ell \cap L_i) \cup (L_\ell^C \cap L_i^C)\right| + 1 < \infty$, after seeing s_i distinct examples, the identifier must have seen an example whose label disagrees with L_i and so will eliminate L_i. Furthermore, L_ℓ will never be eliminated, so after at most $d^*(L_\ell) = \max_{i<\ell} s_i$ distinct examples, the identifier will only output ℓ. Therefore, $\mathcal{L}$ is non-uniformly identifiable in the limit with negative examples.

Part 2 (Necessity): Suppose there exist two languages $L, L' \in \mathcal{L}$ such that $(L \cap L') \cup (L^C \cap L'^C)$ is infinite. Suppose for the sake of contradiction that $\mathcal{L}$ were non-uniformly identifiable. Let $d' = \max(d^*(L), d^*(L'))$. Consider a sequence S consisting of d' distinct examples whose labels agree with both L and L' (examples from $(L \cap L') \cup (L^C \cap L'^C)$). This sequence can be continued to an enumeration consistent with L being the target, which means the identifier must output the index of L in all rounds after d' once it sees S since it claims to non-uniformly identify L in the limit. However, this contradicts the identifier non-uniformly identifying L', since S can also be continued to an enumeration consistent with L' being the target and so the identifier should output the index of L' in all rounds after d'. Therefore, if Condition (1) does not hold, then $\mathcal{L}$ is not non-uniformly identifiable in the limit with negative examples. □

Next, we state and prove the analogous uniform characterization:

Theorem A.3. *Language class $\mathcal{L}$ is uniformly identifiable in the limit with negative examples if and only if there exists some $d \in \mathbb{N}$ such that for all distinct $L, L' \in \mathcal{L}$, $\left|(L \cap L') \cup (L^C \cap L'^C)\right| < d$.*

Proof. This is again an easy modification of the non-uniform proof.

Part 1 (Sufficiency): From our proof of the non-uniform case, we get non-uniform bounds $d^*(L_\ell) = \max_{i<\ell} s_i \leq d + 1$, which is uniform over all targets and so $\mathcal{L}$ is uniformly identifiable in the limit.

Part 2 (Necessity): Suppose an identifier claimed to uniformly identify language class $\mathcal{L}$ after seeing d^* distinct examples, but there exists no uniform bound on $\left|(L \cap L') \cup (L^C \cap L'^C)\right|$ for distinct $L, L' \in \mathcal{L}$. Let $L, L' \in \mathcal{L}$ such that $\left|(L \cap L') \cup (L^C \cap L'^C)\right| > d^*$. Consider a sequence S consisting of $d' > d^*$ distinct examples such that $S \subseteq (L \cap L') \cup (L^C \cap L'^C)$. This sequence can be continued to be consistent with either L or L', which means the identifier must fail to achieve its claimed uniform guarantee of d^*. Therefore, if $\mathcal{L}$ does not have such uniformly bounded intersections, then it is not uniformly identifiable in the limit with negative examples. □

Lastly, we note that non-uniformly identifiable classes when using negative examples are a strict subset of those that are non-uniformly identifiable in the standard setting (with only positive examples). If Theorem 3.12 holds, any two distinct languages have finite intersection. Moreover, two finite languages would have infinitely many common negative examples, so $\mathcal{L}$ contains at most one

finite language. Every proper sublanguage $L' \in \mathcal{L}$ of L is finite, since $L' = L \cap L'$, and thus there is at most one such L'. Its size gives the bound required by Theorem 3.1.

We give an example of a language class which is non-uniformly identifiable with only positive examples but not when using negative examples in Section 3.5.

A.3 Full Results and Proofs for Identification in the Limit with Colors

In this section, we present full results and proofs for identification in the limit with colors.

To reason about identification with colors, it will be useful to reduce to identification in the standard setting, which we do with the following construction:

Definition 6. *Let $\mathcal{L}$ be a language class over finite alphabet Σ and let $\{c_L\}_{L\in\mathcal{L}}$ be coloring functions over finite palette P. Let the* dual class*, denoted $\overline{\mathcal{L}}$, be defined as a language class over a finite alphabet encoding pairs consisting of languages $\overline{L} := \{(x, c_L(x)) : x \in L\}$ for all $L \in \mathcal{L}$.*

In other words, the dual class contains languages consisting of ordered pairs of strings together with their colors. Only countably many colored examples occur, so these pairs can be encoded over a finite alphabet. Now it is apparent that the identification game played with $\mathcal{L}$ and colors $\{c_L\}_{L\in\mathcal{L}}$ is equivalent to the standard identification game played with $\overline{\mathcal{L}}$ (since the information given to the identifier is equivalent). Then it is apparent that the characterization of (non-uniform or uniform) identifiability of $\mathcal{L}$ is equivalent to the characterization of (non-uniform or uniform) identifiability of $\overline{\mathcal{L}}$.

Combining this with our characterizations from Theorem 3.1 and Theorem A.1, we directly obtain the characterization in Theorem 3.8 as well as the following analog for uniform identification:

Theorem A.4 (Characterization of uniformly identifiable classes with colors.)**.** *Language class $\mathcal{L}$ is uniformly identifiable in the limit with coloring functions $\{c_L\}_{L\in\mathcal{L}}$ if and only if there exists $d \in \mathbb{N}$ such that $|\{x \in L \cap L' : c_L(x) = c_{L'}(x)\}| < d$ for all distinct $L, L' \in \mathcal{L}$.*

Next, we present our results on colorability. First, we present our positive result for trace colorings that make every class non-uniformly identifiable:

Proof of Theorem 3.9. Fix an ordering $\mathcal{X} = \{x_1, x_2, \dots\}$. Let

$$\mathrm{tr}_L(u) = \begin{cases} \mathbb{1}_{|L|=\infty} & \text{if } |u| = 0, \\ \mathbb{1}_{x_{|u|}\in L} & \text{if } |u| \geq 1. \end{cases}$$

Consider any $L, L' \in \mathcal{L}$ with $L \neq L'$. Since these two languages disagree on some example x_j, their colors will differ on all strings of length at least j, so their colored intersection will be finite. Now consider $L' \subset L$ with $\overline{L'} \subset \overline{L}$. We just showed that the colored intersection of L' and L must be finite, so $|L'| < \infty$. If L is infinite, its color at the empty prefix differs from that of L' on every string, contradicting $\overline{L'} \subset \overline{L}$. Hence there are never any proper colored subsets of infinite languages; those of finite languages have size at most $|L|$. Therefore, under this coloring, all countable collections become non-uniformly identifiable with these trace colors. □

Next, we show that not all collections are colorable with trace colors to be uniformly identifiable in the limit (and leave the full characterization as an open question).

Proposition A.5. *The language class $\mathcal{L} = \{2\mathbb{N}, 2\mathbb{N}\cup\{1\}, 2\mathbb{N}\cup\{3\}, 2\mathbb{N}\cup\{5\}, \dots\}$ is not uniformly identifiable in the limit with any collection dependent trace coloring scheme using a finite palette.*

Proof of Proposition A.5. Suppose $\mathcal{L}$ were uniformly identifiable in the limit by color traces with uniform bound d^*. Choose $d^* + 1$ distinct examples from $2\mathbb{N}$. Since the palette is finite, there are only finitely many ways to color these $d^* + 1$ examples. However, there are infinitely many languages in $\mathcal{L}$ sharing these examples, and so two languages must have colored intersection at least $d^* + 1$, contradicting colored uniform identification with bound d^*. □

Moving on to terminal colorings, we present here a proof of our characterization of classes that are colorable with terminal colorings to be non-uniformly identifiable (which we recall are those collections with finite intersections of $k+1$ intersections).

Proof of Theorem 3.10. We first show necessity. Suppose $L_1, L_2, \ldots, L_{k+1}$ are languages in $\mathcal{L}$ with infinite common intersection. For $i < j$, let $E_{i,j}$ denote the set of all $x \in L_i \cap L_j$ such that $c_{L_i}(x) = c_{L_j}(x)$. We claim that some $E_{i,j}$ must be infinite, which contradicts non-uniform identifiability in the limit with color traces. Consider $x \in L_1 \cap \cdots \cap L_{k+1}$. Since we have $k+1$ languages but only k colors, x must belong to some $E_{i,j}$. Then $L_1 \cap \cdots \cap L_{k+1} \subseteq \bigcup_{i<j} E_{i,j}$. The left-hand side is infinite by assumption, and since there are a finite number of $E_{i,j}$ where $i, j \in [1, k+1]$, some $E_{i,j}$ must be infinite.

Conversely, we present a collection dependent terminal coloring scheme using k colors that makes $\mathcal{L}$ non-uniformly identifiable in the limit with color traces, provided all common intersections of $k+1$ languages in $\mathcal{L}$ are finite. Let $\mathcal{L} = \{L_1, L_2, \ldots\}$. For each $x \in \mathcal{X}$, let $L_{i_1}, L_{i_2}, \ldots$ be the languages containing x, where $i_1 < i_2 < \cdots$. Then we define colorings $c_{L_{i_j}}(x) = j \mod k$. Now consider any $L_a, L_b \in \mathcal{L}$ with $a < b$ and both L_a and L_b infinite. Suppose $x \in L_a \cap L_b$ and $c_{L_a}(x) = c_{L_b}(x)$. By our coloring scheme, this means their positions in this list differ by a multiple of k and there must be at least $k-1$ other languages between a and b that also contain x . In other words, there must be $k+1$ languages between L_a and L_b, inclusive, which all contain x . Since there are a finite number of combinations of $k+1$ languages between L_a and L_b and each one must have finite colored intersection by assumption, we see that L_a and L_b must have finite colored intersection, as desired. Now consider any finite language L_i. Choose a coloring of its strings that differs from the coloring of each infinite $L_j \supseteq L_i$ with $j \leq |L_i|$. Since there are $k^{|L_i|}$ possible colorings of L_i but only $|L_i|$ colorings that are not allowed, a permissible coloring always exists (when $k \geq 2$). Hence for any infinite languages L_j, the finite languages that agree with its coloring must have size at most j. Therefore, this collection dependent terminal coloring scheme makes $\mathcal{L}$ non-uniformly identifiable in the limit with colors. □

We also show that these colorings functions must be collection dependent:

Proposition A.6. *Consider any collection independent terminal coloring scheme $\{c_L\}_{L \subset \Sigma^*}$ over a finite palette of $k \geq 2$ colors. Then one can always find a subset of languages which satisfies every $k+1$ intersections being finite yet is not non-uniformly identifiable with colors given by $\{c_L\}_{L \subset \Sigma^*}$.*

Proof. Consider any $k+1$ distinct languages $L_1, \ldots, L_{k+1}$ which intersect on an infinite set A. For every string in this infinite set, two languages must color this string in the same way. There are only finitely many languages, so there must be two languages L_i, L_j which color infinitely many strings in A the same color, making the two-language collection $\{L_i, L_j\}$ (which satisfies every $k+1$ intersections being finite) not non-uniformly identifiable in the limit. □

Lastly, we show that upgrading the bounds from Theorem 3.10 to be uniform does not achieve uniform identifiability (unlike our other results).

Proof of Proposition 3.11. We want to construct a language class such that each individual triplet of languages intersects on a small subset, but in aggregate there are so many strings that intersections are forced to be unbounded no matter what finite-palette coloring scheme is used.

For each prime p, encode as strings the p^3 polynomials $f(x) = ax^2 + bx + c$ where $a, b, c \in \mathbb{F}_p$ (where $\mathbb{F}_p$ is the field with p elements). Then consider p^2 languages where $L_{(u,v)} = \{f : f(u) = v\}$ for all $u, v \in \mathbb{F}_p$. Do this for all primes (encoding the polynomials such that the strings are disjoint for each prime).

Since three distinct points lie on at most one shared polynomial, intersections of three distinct languages have size at most one and hence three-wise intersections are uniformly bounded in size. However, we now show that even this is not enough to achieve uniform colorability by any finite-palette terminal coloring scheme.

Consider any terminal coloring of this language class using a palette size of k. Consider fixed prime p. Each polynomial f appears in p languages (one for each input u). Let $n_1, \ldots, n_k$ be how many times f is colored by each color within these p languages. Then the number of pairwise intersections f contributes to in the dual language class is

$$\begin{aligned}\sum_{i=1}^{k} \binom{n_i}{2} &= \frac{1}{2}\sum_{i=1}^{k}(n_i^2 - n_i) \\ &= \frac{1}{2}\left(\sum_{i=1}^{k} n_i^2 - p\right) \\ &\geq \frac{1}{2}\left(\frac{p^2}{k} - p\right).\end{aligned}$$

Then multiplying by p^3 counts the total size of pairwise intersections in the dual language class within the p^2 languages created from p. If there existed a uniform bound S on the intersection sizes of the dual language class, then we must have

$$S \cdot \binom{p^2}{2} \geq \frac{p^3}{2}\left(\frac{p^2}{k} - p\right) \implies S \geq \frac{p}{p^2 - 1}\left(\frac{p^2}{k} - p\right).$$

However, the right hand size is unbounded as $p \to \infty$, which shows that no finite-palette terminal coloring scheme can make this language class uniformly identifiable in the limit with colors. □

Note that all languages here are finite. If we wanted a collection of infinite languages, then we could simply add a separate set of infinite strings to each language such that these infinite sets are pairwise disjoint.

A.4 Full Results and Proofs for Approximate Identification in the Limit

In this section, we present full results and proofs for approximate identification in the limit. Recall that stable and unstable identification coincide here. Any stable identifier is also unstable, and any unstable identifier can be made stable by outputting a fixed representative of the equivalence class of its original output under the equivalence relation of having finite symmetric difference. Since an unstable identifier eventually always outputs languages which have finite symmetric difference with the target, these outputs are in the same equivalence class and hence this procedure will make the identifier converge to a single language, satisfying stability.

Therefore, we only need to prove characterizations for unstable identifiers, which is how we will proceed.

Lange et al. (2008) attribute the characterization of classes that are approximately identifiable in the limit to Tabe & Zeugmann (1995). We give its statement and proof below for completeness. This characterization uses a "weak tell-tale", which only has to distinguish L from its proper sublanguages that differ from it on infinitely many strings.

Theorem A.7. *Language class $\mathcal{L}$ is approximately identifiable in the limit (with or without stability) if and only if for every language $L \in \mathcal{L}$, there exists "weak tell-tale" set $T \subseteq L$ of finite size $|T| < \infty$ such that for all $L' \in \mathcal{L}$, if $L' \subset L$ then either $T \not\subseteq L'$ or $|L \setminus L'| < \infty$.*

Proof of Theorem A.7 (Characterization of approximate identifiability). Part 1 (Sufficiency): Suppose the weak tell-tale condition holds. Given language class $\mathcal{L}$, we can partition it into equivalence classes under the equivalence relation of having finite symmetric difference. Now consider the identifier that outputs the lowest indexed representative of the first language consistent with the examples seen so far and for which the weak tell-tale set has been seen. Fix target language L_ℓ. Consider any L_i with $i < \ell$. If $L_i \subset L_\ell$, then the identifier will eventually see an element from $L_\ell \setminus L_i$ and will eliminate L_i. If $L_\ell \subset L_i$ and $|L_\ell \triangle L_i| = \infty$, then the weak tell-tale for L_i cannot be contained in L_ℓ and so the identifier will never guess the equivalence class of L_i. The identifier will thus converge

to the index of the first representative of the target and hence successfully approximately identify $\mathcal{L}$ in the limit (with stability).

Part 2 (Necessity): Suppose the weak tell-tale condition were violated for language L_ℓ. Let the adversary begin enumerating L_ℓ. If an identifier approximately identified L_ℓ in the limit (with or without stability), then it must eventually output the index of a language which has finite symmetric difference with L_ℓ. Denote the examples seen so far as S_a. Then there must be another language $L_a \supset S_a$ such that $|L_\ell \triangle L_a| = \infty$, otherwise S_a would be a weak tell-tale for L_ℓ. The adversary can then continue enumerating L_a, after which at some point the identifier must output the index of a language which has finite symmetric difference with L_a and therefore has infinite symmetric difference with L_ℓ. Then the adversary can again continue enumerating L_ℓ and repeat this procedure, eventually enumerating L_ℓ but forcing the identifier to never stop outputting indices of languages with infinite symmetric difference with L_ℓ. □

Next, we prove our characterization of non-uniform approximate identification.

Proof of Theorem 3.7. We consider identification without stability, where the identifier does not have to converge to a single index (which we have shown is equivalent to identification with stability).

Part 1 (Sufficiency): We can use the algorithm which outputs the first language that is consistent with all examples seen so far such that we have eliminated all proper subsets with infinite symmetric difference. Suppose the target language is L_i. Then consider what happens after seeing $\max(d, \max_{j<i, |L_j \cap L_i| < \infty, L_i \not\subset L_j} |L_j \cap L_i|)$ distinct examples. Seeing more than d distinct examples guarantees that we have ruled out all subsets of L_i that have infinite symmetric difference with L_i. Seeing more than $\max_{j<i, |L_j \cap L_i| < \infty, L_i \not\subset L_j} |L_j \cap L_i|$ examples guarantees that we have ruled out all languages before L_i that intersect finitely with L_i and are not subsets or supersets of L_i. Finally, we will never guess a superset of L_i unless it has finite symmetric difference with L_i. Hence our identifier will always output the index of a language with finite symmetric difference with L_i after this point.

Part 2 (Necessity): First, we show necessity of Condition (1). Suppose for the sake of contradiction that Condition (1) were violated for language L but an identifier claims to approximately identify L with uniform bound $d^*(L)$. The adversary can select some $L' \subset L$ with $|L \setminus L'| = \infty$ such that $|L'| > d^*(L)$ and output $d^*(L)$ examples from $|L' \cap L|$. Then the identifier must output the index of a language that has finite symmetric difference from L in all future rounds. However, the adversary could then continue this sequence to an enumeration of L' and force the identifier to fail to approximately identify L'. Next, we show necessity of Condition (2). Suppose there existed languages $L, L' \in \mathcal{L}$ with infinite intersection and infinite symmetric difference. We can show the identifier $\max(d_L, d_{L'})$ examples from $|L \cap L'|$, after which it must output only languages with finite symmetric difference with both L and L' since this sequence can be extended to an enumeration of either. However, this is not possible by transitivity of the finite symmetric difference. □

We then present and prove the analogous characterization for uniform approximate identification:

Theorem A.8. *Language class $\mathcal{L}$ is uniformly approximately identifiable in the limit (with or without stability) if and only if there exists some $d \in \mathbb{N}$ such that both of the following hold:*

1. *For each language $L \in \mathcal{L}$, $|L'| < d$ for all $L' \subset L$ with $|L \setminus L'| = \infty$.*
2. *For all $L, L' \in \mathcal{L}$, either $|L \cap L'| < d$ or $|L \triangle L'| < \infty$.*

Proof. Again we make a modification to our non-uniform proof.

Part 1 (Sufficiency): From our non-uniform case, we obtain bounds $d^*(L_i) = \max(d, \max_{j<i, |L_j \cap L_i| < \infty, L_i \not\subset L_j} |L_j \cap L_i|)$ distinct examples. If the conditions in Theorem A.8

are satisfied, then $d^*(L_i) = \max(d, \max_{j<i, |L_j \cap L_i|<\infty, L_i \not\subset L_j} |L_j \cap L_i|) \leq \max(d, d)$, which is language independent and hence we get uniform identification.

Part 2 (Necessity): First, we show necessity of Condition (1). Suppose for the sake of contradiction that Condition (1) were violated but an identifier claims to approximately identify $\mathcal{L}$ with uniform bound d^*. The adversary can select some $L' \subset L$ with $|L \setminus L'| = \infty$ such that $|L'| > d^*$ and output d^* examples from $|L' \cap L|$. Then the identifier must output the index of a language that has finite symmetric difference from L in all future rounds. However, the adversary could then continue this sequence to an enumeration of L' and force the identifier to fail to approximately identify L'. Next, we show necessity of Condition (2). Suppose there existed languages $L, L' \in \mathcal{L}$ with intersection greater than d^* and infinite symmetric difference. We can show the identifier $d^* + 1$ examples from $|L \cap L'|$, after which it must output only languages with finite symmetric difference with both L and L' since this sequence can be extended to an enumeration of either. However, this is not possible by transitivity of the finite symmetric difference. □

Lastly, we show why the language class of one representative from each equivalence class can be non-uniformly identifiable, but the original language class may fail to be non-uniformly approximately identifiable. Consider the language class $\{\mathbb{N}\} \cup \{A \subset \mathbb{N} : A \text{ is finite}\}$. Then $\widetilde{\mathcal{L}} = \{\mathbb{N}, \{1\}\}$ is a language class consisting of one representative from each equivalence class. $\widetilde{\mathcal{L}}$ is clearly non-uniformly identifiable in the limit, but the original language class does not satisfy the conditions of Theorem 3.7 (or even those of Theorem A.7). Hence the structure within and between the entire equivalence class matters and we cannot just take one representative from each equivalence class.

A.5 Full Results and Proofs for k-list Identification in the Limit

In this section, we present full results and proofs for our results on k-list identification in the limit. We begin with a proof of our characterization for non-uniform k-list identification.

Proof of Theorem 3.5. Fix any ordering $\mathcal{L} = \{L_1, L_2, \dots\}$. We divide the proof into two parts.

Part 1 (Sufficiency): Suppose all intersections of $k+1$ languages are finite. Consider the following algorithm: the identifier in round t, after seeing S_t, outputs the k lowest-indexed languages consistent with S_t if $S_t \notin \mathcal{L}$, and otherwise outputs S_t followed by the $k-1$ lowest-indexes languages consistent with S_t.

Let L_ℓ be the target language. If L_ℓ is finite, then it is always included after seeing $|L_\ell|$ distinct examples. Otherwise, let $d = \max_{i_1,\dots,i_k<\ell} |L_{i_1} \cap \dots \cap L_{i_k} \cap L_\ell|$. We know $|L_{i_1} \cap \dots \cap L_{i_k} \cap L_\ell|$ is finite since it is an intersection of $k+1$ languages. After $d^*_{L_\ell} := d+1$ rounds, no set of k languages $L_{i_1}, \dots, L_{i_k}$ from the first $\ell - 1$ languages in $\mathcal{L}$ can be simultaneously consistent with L_ℓ, since we have seen more elements than the number of elements simultaneously consistent with $L_{i_1}, \dots, L_{i_k}, L_\ell$ (their intersection). Therefore, there are at most $k-1$ languages with indices less than ℓ that are consistent with S_t for all $t \geq d^*_{L_\ell}$. Furthermore, the size of finite languages still consistent with these k languages must have some bounded size B which doesn't depend on the enumeration (by condition (2), so we eliminate all finite languages after seeing at least B. Hence the first k languages consistent with S_t must contain L_ℓ for all $t \geq \max(d^*_{L_\ell}, B+1)$, showing that $\mathcal{L}$ is non-uniformly k-list identifiable in the limit.

Part 2 (Necessity): We first show necessity of Condition (1). Suppose languages $L_{i_1}, \dots, L_{i_{k+1}} \in \mathcal{L}$ have infinite intersection. Assume for the sake of contradiction that $\mathcal{L}$ is non-uniformly k-list identifiable in the limit, and let $d' = \max(d^*_{L_{i_1}}, \dots, d^*_{L_{i_{k+1}}})$. Consider a sequence of d' examples $S_{d'}$ such that $S_{d'} \subseteq L_{i_1} \cap \dots \cap L_{i_{k+1}}$. This can be continued to be an enumeration for all L_{i_ℓ} for $\ell = 1, \dots, k+1$. Hence, for all $t \geq d'$, we should have $L_\ell \in \mu_t$ for all $\ell = 1, \dots, k+1$ by the definition of non-uniform identifiability of each L_ℓ. However, this is not possible, since there are at most k languages in μ_t but there are $k+1$ possible target languages L_ℓ. This means if $\mathcal{L}$ does not have finite intersections of all sets of $k+1$ languages, then it is not non-uniformly identifiable in the limit.

Next, we show the necessity of Condition (2). Suppose there are k distinct languages $L_1, \ldots, L_k$ such that $\sup_{L \in L_1 \cap \cdots \cap L_k} |L| = \infty$. Suppose a k-list identifier claims to non-uniformly identify the collection. However, the adversary can present arbitrarily long enumerations of $L_1 \cap \cdots \cap L_k$ that are consistent with some other language $L' \in L_1 \cap \cdots \cap L_k$. When the length of this enumeration exceeds the claimed non-uniform bounds for $L_1, \ldots, L_k$, the identifier fails to non-uniformly identify the collection. □

Theorem A.9. *A countable language class $\mathcal{L}$ is uniformly k-list identifiable in the limit with or without stability if and only if there exists $d \in \mathbb{N}$ such that both of the following conditions hold:*

1. *Every intersection of $k+1$ distinct languages has size less than d.*
2. *For every k distinct languages $L_1, \ldots, L_k$, we have $\sup_{L \in \mathcal{L}: L \subseteq L_1 \cap \cdots \cap L_k, L \neq L_i} |L| < d$.*

Proof. We again extend the proof of Theorem 3.5 to the uniform case.

Part 1 (Sufficiency): In the non-uniform case, we get bounds $d^*(L_\ell) = \max(d^*_{L_\ell}, B+1)$. When the conditions in Theorem A.9 are satisfied, this becomes $d^*(L_\ell) = \max(d+1, d+1)$, which doesn't depend on the target language and so we get uniform k-list identification.

Part 2 (Necessity): Suppose an identifier claims to uniformly identify $\mathcal{L}$ with bound d^*. We first show necessity of Condition (1). Suppose languages $L_{i_1}, \ldots, L_{i_{k+1}} \in \mathcal{L}$ have intersection greater than d^*. Consider a sequence of d^*+1 examples $S_{d'}$ such that $S_{d'} \subseteq L_{i_1} \cap \cdots \cap L_{i_{k+1}}$. This can be continued to be an enumeration for all L_{i_ℓ} for $\ell = 1, \ldots, k+1$. Hence, for all $t \geq d'$, we should have $L_\ell \in \mu_t$ for all $\ell = 1, \ldots, k+1$ by the definition of non-uniform identifiability of each L_ℓ. However, this is not possible, since there are at most k languages in μ_t but there are $k+1$ possible target languages L_ℓ. This means if $\mathcal{L}$ does not have uniformly bounded intersections of all sets of $k+1$ languages, then it is not uniformly identifiable in the limit.

Next, we show the necessity of Condition (2). Suppose there are k distinct languages $L_1, \ldots, L_k$ such that $\sup_{L \in L_1 \cap \cdots \cap L_k} |L| \geq d^*$. However, the adversary can present d^*+1 examples from $L_1 \cap \cdots \cap L_k$ that are consistent with some other language $L' \in L_1 \cap \cdots \cap L_k$. Hence the identifier fails to uniformly identify the collection with bound d^*. □

Finally, we show a separation between stable and unstable non-uniform k-list identification.

Proof of Proposition 3.6. Let $\mathcal{L} = \{\mathbb{N}\} \cup \{\{1\}, \{1,2\}, \{1,2,3\}, \ldots\}$. We see that $\mathcal{L}$ satisfies the conditions from Theorem 3.5 with $k=2$, so is non-uniformly k-list identifiable without stability. However, suppose an identifier stably identified the first language $\mathbb{N}$ with bound d^*. Then an adversary could present examples $\{1, \ldots, d^*\}$, after which the list given by the identifier must not change. However, there are an infinite number of possible continuations of this set of examples, while the identifier only has a finite list. Hence no identifier can non-uniformly k-list identify $\mathcal{L}$ with stability. □