https://opennovelty.org/

Novelty Assessment Report

Paper: Automata Learning and Identification of the Support of Language Models

PDF URL: https://openreview.net/pdf?id=L8SMNWsxfK

Venue: ICLR 2026 Conference Submission

Year: 2026

Report Generated: 2025-12-27

Abstract

We study the learnability of languages in the Next Symbol Prediction (NSP) setting, where a learner receives only positive examples from a language together with, for every prefix, (i) whether the prefix itself is in the language and (ii) which next symbols can lead to an accepting string. This setting has been used in prior work to empirically analyze neural sequence models, and additionally, we observe that efficient algorithms for the NSP setting can be used to learn the (truncated) support of language models. We first show that the class of DFAs with at most N states is identifiable from positive examples augmented with these NSP labels. Nevertheless, even with this richer supervision, we show that PAC-learning DFAs remains computationally hard, and exact identification using only membership queries cannot be achieved in polynomial time. We then present $\mathrm{L_{nsp}^{\star}}$, an extension of Angluin's $\mathrm{L}^{\star}$ algorithm, and show that DFAs can be PAC-learned efficiently using a language-model–based teacher that answers membership queries and generates valid strings conditioned on prefix prompts. Finally, we conduct a comprehensive experimental evaluation on 11 regular languages of varying complexity. Using $\mathrm{L}^{\star}_{\text{nsp}}$, we extract DFAs from Transformer-based language models trained on regular languages to evaluate the algorithm's effectiveness and identify erroneous examples.

Core Task Landscape

This paper addresses: **Learning DFAs from Next Symbol Prediction Labels**

A total of **15 papers** were analyzed and organized into a taxonomy with **11 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Theoretical Foundations and Complexity Analysis**
- **Algorithmic Approaches for DFA Extraction**
- **Neural Network Analysis and Automata Extraction**
- **Application Domains and Specialized Extensions**

Complete Taxonomy Tree

- Learning DFAs from Next Symbol Prediction Labels Survey Taxonomy
- Theoretical Foundations and Complexity Analysis
  - PAC-Learnability and Identifiability Results ★ (3 papers)
  - [0] Automata Learning and Identification of the Support of Language Models (Anon et al., 2026) View paper
  - [2] Hardness of Learning Regular Languages in the Next Symbol Prediction Setting (Bhattamishra, 2025) View paper
  - [6] The Next Symbol Prediction Problem: PAC-learning and its relation to Language Models (S Bhattamishra, 2023) View paper
  - Average-Case and Probabilistic Learning Models (1 papers)
  - [14] Efficient learning of typical finite automata from random walks (Yoav Freund, 1993) View paper
- Algorithmic Approaches for DFA Extraction
  - Query-Based Learning Algorithms (1 papers)
  - [13] PDFA Distillation via String Probability Queries (Baumgartner, 2024) View paper
  - Passive Learning from Positive Examples (1 papers)
  - [9] Learning DFAs from sparse data (Guillaumier, 2020) View paper
  - Spectral and Algebraic Extraction Methods (2 papers)
  - [3] Distillation of weighted automata from recurrent neural networks using a spectral approach (R. Eyraud, 2024) View paper
  - [10] Extraction of rules from discrete-time recurrent neural networks (Christian W. Omlin, 1996) View paper
- Neural Network Analysis and Automata Extraction
  - Transformer Mechanism Analysis (1 papers)
  - [11] Mechanics of Next Token Prediction with Self-Attention (Li, 2024) View paper
  - World Model Evaluation and Implicit Structure Recovery (2 papers)
  - [1] Evaluating the world model implicit in a generative model (Justin Chen, 2024) View paper
  - [15] World Models: When Models Train Models (Mackensen, n.d.) View paper
  - Inference-Time Learning and Emergent Capabilities (2 papers)
  - [4] Causal language modeling can elicit search and reasoning capabilities on logic puzzles (Nishanth Dikkala, 2024) View paper
  - [12] Inference-Time Learning Algorithms of Language Models (Akyurek, 2025) View paper
- Application Domains and Specialized Extensions
  - Temporal Logic and Neurosymbolic Integration (1 papers)
  - [7] Neurosymbolic integration of linear temporal logic in non symbolic domains (Elena Umili, 2023) View paper
  - Complex Event Recognition and Forecasting (1 papers)
  - [8] Automata learning for complex event recognition and forecasting (Baou, 2025) View paper

- Comparative Sequence Modeling Studies (1 papers)
- [5] An interdisciplinary comparison of sequence modeling methods for next-element prediction (Tax, 2020) View paper

Narrative

Core task: learning DFAs from next symbol prediction labels. This field explores how to extract interpretable finite automata from models trained to predict the next symbol in a sequence, bridging neural learning and symbolic representation. The taxonomy organizes work into four main branches: Theoretical Foundations and Complexity Analysis examines the fundamental limits and sample complexity of learning automata from prediction oracles, including PAC-learnability results and hardness characterizations (e.g., Hardness Regular Languages[2], Next Symbol PAC[6]); Algorithmic Approaches for DFA Extraction develops concrete methods for constructing automata from trained predictors, often leveraging state merging or spectral techniques (e.g., Spectral Automata Distillation[3], DFAs Sparse Data[9]); Neural Network Analysis and Automata Extraction investigates how neural architectures encode regular structure and how to distill symbolic rules from them (e.g., Rule Extraction RNN[10], Next Token Mechanics[11]); and Application Domains and Specialized Extensions applies these ideas to tasks like temporal logic synthesis, complex event processing, and world model evaluation (e.g., Neurosymbolic Temporal Logic[7], Automata Complex Events[8], Evaluating World Model[1]).

A central tension runs through the field: while next-symbol prediction is a natural and widely used training signal, theoretical work reveals that learning exact DFAs from such labels can be computationally hard or require exponentially many queries in certain settings, as highlighted by Hardness Regular Languages[2]. Practical extraction algorithms must therefore balance expressiveness, sample efficiency, and computational tractability, often employing heuristics or restricting to subclasses of automata. Automata Learning Support[0] sits squarely within the Theoretical Foundations branch, focusing on PAC-learnability and identifiability results. Its emphasis on formal guarantees places it alongside Next Symbol PAC[6], which also investigates sample complexity bounds, though Automata Learning Support[0] may explore different oracle models or tighter characterizations. This contrasts with more algorithm-oriented neighbors like Spectral Automata Distillation[3], which prioritizes practical extraction methods over worst-case complexity, illustrating the field's dual commitment to rigorous theory and deployable techniques.

Related Works in Same Category

The following **2 sibling papers** share the same taxonomy leaf node with the original paper:

1. Hardness of Learning Regular Languages in the Next Symbol Prediction Setting

Authors: Bhattamishra, Satwik, Blunsom, Phil, Kanade, et al. (6 authors total) | **Year/Venue**: 2025 • arXiv.org | **URL**: View paper

Abstract

We study the learnability of languages in the Next Symbol Prediction (NSP) setting, where a learner receives only positive examples from a language together with, for every prefix, (i) whether the prefix itself is in the language and (ii) which next symbols can lead to an accepting string. This setting has been used in prior works to empirically analyze neural sequence models, and additionally, we observe that efficient algorithms for the NSP setting can be used to learn the (truncated) support ...

⚠ Similarity Notice

This paper appears to be a highly similar variant or earlier version of the original paper. Both papers share nearly identical titles (differing only in 'Automata Learning and Identification of the Support of Language Models' vs 'Hardness of Learning Regular Languages in the Next Symbol Prediction Setting'), describe the same NSP setting with identical formal definitions, and present the same core hardness results regarding PAC-learning DFAs from NSP labels. The abstracts and technical content overlap substantially, suggesting these are likely different versions of the same work rather than independent contributions.

---

2. The Next Symbol Prediction Problem: PAC-learning and its relation to Language Models

Authors: S Bhattamishra, P Blunsom, V Kanade | **Year/Venue**: 2023 | **URL**: View paper

Abstract

â¦ We show that Algorithm 2 finds the minimum consistent DFA for a given set of examples with the labels corresponding to the NSP setting. The algorithm works as follows: â¦

Relationship Analysis

Both papers belong to the PAC-Learnability and Identifiability Results category, analyzing whether DFA classes are efficiently learnable from NSP labels with polynomial complexity guarantees. They share overlapping focus on computational hardness results (both prove PAC-learning DFAs remains hard despite NSP labels) and identifiability from positive examples with NSP supervision. The key difference is that the original paper extends analysis to practical learning with language-model teachers via an L⋆-based algorithm and comprehensive empirical evaluation, while the candidate paper focuses primarily on theoretical PAC-learnability results and state-merging algorithms for general DFAs and acyclic DFAs without the language-model teacher framework or extensive experiments.

Contributions Analysis

Overall novelty summary. The paper contributes formal learnability results for DFAs in the Next Symbol Prediction setting, establishing identifiability from positive examples with NSP labels and proving computational hardness for PAC-learning. It resides in the 'PAC-Learnability and Identifiability Results' leaf alongside two sibling papers within the 'Theoretical Foundations and Complexity Analysis' branch. This leaf represents a focused but not overcrowded research direction, with only three papers addressing formal sample and time complexity guarantees for NSP-based DFA learning, suggesting the work enters a relatively sparse theoretical niche.

The taxonomy reveals neighboring research directions that contextualize this contribution. The sibling leaf 'Average-Case and Probabilistic Learning Models' explores distributional assumptions rather than worst-case analysis, while the broader 'Algorithmic Approaches for DFA Extraction' branch contains query-based and passive learning methods that implement extraction procedures. The paper bridges theoretical complexity analysis with algorithmic design through its L-star extension, connecting the 'Theoretical Foundations' branch to practical 'Query-Based Learning Algorithms'. The taxonomy's scope and exclude notes clarify that this work focuses on formal guarantees rather than empirical extraction heuristics or neural architecture analysis.

Among 25 candidates examined across three contributions, the identifiability result shows one refutable candidate from nine examined, indicating some prior overlap in this specific theoretical direction. The computational hardness results examined ten candidates with zero refutations, suggesting this contribution may occupy less-explored territory within NSP complexity theory. The L-star-nsp algorithm examined six candidates without refutation, though the limited search scope means undiscovered related query-based methods may exist. The statistics reflect a targeted literature search rather than exhaustive coverage, with contribution-level examination ranging from six to ten papers per claim.

Based on top-25 semantic matches plus citation expansion, the work appears to make substantive theoretical contributions, particularly in hardness characterization and algorithmic design for NSP settings. The identifiability result encounters some prior work overlap within the examined scope, while the hardness and algorithmic contributions show less direct precedent among candidates reviewed. The analysis covers formal learnability questions but does not exhaustively survey all automata extraction methods or neural analysis techniques in adjacent taxonomy branches.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: Identifiability of DFAs from positive NSP-labeled examples

Description: The authors prove that positive examples with Next Symbol Prediction labels are information-theoretically sufficient to uniquely identify minimal DFAs. They show that distinct minimal DFAs always disagree on NSP labeling of some positive string, establishing finite teaching sets and well-defined equivalence oracles.

This contribution was assessed against **9 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Learning fallible deterministic finite automata

URL: View paper

Brief Assessment

Fallible DFA Learning[16] addresses learning DFAs from a fallible expert with persistent errors, not identifiability from positive NSP-labeled examples. The candidate focuses on error correction rather than NSP supervision.

---

2. Quantifying the difference between black boxes and their automata approximations

URL: View paper

Brief Assessment

Black Box Approximations[17] focuses on measuring similarity between trained neural networks and extracted automata approximations, not on identifiability theory from labeled examples. The candidate addresses a different problem: quantifying differences between black-box models and their automata substitutes.

---

3. Extraction of rules from discrete-time recurrent neural networks

URL: View paper

Brief Assessment

Rule Extraction RNN[10] focuses on extracting DFAs from trained recurrent neural networks using clustering algorithms in the output space of recurrent state neurons, not on the theoretical identifiability of DFAs from positive examples with next symbol prediction labels. The candidate addresses a different problem domain (neural network rule extraction) rather than the information-theoretic identifiability question posed in the original contribution.

---

4. Efficient learning of typical finite automata from random walks

URL: View paper

Brief Assessment

Random Walk Automata[14] focuses on learning from random input sequences in an average-case setting with passive observation, not on identifiability from positive examples with NSP labels or teaching set existence.

---

5. An interdisciplinary comparison of sequence modeling methods for next-element prediction

URL: View paper

Brief Assessment

Sequence Modeling Comparison[5] focuses on next-element prediction using various sequence modeling methods (neural networks, Markov models, process mining techniques) on real-world datasets. It does not address the theoretical problem of identifiability of DFAs from positive examples with NSP labels or provide any formal learning-theoretic results about DFA identification.

---

6. Learning of context-sensitive language acceptors through regular inference and constraint induction

URL: View paper

Brief Assessment

Context Sensitive Learning[18] focuses on learning context-sensitive languages through regular inference and constraint induction using second-order RNNs for next-symbol prediction. The candidate does not address the theoretical identifiability properties of DFAs from positive examples with NSP labels that the original paper establishes.

---

7. Hardness of Learning Regular Languages in the Next Symbol Prediction Setting

URL: View paper

Brief Assessment

Hardness Regular Languages[2] focuses on computational hardness of learning DFAs in the NSP setting, not on identifiability from positive examples. The candidate establishes that learning remains hard despite richer labels, which is orthogonal to proving that distinct minimal DFAs can be uniquely identified.

---

8. The Next Symbol Prediction Problem: PAC-learning and its relation to Language Models

URL: View paper

Prior Art Analysis

Next Symbol PAC[6] demonstrates that positive examples with NSP labels are sufficient to uniquely identify DFAs, establishing that distinct minimal DFAs always disagree on NSP labeling of some positive string. The candidate paper proves the same identifiability result (Proposition 3.1) and explicitly shows that finite teaching sets exist and equivalence oracles are well-defined in the NSP setting. Both papers establish that NSP labels provide information-theoretic sufficiency for identifying minimal DFAs from positive examples alone, with nearly identical theoretical frameworks and proofs.

Evidence

Evidence 1 - **Rationale**: Both papers establish the same core identifiability result: that positive examples with NSP labels are information-theoretically sufficient to identify minimal DFAs. The original paper's claim of novelty is directly challenged by this prior work. - **Original**: we show that positive examples augmented with nsp labels are information-theoretically sufficient to identify minimal dfas. concretely, distinct minimal dfas always disagree on thensp labeling of some positive string, yielding finite teaching sets and a well-defined equivalence oracle in the nsp mod... - **Candidate**: we show that positive examples augmented with nsp labels are information-theoretically sufficient to uniquely identify minimal dfas. in the nsp setting, the learning algorithm is granted access to an example oracle, capable of supplying positively labeled strings. for every prefix of a given string,...

---

9. Learning DFAs from sparse data

URL: View paper

Brief Assessment

DFAs Sparse Data[9] is a PhD dissertation focused on learning DFAs from sparse data. The provided context only contains the title page and declaration, with no technical content about identifiability, NSP labels, or positive examples. Without access to the dissertation's technical chapters, comparison is impossible.

Contribution 2: Computational hardness results for NSP learning

Description: The authors demonstrate that despite richer NSP supervision, efficient PAC-learning of DFAs remains computationally intractable under cryptographic assumptions. They prove this through a reduction showing that NSP learning of acyclic DFAs is as hard as conventional binary classification learning.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Learning DFA from simple examples

URL: View paper

Brief Assessment

DFA Simple Examples[26] focuses on PAC-learning DFAs from simple examples under specific distributions (universal distribution), not on the computational hardness of NSP learning with membership queries as studied in the original paper.

2. Learning fallible deterministic finite automata

URL: View paper

Brief Assessment

Fallible DFA Learning[16] proves hardness of learning DFAs from a fallible expert under uniform distribution, not hardness of NSP learning with membership queries as in the original paper.

3. The query complexity of learning DFA

URL: View paper

Brief Assessment

DFA Query Complexity[23] focuses on query complexity bounds for learning DFAs with membership queries, not on computational hardness under cryptographic assumptions or the NSP setting with continuation labels.

4. Analyzing Robustness of Angluin's L* Algorithm in Presence of Noise

URL: View paper

Brief Assessment

Noisy Angluin Robustness[19] focuses on robustness of Angluin's L* algorithm under noise in membership queries, not on computational hardness of PAC-learning DFAs with membership queries or NSP supervision.

5. Learning Finite-State Machines

URL: View paper

Brief Assessment

Learning Finite State[24] focuses on PAC-learning probabilistic automata and DFAs using state-merging and spectral methods, not on proving computational hardness results for NSP learning specifically.

6. Learning stochastic finite automata

URL: View paper

Brief Assessment

Stochastic Automata Learning[25] addresses learning stochastic finite automata (SDFA), which are probabilistic models, whereas the original paper studies computational hardness of PAC-learning deterministic finite automata (DFAs) in the NSP setting. The candidate's focus on stochastic/probabilistic automata represents a fundamentally different learning problem from the deterministic setting with NSP supervision studied in the original work.

7. Recent advances of grammatical inference

URL: View paper

Brief Assessment

Grammatical Inference Advances[21] discusses general hardness results for DFA learning but does not address the specific NSP (Next Symbol Prediction) setting with continuation labels that the original paper introduces.

8. -Based Learning of Markov Decision Processes

URL: View paper

Brief Assessment

MDP Learning[20] focuses on learning Markov Decision Processes using membership queries, not on the computational hardness of PAC-learning DFAs with NSP supervision or cryptographic reductions for automata learning.

9. PAC Learning of Deterministic

URL: View paper

Brief Assessment

PAC Learning Deterministic[27] focuses on PAC learning of DFAs with membership queries in a standard setting, not the Next Symbol Prediction (NSP) framework with continuation labels that the original paper introduces and analyzes.

10. Learning a random DFA from uniform strings and state information

URL: View paper

Brief Assessment

Random DFA Learning[22] focuses on learning random DFAs from uniform strings with state information in a statistical query model, not on computational hardness of PAC-learning DFAs with membership queries or NSP supervision.

Contribution 3: L-star-nsp algorithm for learning with language model teachers

Description: The authors develop L-star-nsp, an extension of the classical L-star algorithm that uses membership and generative queries to extract DFAs from language models. The algorithm provides distribution-specific PAC guarantees with respect to the teacher's distribution and can identify the truncated support of language models.

This contribution was assessed against **6 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Learning minimal deterministic automata from inexperienced teachers

URL: View paper

Brief Assessment

Inexperienced Teachers Learning[32] focuses on learning from inexperienced teachers and uses weak versions of queries, which is a different learning setting than the original paper's L-star-nsp that uses membership and generative queries from language models.

2. An Active Learning Algorithm

URL: View paper

Brief Assessment

Active Learning Algorithm[33] presents an L-style algorithm for learning DFAs with a teacher providing membership queries, but focuses on standard L framework rather than the NSP (next symbol prediction) setting with language model teachers and generative queries that the original paper introduces.

3. Learning Deterministic Weighted Automata with Queries and Counterexamples

URL: View paper

Brief Assessment

Weighted Automata Queries[31] focuses on learning probabilistic DFAs (PDFAs) from black-box models using conditional probabilities and noise tolerance, whereas the original paper develops L-star-nsp specifically for the next symbol prediction (NSP) setting with membership and generative queries to extract DFAs (not PDFAs) from language models with distribution-specific PAC guarantees.

4. Automata Extraction from Transformers

URL: View paper

Brief Assessment

Automata Extraction Transformers[29] focuses on extracting DFAs from transformer models using the classical L algorithm with black-box tracking of latent representations. The original paper's L-star-nsp extends L specifically for the NSP setting with membership and generative queries, providing distribution-specific PAC guarantees—a fundamentally different learning framework and theoretical contribution.

5. : Learning Automata from Examples using Natural Language Oracles

URL: View paper

Brief Assessment

Natural Language Oracles[28] focuses on learning DFAs from natural language descriptions and demonstrations using LLMs as membership oracles, not on the next symbol prediction (NSP) setting with distribution-specific PAC guarantees that L-star-nsp addresses.

6. L*LM: Learning Automata from Demonstrations, Examples, and Natural Language

URL: View paper

Brief Assessment

Learning from Demonstrations[30] (L*LM) focuses on learning DFAs from natural language and demonstrations using membership queries to LLMs, while the original paper's L⋆-nsp uses membership and generative queries with next-symbol prediction labels from language model teachers. The query types and supervision signals differ fundamentally.

Appendix: Text Similarity Detection

Textual similarity detection checked 24 papers and found 3 similarity segment(s) across 1 paper(s).

The following **1 paper(s)** were detected to have high textual similarity with the original paper. These may represent different versions of the same work, duplicate submissions, or papers with substantial textual overlap. Readers are advised to verify these relationships independently.

1. Hardness of Learning Regular Languages in the Next Symbol Prediction Setting

Detected in: Core Task (sibling), Contribution: contribution_1

⚠ **Note**: This paper shows substantial textual similarity with the original paper. It may be a different version, a duplicate submission, or contain significant overlapping content. Please review carefully to determine the nature of the relationship.

References

- [0] Automata Learning and Identification of the Support of Language Models View paper
- [1] Evaluating the world model implicit in a generative model View paper
- [2] Hardness of Learning Regular Languages in the Next Symbol Prediction Setting View paper
- [3] Distillation of weighted automata from recurrent neural networks using a spectral approach View paper
- [4] Causal language modeling can elicit search and reasoning capabilities on logic puzzles View paper
- [5] An interdisciplinary comparison of sequence modeling methods for next-element prediction View paper
- [6] The Next Symbol Prediction Problem: PAC-learning and its relation to Language Models View paper
- [7] Neurosymbolic integration of linear temporal logic in non symbolic domains View paper
- [8] Automata learning for complex event recognition and forecasting View paper
- [9] Learning DFAs from sparse data View paper
- [10] Extraction of rules from discrete-time recurrent neural networks View paper
- [11] Mechanics of Next Token Prediction with Self-Attention View paper
- [12] Inference-Time Learning Algorithms of Language Models View paper
- [13] PDFA Distillation via String Probability Queries View paper
- [14] Efficient learning of typical finite automata from random walks View paper
- [15] World Models: When Models Train Models View paper
- [16] Learning fallible deterministic finite automata View paper
- [17] Quantifying the difference between black boxes and their automata approximations View paper
- [18] Learning of context-sensitive language acceptors through regular inference and constraint induction View paper

- [19] Analyzing Robustness of Angluin's L* Algorithm in Presence of Noise View paper
- [20] -Based Learning of Markov Decision Processes View paper
- [21] Recent advances of grammatical inference View paper
- [22] Learning a random DFA from uniform strings and state information View paper
- [23] The query complexity of learning DFA View paper
- [24] Learning Finite-State Machines View paper
- [25] Learning stochastic finite automata View paper
- [26] Learning DFA from simple examples View paper
- [27] PAC Learning of Deterministic View paper
- [28] : Learning Automata from Examples using Natural Language Oracles View paper
- [29] Automata Extraction from Transformers View paper
- [30] L*LM: Learning Automata from Demonstrations, Examples, and Natural Language View paper
- [31] Learning Deterministic Weighted Automata with Queries and Counterexamples View paper
- [32] Learning minimal deterministic automata from inexperienced teachers View paper
- [33] An Active Learning Algorithm View paper