IMPLICIT LANGUAGE REPRESENTATION OF LLMS: A FORMAL LANGUAGE LEARNING PERSPECTIVE

Anonymous authors
Paper under double-blind review

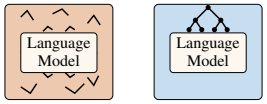
ABSTRACT

When a large language model (LLM) is trained to learn a language, is it learning the underlying language representation (henceforth, also *structure*) precisely, or is it also becoming proficient in other related languages? This question offers insight into the LLM's implicit world model. To answer this question precisely, we need a learning environment for LLMs where we can define the language and its underlying structure. We propose *probabilistic formal language* as the learning environment, which provides greater control than natural language. Next, we need a proficiency test that considers the possibility of generating new strings both inside and outside the language. Our answer is a *local discriminative test* that operates on individual strings: a string is learned if the LLM generates it with higher probability than every neighboring out-of-language string, and language proficiency is the fraction of strings learned this way.

Empirically, within a language, an LLM learns strings better if they are structurally closer to the training data. Beyond individual strings, we characterize the LLM's implicit structure, i.e., its formal grammar here, by *the languages it generalizes to beyond the target*. We find that the LLM generalizes better to out-of-distribution languages that relax global structure than to ones that relax local structure. Simultaneously, the implicit grammar is close to grammar rules governing local structure, but deviates from rules governing global structure, a pattern we also confirm empirically in natural language, where it learns local word structure better than global sentence- or paragraph-level structure.

1 INTRODUCTION

"An ant, viewed as a behaving system, is quite simple. The apparent complexity of its behavior over time is largely a reflection of the complexity of the environment in which it finds itself." – Herbert Simon in The Sciences of the Artificial, 1996



Large language models (LLMs) are probabilistic generative models that can generate fluent, natural-sounding text (Brown et al., 2020). Yet, it is hard to precisely determine *what implicit representation or structure of a language an LLM learns* (Bender et al., 2021; Hewitt & Manning, 2019; Vafa et al., 2024). Ideally, we want to study the question in natural language, which lacks a well-defined structure (Shieber, 1985). We therefore take an orthogonal perspective: we teach a *probabilistic formal language* to an LLM, where the language structure is defined by a *probabilistic formal grammar* (Chomsky, 1956; Booth, 1969). The setup also lets us isolate the LLM's complexity from that of its environment, the formal language, following Simon's parable (Simon, 1996).

Implicit Language Representation via Out-of-Distribution Generalization. Our goal is to study the implicit language structure of an LLM when it learns a language from sampled strings. Formally, given a language L represented by grammar G, does the LLM's implicit grammar match G exactly, and if not, how does it differ? Fully specifying the LLM's implicit world model is extremely challenging; instead, we study some of its characteristics via what other related languages it also learns. To do so, we study hierarchical probabilistic context-free grammars (HPCFG), which are inspired by the recursive structure of natural language (Allen-Zhu & Li, 2023; Ghosh et al., 2026). Such grammars present a hierarchy of grammar rules, where low-level rules govern how nearby tokens combine, capturing *local* structure, and high-level rules govern how larger substrings

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.

combine, capturing *global* structure. We therefore have an opportunity to *relax* the language's structure (i.e., inserting, deleting, or replacing a symbol in a grammar rule) and study how well the LLM learns the resulting out-of-distribution languages, using the proficiency test defined next. This out-of-distribution generalization is our key insight for characterizing an LLM's implicit grammar.

In Section 5, we find that the LLM generalizes better to out-of-distribution languages that relax global structure than to ones that relax local structure. Simultaneously, the LLM's implicit grammar is close to G on rules governing local structure, but deviates on rules governing global structure. This asymmetry also holds empirically in natural language, where an LLM learns local word-level structure better than global sentence- or paragraph-level structure.

Language Proficiency via a Local Discriminative Test. To study the implicit grammar, we revisit a more fundamental question: *how can we measure an LLM's language proficiency*? A generative test considers only strings inside the language, measuring proficiency by generation probability – higher probability means higher proficiency (Allen-Zhu & Li, 2023; Suzgun et al., 2019a;b; Bhattamishra et al., 2020; Delétang et al., 2023). However, an LLM assigns some probability to every string, including those *outside* the language, making the generative test less informative. A global discriminative test instead checks whether strings inside the language receive higher generation probability than strings outside it (Ghosh et al., 2026). But the global test is also insufficient, since different strings are learned unevenly. Our answer is therefore a *local discriminative test*: a string is learned if the LLM generates it with higher probability than every neighboring out-of-language string (obtained by inserting, deleting, or replacing a token), and language proficiency is the fraction of strings learned this way. Applying this test, we show that strings structurally closer to the training data are learned faster and remain learned longer (Section 4). Importantly, local discrimination does not always correlate with generation probability, i.e, a string generated with higher probability is not necessarily learned or locally discriminated.

Our results open up new avenues for studying an LLM's implicit grammar through *succinctness* (how compactly it represents a language) and *creativity* (its capacity to generate beyond the target while staying faithful to local structure). Finally, we hope that our work motivates the community to adopt probabilistic formal languages as an environment simulator for studying LLMs, which themselves are probabilistic generators.

2 Motivation and Related Work

Implicit World Models of LLMs. A growing body of work asks whether the next-token generation by an LLM induces an internal world model of the sequence-generating process. Most works consider deterministic ground-truth world models, such as Othello and chess (Li et al., 2023; Nanda et al., 2023; Toshniwal et al., 2022) or real-world facts (Gurnee & Tegmark, 2024), via internal probing, finding that transformers trained only on move sequences learn to encode such state internally. Closer to our question, Hewitt & Manning (2019); Tenney et al. (2019) probe whether models encode syntactic structure specifically. Vafa et al. (2024) instead evaluate against deterministic finite automata, finding recovered world models far less coherent than their surface behavior suggests. A methodological difference between these works and ours is the following: we infer the characterization of the LLM's implicit grammar by studying what additional languages the LLM is able to learn, in direct contrast to the binary question of whether the implicit grammar matches the target exactly.

Formal Language and LLMs. Formal language offers a controlled setting for studying the representational expressiveness of sequence models: what classes of languages a given architecture can represent or recognize at all. This question has been studied for non-transformer architectures such as RNNs and LSTMs (Suzgun et al., 2019a), and for transformers (Bhattamishra et al., 2020; Delétang et al., 2023). Another line of work asks what computation model explains transformers' inner workings: Weiss et al. (2021) propose RASP, a restricted programming language that captures what transformer encoders can compute. Across this work, models are typically small, single-task, and not decoder-only, unlike the widely used LLMs we study. Allen-Zhu & Li (2023) instead train large transformers on an HPCFG designed to mimic natural language, motivating our own use of such a grammar (Section 3). A separate computation-free line of work asks a theoretical question involving formal languages: does a learner identify a language in the limit (Gold, 1967)? Kleinberg & Mullainathan (2024) recently revisit this question as language generation in the limit. Our paper instead studies an actual, already-trained LLM, not an idealized learner unconstrained by architecture.

Measuring Language Proficiency. Prior tests of language proficiency are mostly generative, focusing on grammatically correct strings accepted in the language (Allen-Zhu & Li, 2023; Suzgun et al., 2019a;b; Bhattamishra et al., 2020; Delétang et al., 2023). These works essentially compare the set of accepted strings in the language with the set of strings generated by the LLM. An LLM, however, generates any string with some probability, yielding a critical mismatch between comparing a set and a distribution, as rightly pointed out by Borenstein et al. (2024). Therefore, Borenstein et al. (2024); Allen-Zhu & Li (2023) propose comparing the language distribution with the LLM distribution using the Kullback-Leibler (KL) divergence. We, however, show that KL divergence carries no signal from grammatically incorrect strings. Closest to our work, Ghosh et al. (2026) propose a discriminative test based on relative probability between grammatically correct and incorrect strings. Their test is global: treating all strings as one classification problem. However, strings are learned unevenly, which requires a local per-string discriminative test (Section 4).

We extend related work in Appendix A.

3 Research Questions and Setup

Our goal is to study the implicit language structure of an LLM. To this end, we propose a measure for language proficiency, which we use to study this structure. Our research questions (RQs) are:

- **RQ 1:** How can we measure the proficiency of an LLM in a language?
- **RQ 2:** When an LLM learns a language, what can we infer about its implicit language structure?

We answer **RQ 1** and **RQ 2** in Section 4 and Section 5, respectively. In this section, we first discuss the preliminaries for our experimental setup, then explain why formal language is an effective environment simulator for studying an LLM's learning.

3.1 Setup and Preliminaries

Large Language Models (LLMs). We study an LLM through the generation probability it assigns to strings, where a string is a sequence of tokens. We consider auto-regressive LLMs, which generate each token in a string conditioned on all previous tokens, called the prefix. Let $\boldsymbol{V}$ be the LLM's token vocabulary and $\boldsymbol{V}^*$ be the universe of all finite strings over $\boldsymbol{V}$. An LLM with parameters θ induces a distribution $P_\theta : \boldsymbol{V}^* \to [0, 1]$ over $\boldsymbol{V}^*$, factorized auto-regressively as $P_\theta(s) = \prod_{i=1}^{|s|} P_\theta(s^{(i)} \mid s^{(<i)})$. Here, $s^{(<i)} = s^{(1)} \cdots s^{(i-1)}$ is the prefix for token $s^{(i)}$. *Most fundamentally, the LLM can generate any string in the universe $\boldsymbol{V}^*$ with some (possibly vanishing) probability.* In practice, we compute per-token cross-entropy loss as $\texttt{loss}_\theta(s) = -\frac{1}{|s|} \log P_\theta(s)$, which is inversely related to the probability and comparable across strings of different lengths.

$S \to A18\ A16$ [0.50]
$S \to A16\ A18\ A17$ [0.50]
} Level 4

$A18 \to A15\ A14\ A13$ [0.50]
$A18 \to A14\ A15\ A13$ [0.50]
$A17 \to A14\ A13\ A15$ [0.50]
$A17 \to A13\ A14\ A15$ [0.50]
$A16 \to A14\ A15$ [0.50]
$A16 \to A15\ A14$ [0.50]
} Level 3

$A15 \to A11\ A12\ A10$ [0.50]
$A15 \to A12\ A10\ A11$ [0.50]
$A14 \to A11\ A10\ A12$ [0.50]
$A14 \to A10\ A11\ A12$ [0.50]
$A13 \to A10\ A12\ A11$ [0.50]
$A13 \to A12\ A11\ A10$ [0.50]
} Level 2

$A12 \to 9\ 8\ 7$ [0.50]
$A12 \to 8\ 7$ [0.50]
$A11 \to 6\ 5$ [0.50]
$A11 \to 6\ 4\ 5$ [0.50]
$A10 \to 3\ 1$ [0.50]
$A10 \to 1\ 2\ 3$ [0.50]
} Level 1

Figure 1: An HPCFG.

Probabilistic Formal Language. Formal language provides the learning environment for LLMs. Let $\boldsymbol{\Sigma} \subseteq \boldsymbol{V}$ be the vocabulary of the environment. A *probabilistic formal language* is a pair (L, P_L), where $L \subseteq \boldsymbol{\Sigma}^*$ is the set of *grammatically correct* strings and P_L is a probability distribution over L, with $\sum_{s \in L} P_L(s) = 1$. We extend P_L to a distribution over the universe $\boldsymbol{V}^*$ by setting $P_L(s) = 0$ for every *grammatically incorrect* string $s \notin L$, including strings using tokens outside $\boldsymbol{\Sigma}$. How well the LLM generates strings from L relative to strings outside L is the key desideratum of **RQ 1**.

Probabilistic Formal Grammar. A formal grammar defines the structure of a formal language. In our setup, the LLM never sees the grammar, only strings sampled from it; yet we ask what implicit grammar the LLM learns (**RQ 2**). We consider hierarchical probabilistic context-free grammars (HPCFGs), formally defined as $G = (\boldsymbol{N}, \boldsymbol{\Sigma}, \boldsymbol{R}, S, \pi)$. $\boldsymbol{N}$ is a finite set of non-terminals, $\boldsymbol{R}$ is a

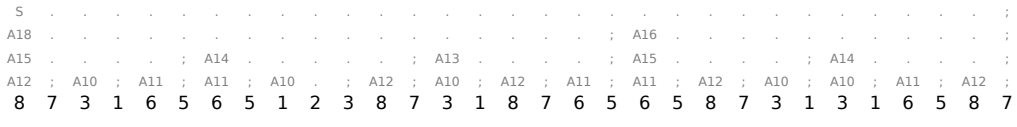


Figure 2: A string from the grammar in Figure 1, with probability $(0.5)^{23}$ and its derivation tree.

finite set of production rules of the form $A \to \beta$ with $A \in \boldsymbol{N}$ and $\beta \in (\boldsymbol{N} \cup \boldsymbol{\Sigma})^*$, i.e., β is a (possibly empty) sequence of zero or more terminals or non-terminals, $S \in \boldsymbol{N}$ is the start non-terminal, and π assigns a probability to each rule such that $\sum_{\beta\,:\,(A\to\beta)\in\boldsymbol{R}} \pi(A \to \beta) = 1$ for every non-terminal $A \in \boldsymbol{N}$. In this work, we are mostly concerned with a grammar rule $A \to \beta$, which denotes how a non-terminal A is expanded into a (sub)string with terminals or non-terminals in β while deriving the complete string. In fact, the language generation probability of a string $P_L(s)$ is the sum of products of rule probabilities over all derivations in the grammar G.

An Illustration. Figure 1 shows an HPCFG in which each non-terminal expands to a distinct set of non-terminals at the level below; Figure 2 shows a sampled string with its derivation tree. Rules at different levels impose different structure on the resulting strings. A lower level rule specifies *local structure*, e.g., $A11 \to 6\ 4\ 5$ at level 1 specifies how tokens $\{4, 5, 6\}$ appear together in a string. The non-terminal $A11$ on the LHS of the rule carries no meaning beyond this expansion; it is only referenced by a higher-level rule, e.g., $A15 \to A11\ A12\ A10$ at level 2. In contrast, a higher level rule specifies *global structure* in a string, e.g., $S \to A18\ A16$ at level 4 specifies that the tokens expanded by $A18$ precede those expanded by $A16$. Analyzing rule-proficiency thus leads us to the implicit world model of an LLM, which we study in **RQ 2**.

Construction of Grammatically Incorrect Strings. Despite attempting to learn L, an LLM may generate strings outside L as well. To examine this, we construct grammatically incorrect strings in two ways. **String edits:** We edit a string $s \in L$ to $s_{\text{incorrect}}$ by randomly inserting, deleting, or replacing tokens from $\boldsymbol{\Sigma}$ in at most k positions, e.g., $k \in \{1, 2, 3\}$. **Rule edits:** Given a production rule $R \triangleq A \to \beta$ and a correct string $s \in L$ that uses the rule R in its derivations, we edit R by randomly inserting, deleting, or replacing terminals or nonterminals in the right-hand side β, up to k times, to produce a new rule $R_{\text{incorrect}} \triangleq A \to \beta_{\text{incorrect}}$ such that $R_{\text{incorrect}} \notin \boldsymbol{R}$. Then, we substitute $R_{\text{incorrect}}$ for R to derive a new string $s_{\text{incorrect}}$. In both cases, each edited string $s_{\text{incorrect}}$ is later verified to be a non-member, i.e., $s_{\text{incorrect}} \notin L$, using a grammar parser. String-level edits are used to measure language proficiency (**RQ 1**) while rule-level edits are used to investigate the underlying language structure learned by the LLM (**RQ 2**). Unless otherwise stated, we use $k = 1$.

3.2 Advantages of Formal Language over Natural Language

To precisely answer our research questions, the learning environment should satisfy the following properties, which formal language provides, but natural language does not.

- **Precise Specification of a Language.** We can precisely determine whether a string belongs to the formal language or not, which is generally hard in natural language. Such precision is essential to measure language proficiency and to determine what other languages the LLM learns. In a formal language, grammatical strings carry known structure (their derivation trees), letting us fully characterize their structural properties. Strings that do not belong to the language likewise admit no valid derivation, letting us pinpoint exactly where and how they diverge from the grammar.
- **Exact Sampling.** Exact sampling of training and test strings from the same distribution is only an assumption in natural language, but guaranteed in formal language. Unless sampling is exact and i.i.d., any statistical analysis of the LLM's performance is subject to bias.
- **Elimination of Semantic Ambiguity and Data Contamination.** A formal language contains only syntax, eliminating the semantic ambiguity inherent to natural language. In a formal language, the generation of the next token given a prefix is entirely due to syntactic pattern recognition, independent of semantic interpretation, allowing us to be precise about what the LLM learns. Furthermore, formal language is synthetic, meaning that it is unlikely to be seen by pre-trained LLMs. Evaluations of LLMs on publicly available natural language are often contaminated, making them unreliable for a fair assessment of language proficiency (Xu et al., 2024).

- **Effective Control over Grammar Structure.** Knowing the underlying grammar allows us to control the complexity of the language and study the relationship between complexity and proficiency. Pinpointing the complexity in any natural language dataset is difficult.

Simultaneously satisfying all of the above advantages is hard to achieve in natural language, which is why *formal language serves as an effective learning environment for LLMs*. Nevertheless, we demonstrate how our observations on formal language extend to natural language.

4 Language Proficiency of an LLM

The language proficiency of an LLM cannot solely be defined by the generation loss of strings inside the language, as predominantly done in many *generative* language proficiency tests: lower the generation loss of strings inside the language, the more proficient the LLM is (Appendix B). An LLM can also generate strings outside the language, motivating a recent *global discriminative* test (Ghosh et al., 2026). Global discrimination is the LLM's ability to generate strings inside the language with lower loss than grammatically incorrect ones outside it, turning language proficiency into a grammatically correct vs. incorrect classification task. However, *individual* strings in the language are not equally learnable, and a global classification test does not answer when an individual string is learned, making the need for a *local discriminative test*. In the following, we start with defining when a string is learned by an LLM, followed by a quantitative measure of language proficiency. We end with applying the test on a training dynamic.

4.1 Local Discriminative Test

String Proficiency. We define a string as learned if an LLM not only generates it, but also detects and corrects any small edits that would turn it into a grammatically incorrect string. We emphasize small edits because they are harder to detect than obvious and large ones. Thus, *a string is learned if the LLM generates it with lower loss than neighboring grammatically incorrect strings*. Formally, given a string s, let $\mathcal{N}_k(s)$ be the set of grammatically incorrect strings within edit distance k of s. We define the local boundary of s as the lowest loss any string in $\mathcal{N}_k(s)$ ever achieves over the course of training,

$$\texttt{boundary-loss}_k(s) \triangleq \min_{s_{\text{incorrect}} \in \mathcal{N}_k(s)} \min_{\theta} \texttt{loss}_\theta(s_{\text{incorrect}}).$$

A string is learned (at a given epoch) if it is a *local minimum*: its current loss is below the boundary.

$$\texttt{loss}_\theta(s) < \texttt{boundary-loss}_k(s)$$

Once learned, a string is *forgotten* if its loss eventually reaches higher than the boundary. Also, the local-minimum condition makes the informal notion of *correction* precise. Suppose a grammatically correct and *learned* string lies within k edits of an incorrect string $s_{\text{incorrect}}$; editing $s_{\text{incorrect}}$ to reach s strictly reduces loss. If string s is unique within k edits of $s_{\text{incorrect}}$, we reach the local minimum unambiguously. If several correct and learned strings exist, we can still correct to one of them.

The discriminative test is also agnostic to the intrinsic complexity of generating s, i.e., its absolute loss. If a string is structured and obtains lower loss, so do its neighboring incorrect strings, and the string is learned only if it is a local minimum. In contrast, a generative test would naively call a lower loss a better learned string, even if it is not a local minimum.

Language Proficiency. A language is learned if its constituent strings are learned. Different strings, however, can be learned at different epochs during training. Therefore, we define language proficiency as the expected fraction of strings from L that are learned,

$$\mathbb{E}_{s \sim P_L}\left[\mathbb{1}_\theta[s \text{ is learned}]\right],$$

which we operationalize on D_{test} as $\frac{1}{|D_{\text{test}}|} \sum_{s \in D_{\text{test}}} \mathbb{1}_\theta[s \text{ is learned}]$.

In Appendix B, we elaborately compare our local discriminative test with (i) a generative test, (ii) a distributional-difference-based test, and (iii) a global discriminative test. Unlike ours, (i) and (ii) consider strings inside the language, ignoring LLMs' generation ability outside the language. Moreover, (ii) and (iii) cannot answer when an *individual* string is learned, due to their global nature; our local discriminative test can.

4.2 Results on Language Proficiency

We apply the local discriminative test in the training dynamic to measure string and language proficiency. The results below are for the Qwen-2.5-1.5B model, which is trained on D_{train} and evaluated on D_{test}; both are from the formal language L in Figure 1. We additionally evaluate the generation loss on a set of grammatically incorrect strings $D_{\text{incorrect-test}}$ at increasing edit distances from D_{test}. [1]

LLMs learn to generate not only grammatically correct strings, but also nearby grammatically incorrect strings. In Figure 3, training loss decreases with repeated training, while test loss of grammatically correct strings decreases to a minimum value before rising again due to over-fitting. In the beginning, both grammatically correct and incorrect strings are generated with closer loss and cannot be distinguished by the LLM. Later, the loss gap between correct and incorrect strings widens over training, and is larger for incorrect strings at greater edit distance; random strings incur the largest gap. *Therefore, the LLM progressively becomes better at discriminating correct strings from incorrect ones: the higher the edit distance, the easier the discrimination*, motivating the global discriminative test (Ghosh et al., 2026). The figure, however, cannot determine when individual test strings are learned, which requires string-specific analysis.

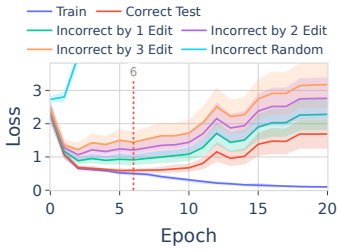
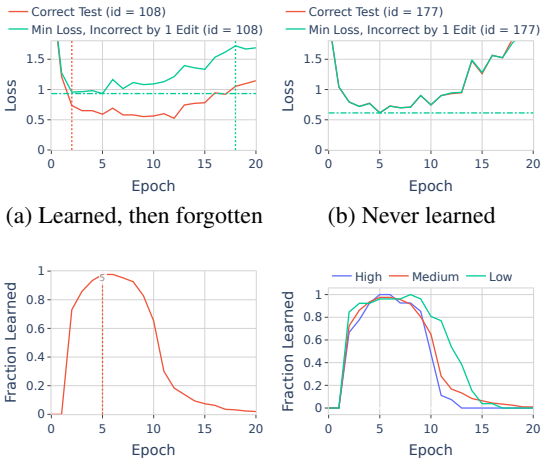


Figure 3: Generation loss of different strings. Average test loss is the lowest at epoch 6.

Not all strings are learned equally. Structurally closer strings to the training set are learned faster and forgotten later. We illustrate the loss dynamics for two representative strings, one that the LLM learns and later forgets (Figure 4a), and another it never learns (Figure 4b). Since local boundary is a string-adaptive threshold for learning, we observe such learning inequality across strings. We also report the fraction of learned strings in the language over epochs (Figure 4c), which rises sharply, peaks around epoch 5 (optimal learning by discrimination), and then declines as training continues, mirroring the overfitting pattern in Figure 3. To further analyze the learning inequality, we compute the average tree-edit distance between a test string and all training strings, using their derivation trees in the grammar. In Figure 4d, test strings structurally closer to the training set (i.e., with lower average tree-edit distance) are learned earlier and remain learned longer, while structurally distant strings are learned later and forgotten sooner. This effect is more pronounced under a skewed language distribution than a uniform one (Figures 22 and 23, Appendix). *Therefore, learning is unequal across strings: those structurally closer to the training set are learned better.*



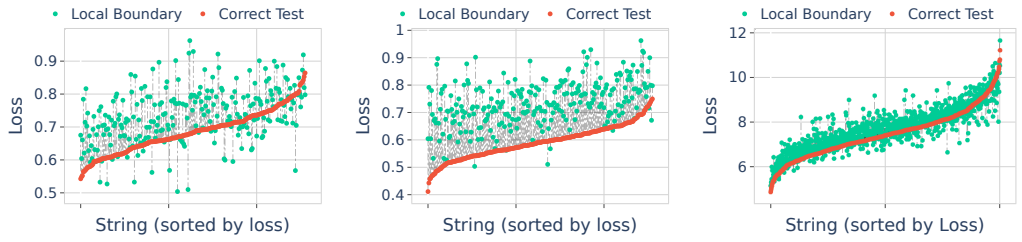
(a) Learned, then forgotten (b) Never learned

(c) Language proficiency (d) Split by tree-edit distance

Figure 4: String and language proficiency. In Figure 4a, vertical lines mark when the string is learned and eventually forgotten, determined by comparing its loss to the local boundary (shown as a horizontal line). Figure 4b shows a string that is never learned. Figure 4c shows proficiency in L, and Figure 4d further breaks down proficiency by the average tree edit distance of a test string to all training strings.

Strings that are easier to generate are not necessarily discriminated well. Today, the perception is that a string with lower loss is learned better (Delétang et al., 2023). In Figure 5, we, however, show that having lower loss does not necessarily mean that the string has lower loss than its local boundary, i.e., the LLM can fix any small grammatical mistakes within the string. Both formal language and natural language (Warstadt et al., 2020) support this claim. *Therefore, our local dis-*

[1] We provide details on experiments in Appendix C. The main paper contains results for a single language. The results for multiple languages are in the Appendix D.



(a) Formal language (optimal epoch / 2)

(b) Formal language (optimal epoch)

(c) English language (inference on pre-trained LLM)

Figure 5: Loss of a string and its local boundary, connected by a gray vertical line. Lower generation loss of a string does not guarantee that the loss is lower than its local boundary.

criminative test asks something more informative: generating a string well is not sufficient for it to be learned; it must be generated well enough that the LLM can also correct any small grammatical mistakes within it.

Takeaway. Local discrimination asks: can an LLM generate an individual string in the language better than nearby grammatically incorrect strings? We find that strings are not learned equally: structurally closer strings to the training data are learned faster and forgotten later. Generation loss alone cannot decide whether a string succeeds or fails in the local discrimination test. Language proficiency increases with training epochs, but decreases due to over-fitting. These results raise a deeper question: if an LLM has proficiency in language L, how does the implicit grammar of the LLM compare with the underlying grammar G? We investigate this question in the next section.

5 Implicit World Model of an LLM

Fully specifying the implicit world model of an LLM is a challenging problem. Here, we instead infer some of its characteristics by observing what other languages beyond language L the LLM is capable of learning. If the LLM's implicit grammar is faithful to G, any out-of-distribution (OOD) grammars that are equally different (by some well-defined measure) from G, should be equally learnable. To the best of our knowledge, this OOD perspective, which we introduce, is absent from prior work on implicit world models (Section 2). We discuss our results below.

The LLM generalizes better to out-of-distribution hierarchical languages that relax global structure than to ones that relax local structure. We start with a hierarchical grammar G and make one edit to different rules: editing a low-level rule, which captures local structure, yields a *local relaxation of the grammar*; editing a high-level rule, which captures global structure, yields a *global relaxation*. To measure language proficiency in the resulting OOD grammar G_{OOD}, we apply the local discriminative test to individual strings in G_{OOD} against their respective k-neighborhoods (Section 4). Figure 6 shows a striking pattern: initially all OOD languages are learned almost equally, but language proficiency is only retained when edits are made to higher-level rules. *Therefore, the LLM is not only learning language L but also learning nearby languages that relax global structure.* Since OOD generalization is unequal in specific ways, the implicit grammar of the LLM is not exactly aligned with G, as investigated next.

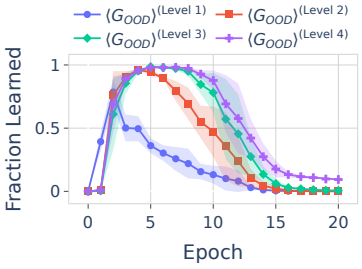
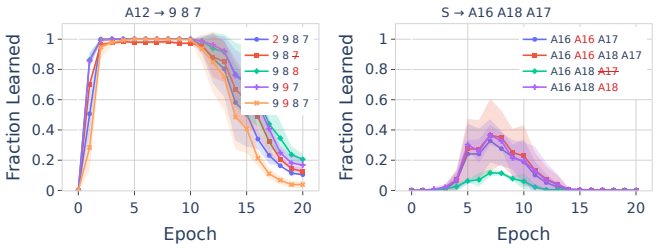


Figure 6: Language proficiency across out-of-distribution languages induced by $\{G_{\text{OOD}}\}^{(\text{Level } \ell)}$, the family of grammars each editing one rule of G at level ℓ. Proficiency is better when editing higher-level rules, i.e., relaxing the global structure of strings in G. Variance is w.r.t. three seeds of the training dataset.

Within a language, the LLM learns local structure better than global structure. In Figure 7, we demonstrate the proficiency of a rule against an edit. This adapts the local discriminative test from Section 4 to individual rules: a string that uses the rule in its derivation is learned if it is a local minimum against all neighboring strings derived using the edited rule. At level

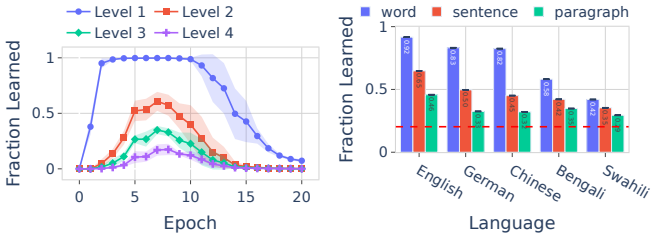
1, proficiency reaches its maximum against all edited rules we have tested. At level 4, proficiency drops: for example, the LLM cannot discriminate the rule $S \rightarrow A16\ A18\ A17$ from an edited rule $S \rightarrow A16\ A18$, reaching a maximum proficiency of 0.12. Thus, the strings generated by either rule are nearly indistinguishable to the LLM, as if the LLM has learned both rules equally. *Therefore, the LLM learns local structure better than global structure within a language.*

This local-global asymmetry generalizes across languages. Figure 8a aggregates proficiency across rules by levels: proficiency is high at lower levels and low at higher levels. The same asymmetry appears in natural language (Figure 8b), where, across five languages (English, German, Chinese, Bengali, and Swahili), discrimination declines monotonically as edits move from words to sentences to paragraphs, essentially moving from local to global. *Therefore, in both formal and natural language, the LLM learns local structure better than global structure.*



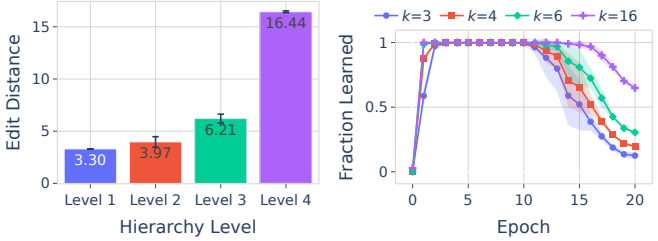
(a) Level 1 (local structure) (b) Level 4 (global structure)

Figure 7: Language proficiency when we edit a rule (particularly, the RHS symbols) and attempt to discriminate strings generated using the edited rule. Red color denotes the edit, where a terminal or non-terminal in the RHS is edited within the same level.



(a) Formal language (b) Natural language

Figure 8: Average proficiency across rules by level. Lower level rules in the formal grammar (8a) are analogous to local structure such as word structure in natural language; higher level rules are analogous to global structure such as sentence- or paragraph-level structure (8b).



(a) Edits on resulting strings (b) Edits at random positions

Figure 9: (a) Number of edits on resulting strings when equally editing rules at different levels in Figure 8a. (b) When a matching number of edits is applied to strings at random positions instead, the fraction of learned strings increases with edit distance, the reverse of the trend in Figure 8.

Understanding the connection between string and rule edits. An edit to a rule rewrites its entire derived substring. Although we perform an equal edit across rules, global rule edits result in larger string edits than local rule edits, as shown in Figure 9a. In general, we would expect that higher string edits lead to better discrimination (Figure 9b). However, this is not always the case. Higher string edits that are a consequence of global rule edits lead to less discrimination, as we consistently observe in Figure 8a. One explanation is that local structure is already learned well by the LLM, and global edits recombine already learned local structure. Such coherence at a global level makes discrimination less likely to succeed, resulting in out-of-distribution generalization that relaxes global structure. Conversely, relaxing local structure hurts out-of-distribution generalization, since the LLM already learns local structure well.

Takeaway: The results above show that within a language, local structures captured using low-level rules are easily discriminated, while global structures captured using high-level rules are difficult to discriminate. Thus, a high-level rule is difficult to distinguish from its edited (nearby) counterpart. Therefore, the LLM

performs better on out-of-distribution languages where global structure is relaxed, but local structure is preserved. This suggests that the LLM's implicit world model is locally close to G, but globally relaxed relative to it.

6 IMPLICATIONS

Our results have the following implications.

Rethinking the Implicit World Model of an LLM. Our key insight into the world model question is that when an LLM learns a language, it is capable of learning other related languages as well. Therefore, validating the implicit world model of an LLM with a *fixed deterministic target*, as predominantly performed before (Section 2), is too restrictive and may not capture the LLM's inductive bias. Our discriminative test reveals the implicit world model through the LLM's uneven discrimination: the LLM discriminates less well on global structure, and correspondingly generalizes better to out-of-distribution languages that relax the global structure. We believe this out-of-distribution generalization holds in domains beyond formal language learning, and characterizing an LLM's implicit world model requires accounting for it.

Learning Succinct Representations. Succinctness is the ability of a learner to compactly express a language (Meyer & Fischer, 1971). Recent theoretical results show that transformers are more succinct than alternate formalisms such as LTL or automata (Bergsträßer et al., 2026). However, transformers' succinctness result has not been connected to learnability. In our work, we focus on hierarchical grammars, where an implicit notion of succinctness is the number of rules in the grammar: a hierarchical grammar is more succinct than a flattened one, since it reuses the same non-terminals across many derivations instead of enumerating each string from scratch. We empirically show that an LLM trained on a hierarchical grammar generalizes to other out-of-distribution hierarchical grammars, and language proficiency correlates with which level of the hierarchy the relaxed rule belongs to. Such generalization suggests that the LLM may have implicitly learned a succinct hierarchical grammar. A full exploration of succinct learning is out of scope for this paper.

Creativity in Language Learning. Creativity in LLMs is usually studied at inference time, e.g., via prompts or constrained decoding; we instead ask what an LLM's training reveals about it. Boden (2003) defines creativity as the ability to come up with ideas or artifacts that are new, surprising, and valuable, and identifies three forms it can take: combinatorial, exploratory, and transformational; Franceschelli & Musolesi (2024) apply this taxonomy to LLMs. Our results show that when an LLM learns a language, it generalizes to out-of-distribution languages that relax global structure, i.e., despite small edits to global rules, language proficiency in the resulting language is still retained. Therefore, an LLM's creativity may be tied to its internal representation, which is developed during training. Our future work focuses on formally establishing such a connection between creativity and language learning.

7 CONCLUSION

We are motivated by the question of an LLM's implicit world model: when an LLM learns a language, does it learn the same underlying structure of the language? Our approach is to observe what other languages the LLM learns beyond the target. To this end, we define the learning environment based on a probabilistic formal language, generated by a hierarchical context-free grammar whose structure is known and controllable. We then define a local discriminative test for measuring language proficiency, granular at the level of individual strings and based on a string's generation probability relative to its neighboring incorrect strings. Applying the test to the learning environment, we empirically find that an LLM not only learns the target language but also out-of-distribution languages that relax the global language structure while preserving the local structure. Simultaneously, the LLM learns local structure in the language better than global structure, suggesting a world model that is locally close to the target but globally relaxed. The local-global asymmetry is also validated in natural language. We hope this work encourages the community to adopt formal language as an environment simulator for studying LLMs.

AI USE STATEMENT

In this work, we used a generative AI tool (Claude) to assist with editing and polishing the text throughout the manuscript. We did not use generative AI tools to generate the research ideas, experimental results, or scientific claims presented in this work. All AI-assisted texts were reviewed by the authors before inclusion. We take responsibility for the final content of this work.

ETHICS STATEMENT

Our experiments do not involve human subjects. Most of our experiments use synthetic, probabilistic formal languages that we collect and adopt from Ghosh et al. (2026), avoiding the privacy, contamination, and bias concerns associated with scraped natural-language corpora. Our natural-language experiments use BLiMP (Warstadt et al., 2020), a publicly released, established benchmark of linguistically-constructed minimal pairs, as well as text sampled from FineWeb-Edu (Lozhkov et al., 2024) and FineWeb2 (Penedo et al., 2025), both publicly released web-text datasets; none of these contain personal or sensitive content. We fine-tune publicly available, open-weight LLMs (Appendix C.1) and do not release any new model weights. We do not foresee harmful applications specific to this work.

REPRODUCIBILITY STATEMENT

We fully specify the 12 grammars used in our experiments in Appendix C.3, allowing exact reconstruction of the languages and derivation trees studied throughout the paper. Our local discriminative test (Section 4) and rule proficiency measure (Appendix D.3) are defined formally enough to be re-implemented directly from the paper. Appendix C.1 and Appendix C.2 list the models, learning rates, training/evaluation set sizes, number of epochs, and hardware used for all experiments. We include our code for grammar generation, training, and evaluation as supplementary material.

REFERENCES

Zeyuan Allen-Zhu and Yuanzhi Li. Physics of language models: Part 1, learning hierarchical language structures. *arXiv preprint arXiv:2305.13673*, 2023.

Idan Attias, Lev Reyzin, Nathan Srebro, and Gal Vardi. On the hardness of learning regular expressions. *arXiv preprint arXiv:2510.04834*, 2025.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pp. 610–623, 2021.

Pascal Bergsträßer, Ryan Cotterell, and Anthony W. Lin. Transformers are inherently succinct. In *International Conference on Learning Representations (ICLR)*, 2026.

Satwik Bhattamishra, Kabir Ahuja, and Navin Goyal. On the ability and limitations of transformers to recognize formal languages. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7096–7116, 2020.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar Van Der Wal. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning (ICML)*, pp. 2397–2430. PMLR, 2023.

Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K. Warmuth. Occam's razor. *Information Processing Letters*, 24(6):377–380, 1987.

Margaret A. Boden. *The Creative Mind: Myths and Mechanisms*. Routledge, 2003.

Taylor L. Booth. Probabilistic representation of formal languages. In *Conference Record of the 10th Annual Symposium on Switching and Automata Theory (SWAT 1969)*, pp. 74–81. IEEE, 1969.

Nadav Borenstein, Anej Svete, Robin Shing Moon Chan, Josef Valvoda, Franz Nowak, Isabelle Augenstein, Eleanor Chodroff, and Ryan Cotterell. What languages are easy to language-model? a perspective from learning probabilistic regular languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15115–15134, 2024.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 1877–1901, 2020.

Noam Chomsky. Three models for the description of language. *IRE Transactions on Information Theory*, 2(3):113–124, 1956.

Grégoire Delétang, Anian Ruoss, Jordi Grau-Moya, Tim Genewein, Li Kevin Wenliang, Elliot Catt, Chris Cundy, Marcus Hutter, Shane Legg, Joel Veness, and Pedro A. Ortega. Neural networks and the Chomsky hierarchy. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023.

Giorgio Franceschelli and Mirco Musolesi. On the creativity of large language models. *AI & Society*, 2024.

Gemma Team. Gemma 2: Improving open language models at a practical size, 2024. URL `https://arxiv.org/abs/2408.00118`.

Bishwamittra Ghosh, Soumi Das, Till Speicher, Qinyuan Wu, Mohammad Aflah Khan, Deepak Garg, Krishna P Gummadi, and Evimaria Terzi. Fine-tuning vs. in-context learning in large language models: A formal language learning perspective. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 41651–41697, 2026.

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

Wes Gurnee and Max Tegmark. Language models represent space and time. In *International Conference on Learning Representations (ICLR)*, 2024.

John Hewitt and Christopher D. Manning. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pp. 4129–4138, 2019.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. URL `https://arxiv.org/abs/2310.06825`.

Urvashi Khandelwal, He He, Peng Qi, and Dan Jurafsky. Sharp nearby, fuzzy far away: How neural language models use context. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 284–294, 2018.

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

Kenneth Li, Aspen K. Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Emergent world representations: Exploring a sequence model trained on a synthetic task. In *International Conference on Learning Representations (ICLR)*, 2023.

Llama Team. The llama 3 herd of models, 2024. URL `https://arxiv.org/abs/2407.21783`.

Anton Lozhkov, Loubna Ben Allal, Leandro von Werra, and Thomas Wolf. Fineweb-edu: the finest collection of educational content, 2024. URL https://huggingface.co/datasets/HuggingFaceFW/fineweb-edu.

Albert R. Meyer and Michael J. Fischer. Economy of description by automata, grammars, and formal systems. In *12th Annual Symposium on Switching and Automata Theory (SWAT 1971)*, pp. 188–191. IEEE, 1971.

Neel Nanda, Andrew Lee, and Martin Wattenberg. Emergent linear representations in world models of self-supervised sequence models. In *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 16–30, 2023.

Guilherme Penedo, Hynek Kydlíček, Vinko Sabolčec, Bettina Messmer, Negar Foroutan, Amir Hossein Kargaran, Colin Raffel, Martin Jaggi, Leandro Von Werra, and Thomas Wolf. Fineweb2: One pipeline to scale them all – adapting pre-training data processing to every language, 2025.

Stuart M. Shieber. Evidence against the context-freeness of natural language. *Linguistics and Philosophy*, 8:333–343, 1985.

Herbert A. Simon. *The Sciences of the Artificial*. MIT Press, Cambridge, MA, 3rd edition, 1996.

Mirac Suzgun, Yonatan Belinkov, Stuart Shieber, and Sebastian Gehrmann. LSTM networks can perform dynamic counting. In *Proceedings of the Workshop on Deep Learning and Formal Languages: Building Bridges*, pp. 44–54, Florence, 2019a.

Mirac Suzgun, Yonatan Belinkov, and Stuart M. Shieber. On evaluating the generalization of LSTM models in formal languages. In *Proceedings of the Society for Computation in Linguistics (SCiL) 2019*, pp. 277–286, 2019b.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. Bert rediscovers the classical nlp pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 4593–4601, 2019.

Shubham Toshniwal, Sam Wiseman, Karen Livescu, and Kevin Gimpel. Chess as a testbed for language model state tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 11385–11393, 2022.

Keyon Vafa, Justin Y. Chen, Ashesh Rambachan, Jon Kleinberg, and Sendhil Mullainathan. Evaluating the world model implicit in a generative model. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. Blimp: The benchmark of linguistic minimal pairs for english. *Transactions of the Association for Computational Linguistics*, 8:377–392, 2020. doi: 10.1162/tacl_a_00321. URL https://doi.org/10.1162/tacl_a_00321.

Gail Weiss, Yoav Goldberg, and Eran Yahav. Thinking like transformers. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pp. 11080–11090, 2021.

Cheng Xu, Shuhao Guan, Derek Greene, and M-Tahar Kechadi. Benchmark data contamination of large language models: A survey. *arXiv preprint arXiv:2406.04244*, 2024.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Outline. This appendix is organized as follows.

- Appendix A: Extended Related Work.
- Appendix B: Conceptual Comparison among Language Proficiency Tests.
- Appendix C: Experimental Setup (Models, Training Details, Grammar Definitions).
- Appendix D: Results (Language Proficiency, Proficiency Test Comparisons, Rule-Level Proficiency Analysis, Natural Language).
- Data Preparation: English, showing example edits used for the natural-language experiments.

A Extended Related Work

A.1 Implicit World Models of LLMs

Approaches to an LLM's implicit world model can be divided into open-box and closed-box, where the former pokes inside the model while the latter observes what comes out of the model: our work falls into the latter approach.

On the open-box side, Li et al. (2023) train a GPT variant on Othello move sequences and find that, although the board state is not linearly decodable from its activations, a nonlinear probe recovers it with high accuracy, evidence of an emergent internal world model despite no explicit supervision on the game's rules. Nanda et al. (2023) revisit this setting and show the same board state is in fact linearly decodable once the probing target is readjusted to the game player's view. Gurnee & Tegmark (2024) similarly probe LLMs trained on natural text, rather than a synthetic move-sequence task, and find that linear probes recover real-world spatial and temporal coordinates from internal activations. Closer to language structure specifically, Hewitt & Manning (2019) show that a single linear transformation of BERT's and ELMo's word representations makes squared ℓ_2 distance encode distance in the dependency parse tree and squared ℓ_2 norm encode parse-tree depth, i.e., an entire syntax tree is linearly embedded in the representation space. Tenney et al. (2019) instead probe which layer of BERT resolves which classical NLP task (part-of-speech tagging, constituents, dependencies, semantic roles, coreference), finding that these resolve in roughly the expected pipeline order with layer depth, though the model can revise earlier decisions using later, higher-level information. Across this line of work, the ground-truth world model is deterministic and externally specified (a game's rules, real-world coordinates, a parse tree), and the question is whether it is decodable from a model's activations at all. We ask a different question: not whether our *probabilistic* language is decodable from the LLM's activations, but whether the LLM learns the same underlying structure, and if not, how it differs and whether it becomes proficient in other languages as well.

On the closed-box side, Toshniwal et al. (2022) train language models directly on chess move sequences and evaluate them by reading the board state off the model's own predicted tokens (i.e., predicted legal moves). They find that, given enough training data, the model learns to track piece positions and predict legal moves with high accuracy from move sequences alone, though this ability degrades sharply when full attention over the entire game history is approximated. Vafa et al. (2024) similarly evaluate an LLM's implicit world model against deterministic finite automata (DFA) purely from its generated behavior. Using "compression" and "distinction" metrics inspired by the Myhill-Nerode theorem, they find the recovered world model far less coherent than its surface behavior suggests, even though the LLM performs well on conventional next-token-prediction diagnostics. The two findings (Toshniwal et al.'s and Vafa et al.'s) are not strictly contradictory: an accurate readout of the current state does not guarantee coherence of the full transition structure a model implicitly computes. Our work improves upon both by considering a probabilistic formal language, a suitable object of learning for a language model. Our local discriminative test for language proficiency therefore applies more generally, to any grammar-defined language rather than only those representable as a DFA, unlike Vafa et al. (2024)'s test.

A.2 Formal Language and LLMs

Formal language offers a controlled setting for studying the representational expressiveness of sequence models: what classes of languages a given architecture can represent or recognize at all, independent of whether that architecture is ever trained to do so. This question has been studied

for non-transformer architectures such as RNNs and LSTMs (Suzgun et al., 2019a), and for transformers (Bhattamishra et al., 2020; Delétang et al., 2023): both lines of work ask what a given architecture is theoretically *capable* of representing, rather than what a specific trained instance of it actually learns. Weiss et al. (2021) instead ask what computation model explains transformers' inner workings, proposing RASP, a restricted programming language that captures what transformer encoders can compute. RASP programs are hand-written to match a target computation, rather than learned from data. Across this line of work, models are typically small, single-task, and not decoder-only, unlike the widely used LLMs we study.

Allen-Zhu & Li (2023) instead train large transformers on an HPCFG designed to mimic natural language, motivating our own use of such a grammar (Section 3); unlike the expressiveness and RASP literature above, their transformers are actually trained on data sampled from the grammar, closer to our own setup. Allen-Zhu & Li (2023) also perform attention probing to show that the LLM learns a hierarchical representation.

Bergsträßer et al. (2026) recast the expressiveness question as one of succinctness, proving that fixed-precision transformers are exponentially more succinct in representing a language than LTL formulas and doubly exponentially more succinct than finite automata. The relationship between succinctness and learnability can run either way in classical learning theory: Occam learning shows succinct hypotheses are more PAC-learnable (Blumer et al., 1987), yet regular expressions, though succinct relative to automata, are provably harder to learn (Attias et al., 2025). An empirical approach connecting succinctness with learnability could build on our approach of teaching hierarchical grammars to LLMs (Section 5), though we leave the connection itself to future work (Section 6).

A separate, computation-free line of work asks a theoretical question: does a learner identify a language in the limit (Gold, 1967)? Kleinberg & Mullainathan (2024) recently revisit this question as language generation in the limit, asking whether a learner can eventually generate valid strings rather than recognize them. Both lines of work characterize what an idealized learner could achieve in principle, unconstrained by architecture, training procedure, or computational budget. Our paper instead studies an actual, already-trained LLM under a fixed, realistic training procedure, not an idealized learner unconstrained by architecture.

A.3 Measuring Language Proficiency

A full technical comparison of our local discriminative test against generative, distributional-difference, and global discriminative tests is given in Appendix B.

B Conceptual Comparison among Language Proficiency Tests

We compare our local discriminative test with (i) a generative test, (ii) a distributional-difference-based test, and (iii) a global discriminative test. Unlike ours, (i) and (ii) consider strings inside the language, ignoring LLMs' generation ability outside the language. Also, (ii) and (iii) cannot answer when an *individual* string is learned, due to their global nature; our local discriminative test can.

Generative Tests. The generative test of Allen-Zhu & Li (2023) is based on token generation accuracy: they fix a prefix in a string and declares it learned if the LLM can generate suffix tokens accurately so that the resulting string belongs to the language; language proficiency is measured by what fraction of prefixes can be completed with a grammatically correct suffix. Two concerns remain: (a) deciding the length of the prefix, which is determined in an *ad hoc* manner (e.g., Allen-Zhu & Li (2023) use a prefix of either 0 or 50 tokens); and (b) deciding which token to generate, since the LLM induces a distribution over tokens given a prefix, *requiring a commitment to the sampling strategy* (e.g., Allen-Zhu & Li use multinomial sampling at temperature 1).

To avoid the ad hoc prefix length, Suzgun et al. (2019a;b); Bhattamishra et al. (2020) and Delétang et al. (2023) evaluate every position in a string. This still leaves the second issue: at each position, the LLM only provides a probability for each candidate next token. Thus, scoring a prediction as correct or incorrect requires committing to a conversion rule: either a probability threshold (is the model's probability for each ground-truth token above some cutoff, e.g., 0.5) as in Suzgun et al. (2019a;b); Bhattamishra et al. (2020), or an argmax match (does the model's single highest-probability token equal the ground-truth token) as in Delétang et al. (2023), the same underlying issue as committing

to a sampling strategy. To summarize, all these tests essentially find a set of grammatically correct strings produced by the LLM, while the LLM naturally yields a distribution over both grammatically correct and incorrect strings, a fundamental mismatch; Borenstein et al. (2024) criticize this *set vs. distribution* mismatch and propose a test that directly compares two distributions.

Any generative test based on accuracy can be computed in terms of loss over correct token generation, essentially considering a continuous loss of generation than a binary correctness/incorrectness prediction. In Section 4, we apply loss-based generative tests in our analysis.

Distributional-Difference-Based Tests. An LLM and a probabilistic formal language both yield a probability distribution over strings. So Borenstein et al. (2024); Allen-Zhu & Li (2023) propose to compare them via KL divergence, $D_{KL}(P_L \parallel P_\theta) \triangleq \sum_{s\in\mathbf{\Sigma}^*} P_L(s) \log \frac{P_L(s)}{P_\theta(s)}$, the cost of approximating P_L with P_θ. But, KL divergence is not symmetric and *carries no meaningful signal from grammatically incorrect strings*. For any $s \notin L$, we have $P_L(s) = 0$, and since P_θ has full support over $\boldsymbol{V}^* \supseteq \mathbf{\Sigma}^*$ (Section 3.1), $P_\theta(s) > 0$. The term contributed by such $s \notin L$ in each direction is

$$\underbrace{P_L(s) \log \frac{P_L(s)}{P_\theta(s)}}_{\text{term in } D_{KL}(P_L \| P_\theta)} = 0 \cdot \log \frac{0}{P_\theta(s)} = 0, \qquad \underbrace{P_\theta(s) \log \frac{P_\theta(s)}{P_L(s)}}_{\text{term in } D_{KL}(P_\theta \| P_L)} = P_\theta(s) \log \frac{P_\theta(s)}{0} = +\infty.$$

So $D_{KL}(P_L \parallel P_\theta)$ adds zero cost from grammatically incorrect strings, while $D_{KL}(P_\theta \parallel P_L)$ becomes infinite because of them. This is the critical distinction from our discriminative test, which considers strings both inside and outside L in a meaningful way (local minimum vs. local boundary).

A nuanced yet practical concern arises whenever an already-pretrained LLM is used (fine-tuned or prompted in-context), where the language-vocabulary $\mathbf{\Sigma}$ is generally a strict subset of the LLM-vocabulary $\boldsymbol{V}$ (Section 3.1). Borenstein et al. (2024); Allen-Zhu & Li (2023) avoid this altogether by training the LLM from scratch on a matching vocabulary, $\boldsymbol{V} = \mathbf{\Sigma}$, which is a more restricted, less practical setting than using a pre-trained LLM. When $\mathbf{\Sigma} \subsetneq \boldsymbol{V}$, whether to renormalize P_θ to $\mathbf{\Sigma}$ or use its raw probability over $\boldsymbol{V}$ is unclear, and the two choices yield different KL-divergence estimates. In contrast, our discriminative test avoids this concern by being agnostic to absolute probabilities (or losses), and instead focuses on the relative probabilities between grammatical and non-grammatical strings, which are well-defined regardless of how $\mathbf{\Sigma}$ and $\boldsymbol{V}$ are defined.

Local vs. Global Discriminative Tests. Ghosh et al. (2026) study two learning modes in LLMs: fine-tuning vs. in-context learning, and propose a discriminative test that is comparable between them. In their test, incorrect strings are similarly constructed by editing grammatically correct strings up to edit distance k. But, their test is global rather than local: All correct and incorrect strings are treated as a single binary classification problem based on loss, where the classification metric AUC is used as the proficiency score. At a granular level, AUC is the probability that a randomly sampled correct string has lower loss than a randomly sampled incorrect string, drawn from the entire dataset rather than from a specific neighborhood. *The global test therefore over-estimates our local one*: outperforming a random incorrect string is easier than outperforming the hardest incorrect string within a k-edit neighborhood, which our local discriminative test requires.

C Experimental Setup

C.1 Models

We evaluate 6 LLMs: Qwen2.5-1.5B and Qwen2.5-7B, Pythia-1B, Mistral-7B-v0.3, Gemma-2-2B, and Llama-3.2-1B (Table 1).

Model	HuggingFace Identifier	Learning Rate
Qwen2.5-1.5B (Yang et al., 2024)	`Qwen/Qwen2.5-1.5B`	5×10^{-5}
Qwen2.5-7B (Yang et al., 2024)	`Qwen/Qwen2.5-7B`	5×10^{-5}
Pythia-1B (Biderman et al., 2023)	`EleutherAI/pythia-1b`	1×10^{-5}
Mistral-7B-v0.3 (Jiang et al., 2023)	`mistralai/Mistral-7B-v0.3`	5×10^{-6}
Gemma-2-2B (Gemma Team, 2024)	`google/gemma-2-2b`	5×10^{-5}
Llama-3.2-1B (Llama Team, 2024)	`meta-llama/Llama-3.2-1B`	5×10^{-5}

Table 1: LLMs used in our experiments, with their fine-tuning learning rate.

C.2 Training Details

We fine-tune each model for 20 epochs on training sets of size 16, 64, 256, and 1024 strings sampled from the language, using cross-entropy loss as the training objective. We consider three random seeds when sampling the training set. For evaluation, we use 256 grammatically correct strings sampled from the language, each paired with at least two grammatically incorrect neighbors. Throughout, we consider edit $k = 1$ both on strings and rules. The training and test sets are disjoint and both sampled from the language such that the frequency distribution of strings is preserved. All experiments use Python 3.10.0. We run experiments on 8x Nvidia H100 94GB NVL GPUs and 2x AMD EPYC 9554 CPU @ 3.1 GHz, 2x64 cores, and 24x 96GB RAM.

C.3 Grammar Definitions

The grammars are adapted from the codebase of Ghosh et al. (2026).

Grammar	Probability	Language
G_1^{uniform}	Uniform	L_1
G_1^{skewed}	Skewed	L_1
G_2^{uniform}	Uniform	L_2
G_2^{skewed}	Skewed	L_2
G_3^{uniform}	Uniform	L_3
G_3^{skewed}	Skewed	L_3
G_4^{uniform}	Uniform	L_4
G_4^{skewed}	Skewed	L_4
G_5^{uniform}	Uniform	L_5
G_5^{skewed}	Skewed	L_5
G_6^{uniform}	Uniform	L_6
G_6^{skewed}	Skewed	L_6

Table 2: Notation for the grammars used in the Appendix figures. The subscript indexes the language; the superscript indicates uniform or skewed rule probability. G_i^{uniform} and G_i^{skewed} accept the same language L_i, differing only in rule probability.

$S \to A18\ A16$ [0.50]	Level 4	$S \to A18\ A16$ [0.90]	Level 4
$S \to A16\ A18\ A17$ [0.50]		$S \to A16\ A18\ A17$ [0.10]	
$A18 \to A15\ A14\ A13$ [0.50]	Level 3	$A18 \to A15\ A14\ A13$ [0.90]	Level 3
$A18 \to A14\ A15\ A13$ [0.50]		$A18 \to A14\ A15\ A13$ [0.10]	
$A17 \to A14\ A13\ A15$ [0.50]		$A17 \to A14\ A13\ A15$ [0.90]	
$A17 \to A13\ A14\ A15$ [0.50]		$A17 \to A13\ A14\ A15$ [0.10]	
$A16 \to A14\ A15$ [0.50]		$A16 \to A14\ A15$ [0.90]	
$A16 \to A15\ A14$ [0.50]		$A16 \to A15\ A14$ [0.10]	
$A15 \to A11\ A12\ A10$ [0.50]	Level 2	$A15 \to A11\ A12\ A10$ [0.90]	Level 2
$A15 \to A12\ A10\ A11$ [0.50]		$A15 \to A12\ A10\ A11$ [0.10]	
$A14 \to A11\ A10\ A12$ [0.50]		$A14 \to A11\ A10\ A12$ [0.90]	
$A14 \to A10\ A11\ A12$ [0.50]		$A14 \to A10\ A11\ A12$ [0.10]	
$A13 \to A10\ A12\ A11$ [0.50]		$A13 \to A10\ A12\ A11$ [0.90]	
$A13 \to A12\ A11\ A10$ [0.50]		$A13 \to A12\ A11\ A10$ [0.10]	
$A12 \to 9\ 8\ 7$ [0.50]	Level 1	$A12 \to 9\ 8\ 7$ [0.90]	Level 1
$A12 \to 8\ 7$ [0.50]		$A12 \to 8\ 7$ [0.10]	
$A11 \to 6\ 5$ [0.50]		$A11 \to 6\ 5$ [0.90]	
$A11 \to 6\ 4\ 5$ [0.50]		$A11 \to 6\ 4\ 5$ [0.10]	
$A10 \to 3\ 1$ [0.50]		$A10 \to 3\ 1$ [0.90]	
$A10 \to 1\ 2\ 3$ [0.50]		$A10 \to 1\ 2\ 3$ [0.10]	

Figure 10: G_1^{uniform} (left) and G_1^{skewed} (right).



Figure 11: A string sampled from G_1 (Figure 10), with its derivation tree. Since G_1^{uniform} and G_1^{skewed} accept the same language, one example suffices for both.

$S \to A18\ A17\ A16\ [0.50]$	Level 4	$S \to A18\ A17\ A16\ [0.95]$	Level 4
$S \to A16\ A18\ A17\ [0.50]$		$S \to A16\ A18\ A17\ [0.05]$	
$A18 \to A15\ A14\ A13\ [0.50]$	Level 3	$A18 \to A15\ A14\ A13\ [0.95]$	Level 3
$A18 \to A14\ A15\ A13\ [0.50]$		$A18 \to A14\ A15\ A13\ [0.05]$	
$A17 \to A14\ A13\ A15\ [0.50]$		$A17 \to A14\ A13\ A15\ [0.95]$	
$A17 \to A13\ A14\ A15\ [0.50]$		$A17 \to A13\ A14\ A15\ [0.05]$	
$A16 \to A14\ A15\ [0.50]$		$A16 \to A14\ A15\ [0.95]$	
$A16 \to A15\ A14\ [0.50]$		$A16 \to A15\ A14\ [0.05]$	
$A15 \to A11\ A12\ A10\ [0.50]$	Level 2	$A15 \to A11\ A12\ A10\ [0.95]$	Level 2
$A15 \to A12\ A10\ A11\ [0.50]$		$A15 \to A12\ A10\ A11\ [0.05]$	
$A14 \to A11\ A10\ A12\ [0.50]$		$A14 \to A11\ A10\ A12\ [0.95]$	
$A14 \to A10\ A11\ A12\ [0.50]$		$A14 \to A10\ A11\ A12\ [0.05]$	
$A13 \to A10\ A12\ A11\ [0.50]$		$A13 \to A10\ A12\ A11\ [0.95]$	
$A13 \to A12\ A11\ A10\ [0.50]$		$A13 \to A12\ A11\ A10\ [0.05]$	
$A12 \to 9\ 8\ 7\ [0.50]$	Level 1	$A12 \to 9\ 8\ 7\ [0.95]$	Level 1
$A12 \to 8\ 7\ 9\ [0.50]$		$A12 \to 8\ 7\ 9\ [0.05]$	
$A11 \to 6\ 5\ 4\ [0.50]$		$A11 \to 6\ 5\ 4\ [0.95]$	
$A11 \to 6\ 4\ 5\ [0.50]$		$A11 \to 6\ 4\ 5\ [0.05]$	
$A10 \to 3\ 1\ 2\ [0.50]$		$A10 \to 3\ 1\ 2\ [0.95]$	
$A10 \to 1\ 2\ 3\ [0.50]$		$A10 \to 1\ 2\ 3\ [0.05]$	

Figure 12: G_2^{uniform} (left) and G_2^{skewed} (right).

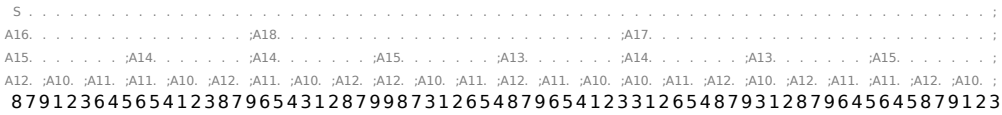


Figure 13: A string sampled from G_2 (Figure 12), with its derivation tree. Since G_2^{uniform} and G_2^{skewed} accept the same language, one example suffices for both.

$S \to A16\ A17\ A16$ [0.50]	Level 4	$S \to A16\ A17\ A16$ [0.95]	Level 4
$S \to A17\ A17\ A16$ [0.50]		$S \to A17\ A17\ A16$ [0.05]	
$A17 \to A14\ A14\ A15$ [0.50]	Level 3	$A17 \to A14\ A14\ A15$ [0.95]	Level 3
$A17 \to A14\ A15\ A14$ [0.50]		$A17 \to A14\ A15\ A14$ [0.05]	
$A16 \to A13\ A13\ A15$ [0.50]		$A16 \to A13\ A13\ A15$ [0.95]	
$A16 \to A13\ A15\ A13$ [0.50]		$A16 \to A13\ A15\ A13$ [0.05]	
$A15 \to A11\ A11\ A10$ [0.50]	Level 2	$A15 \to A11\ A11\ A10$ [0.95]	Level 2
$A15 \to A11\ A10\ A11$ [0.50]		$A15 \to A11\ A10\ A11$ [0.05]	
$A14 \to A11\ A12\ A12$ [0.50]		$A14 \to A11\ A12\ A12$ [0.95]	
$A14 \to A12\ A11\ A12$ [0.50]		$A14 \to A12\ A11\ A12$ [0.05]	
$A13 \to A10\ A10\ A11$ [0.50]		$A13 \to A10\ A10\ A11$ [0.95]	
$A13 \to A10\ A11\ A10$ [0.50]		$A13 \to A10\ A11\ A10$ [0.05]	
$A12 \to 9\ 8\ 7$ [0.50]	Level 1	$A12 \to 9\ 8\ 7$ [0.95]	Level 1
$A12 \to 8\ 7\ 9$ [0.50]		$A12 \to 8\ 7\ 9$ [0.05]	
$A11 \to 6\ 5\ 4$ [0.50]		$A11 \to 6\ 5\ 4$ [0.95]	
$A11 \to 6\ 4\ 5$ [0.50]		$A11 \to 6\ 4\ 5$ [0.05]	
$A10 \to 3\ 1\ 2$ [0.50]		$A10 \to 3\ 1\ 2$ [0.95]	
$A10 \to 1\ 2\ 3$ [0.50]		$A10 \to 1\ 2\ 3$ [0.05]	

Figure 14: G_3^{uniform} (left) and G_3^{skewed} (right).

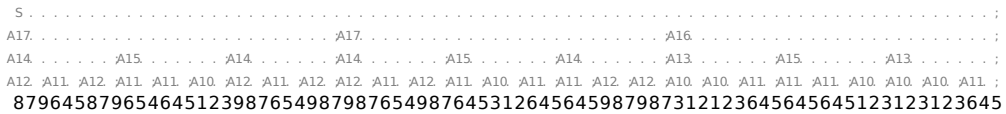


Figure 15: A string sampled from G_3 (Figure 14), with its derivation tree. Since G_3^{uniform} and G_3^{skewed} accept the same language, one example suffices for both.

$S \to A_3^2\ A_3^2\ A_3^2\ [0.50]$	Level 4	$S \to A_3^2\ A_3^2\ A_3^2\ [0.159]$	Level 4
$S \to A_3^0\ A_3^1\ A_3^1\ [0.50]$		$S \to A_3^0\ A_3^1\ A_3^1\ [0.841]$	
$A_3^2 \to A_2^0\ A_2^2\ A_2^0\ [0.50]$	Level 3	$A_3^2 \to A_2^0\ A_2^2\ A_2^0\ [0.841]$	Level 3
$A_3^2 \to A_2^2\ A_2^2\ A_2^2\ [0.50]$		$A_3^2 \to A_2^2\ A_2^2\ A_2^2\ [0.159]$	
$A_3^1 \to A_2^1\ A_2^2\ A_2^2\ [0.50]$		$A_3^1 \to A_2^1\ A_2^2\ A_2^2\ [0.841]$	
$A_3^1 \to A_2^2\ A_2^1\ A_2^2\ [0.50]$		$A_3^1 \to A_2^2\ A_2^1\ A_2^2\ [0.159]$	
$A_3^0 \to A_2^0\ A_2^0\ A_2^2\ [0.50]$		$A_3^0 \to A_2^0\ A_2^0\ A_2^2\ [0.841]$	
$A_3^0 \to A_2^0\ A_2^1\ A_2^0\ [0.50]$		$A_3^0 \to A_2^0\ A_2^1\ A_2^0\ [0.159]$	
$A_2^2 \to A_1^2\ A_1^0\ A_1^0\ [0.50]$	Level 2	$A_2^2 \to A_1^2\ A_1^0\ A_1^0\ [0.841]$	Level 2
$A_2^2 \to A_1^0\ A_1^0\ A_1^1\ [0.50]$		$A_2^2 \to A_1^0\ A_1^0\ A_1^1\ [0.159]$	
$A_2^1 \to A_1^0\ A_1^0\ A_1^2\ [0.50]$		$A_2^1 \to A_1^0\ A_1^0\ A_1^2\ [0.841]$	
$A_2^1 \to A_1^0\ A_1^0\ A_1^1\ [0.50]$		$A_2^1 \to A_1^0\ A_1^0\ A_1^1\ [0.159]$	
$A_2^0 \to A_1^1\ A_1^1\ A_1^1\ [0.50]$		$A_2^0 \to A_1^1\ A_1^1\ A_1^1\ [0.841]$	
$A_2^0 \to A_1^1\ A_1^2\ A_1^2\ [0.50]$		$A_2^0 \to A_1^1\ A_1^2\ A_1^2\ [0.159]$	
$A_1^2 \to 2\ 1\ 1\ [0.50]$	Level 1	$A_1^2 \to 2\ 1\ 1\ [0.841]$	Level 1
$A_1^2 \to 7\ 1\ 6\ [0.50]$		$A_1^2 \to 7\ 1\ 6\ [0.159]$	
$A_1^1 \to 4\ 3\ 7\ [0.50]$		$A_1^1 \to 4\ 3\ 7\ [0.841]$	
$A_1^1 \to 1\ 3\ 9\ [0.50]$		$A_1^1 \to 1\ 3\ 9\ [0.159]$	
$A_1^0 \to 4\ 0\ 0\ [0.50]$		$A_1^0 \to 4\ 0\ 0\ [0.841]$	
$A_1^0 \to 5\ 4\ 5\ [0.50]$		$A_1^0 \to 5\ 4\ 5\ [0.159]$	

Figure 16: G_4^{uniform} (left) and G_4^{skewed} (right).

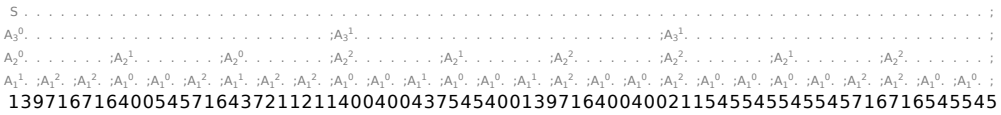


Figure 17: A string sampled from G_4 (Figure 16), with its derivation tree. Since G_4^{uniform} and G_4^{skewed} accept the same language, one example suffices for both.

$S \to A_4^2\ A_4^1\ A_4^0$ [0.50]	Level 5	$S \to A_4^2\ A_4^1\ A_4^0$ [0.159]	Level 5
$S \to A_4^2\ A_4^1\ A_4^0$ [0.50]		$S \to A_4^2\ A_4^1\ A_4^0$ [0.841]	
$A_4^2 \to A_3^2\ A_3^0\ A_3^2$ [0.50]	Level 4	$A_4^2 \to A_3^2\ A_3^0\ A_3^2$ [0.841]	Level 4
$A_4^2 \to A_3^0\ A_3^2\ A_3^1$ [0.50]		$A_4^2 \to A_3^0\ A_3^2\ A_3^1$ [0.159]	
$A_4^1 \to A_3^2\ A_3^0\ A_3^2$ [0.50]		$A_4^1 \to A_3^2\ A_3^0\ A_3^2$ [0.841]	
$A_4^1 \to A_3^2\ A_3^1\ A_3^1$ [0.50]		$A_4^1 \to A_3^2\ A_3^1\ A_3^1$ [0.159]	
$A_4^0 \to A_3^0\ A_3^1\ A_3^1$ [0.50]		$A_4^0 \to A_3^0\ A_3^1\ A_3^1$ [0.841]	
$A_4^0 \to A_3^2\ A_3^2\ A_3^2$ [0.50]		$A_4^0 \to A_3^2\ A_3^2\ A_3^2$ [0.159]	
$A_3^2 \to A_2^0\ A_2^2\ A_2^0$ [0.50]	Level 3	$A_3^2 \to A_2^0\ A_2^2\ A_2^0$ [0.841]	Level 3
$A_3^2 \to A_2^2\ A_2^2\ A_2^2$ [0.50]		$A_3^2 \to A_2^2\ A_2^2\ A_2^2$ [0.159]	
$A_3^1 \to A_2^1\ A_2^2\ A_2^2$ [0.50]		$A_3^1 \to A_2^1\ A_2^2\ A_2^2$ [0.841]	
$A_3^1 \to A_2^2\ A_2^1\ A_2^2$ [0.50]		$A_3^1 \to A_2^2\ A_2^1\ A_2^2$ [0.159]	
$A_3^0 \to A_2^0\ A_2^0\ A_2^2$ [0.50]		$A_3^0 \to A_2^0\ A_2^0\ A_2^2$ [0.841]	
$A_3^0 \to A_2^0\ A_2^1\ A_2^0$ [0.50]		$A_3^0 \to A_2^0\ A_2^1\ A_2^0$ [0.159]	
$A_2^2 \to A_1^2\ A_1^0\ A_1^0$ [0.50]	Level 2	$A_2^2 \to A_1^2\ A_1^0\ A_1^0$ [0.841]	Level 2
$A_2^2 \to A_1^0\ A_1^0\ A_1^1$ [0.50]		$A_2^2 \to A_1^0\ A_1^0\ A_1^1$ [0.159]	
$A_2^1 \to A_1^0\ A_1^0\ A_1^2$ [0.50]		$A_2^1 \to A_1^0\ A_1^0\ A_1^2$ [0.841]	
$A_2^1 \to A_1^0\ A_1^0\ A_1^1$ [0.50]		$A_2^1 \to A_1^0\ A_1^0\ A_1^1$ [0.159]	
$A_2^0 \to A_1^1\ A_1^1\ A_1^1$ [0.50]		$A_2^0 \to A_1^1\ A_1^1\ A_1^1$ [0.841]	
$A_2^0 \to A_1^1\ A_1^2\ A_1^2$ [0.50]		$A_2^0 \to A_1^1\ A_1^2\ A_1^2$ [0.159]	
$A_1^2 \to 2\ 1\ 1$ [0.50]	Level 1	$A_1^2 \to 2\ 1\ 1$ [0.841]	Level 1
$A_1^2 \to 7\ 1\ 6$ [0.50]		$A_1^2 \to 7\ 1\ 6$ [0.159]	
$A_1^1 \to 4\ 3\ 7$ [0.50]		$A_1^1 \to 4\ 3\ 7$ [0.841]	
$A_1^1 \to 1\ 3\ 9$ [0.50]		$A_1^1 \to 1\ 3\ 9$ [0.159]	
$A_1^0 \to 4\ 0\ 0$ [0.50]		$A_1^0 \to 4\ 0\ 0$ [0.841]	
$A_1^0 \to 5\ 4\ 5$ [0.50]		$A_1^0 \to 5\ 4\ 5$ [0.159]	

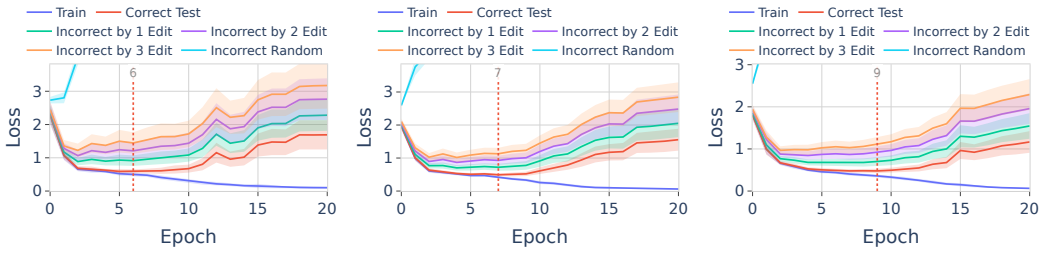
Figure 18: G_5^{uniform} (left) and G_5^{skewed} (right).

$S \to A_5^0\ A_5^0\ A_5^2$ [0.50]	Level 6	$S \to A_5^0\ A_5^0\ A_5^2$ [0.159]	Level 6
$S \to A_5^2\ A_5^1\ A_5^2$ [0.50]		$S \to A_5^2\ A_5^1\ A_5^2$ [0.841]	
$A_5^2 \to A_4^2\ A_4^1\ A_4^1$ [0.50]	Level 5	$A_5^2 \to A_4^2\ A_4^1\ A_4^1$ [0.841]	Level 5
$A_5^2 \to A_4^2\ A_4^0\ A_4^0$ [0.50]		$A_5^2 \to A_4^2\ A_4^0\ A_4^0$ [0.159]	
$A_5^1 \to A_4^1\ A_4^0\ A_4^2$ [0.50]		$A_5^1 \to A_4^1\ A_4^0\ A_4^2$ [0.841]	
$A_5^1 \to A_4^0\ A_4^1\ A_4^2$ [0.50]		$A_5^1 \to A_4^0\ A_4^1\ A_4^2$ [0.159]	
$A_5^0 \to A_4^2\ A_4^1\ A_4^0$ [0.50]		$A_5^0 \to A_4^2\ A_4^1\ A_4^0$ [0.841]	
$A_5^0 \to A_4^2\ A_4^1\ A_4^0$ [0.50]		$A_5^0 \to A_4^2\ A_4^1\ A_4^0$ [0.159]	
$A_4^2 \to A_3^2\ A_3^0\ A_3^2$ [0.50]	Level 4	$A_4^2 \to A_3^2\ A_3^0\ A_3^2$ [0.841]	Level 4
$A_4^2 \to A_3^0\ A_3^2\ A_3^1$ [0.50]		$A_4^2 \to A_3^0\ A_3^2\ A_3^1$ [0.159]	
$A_4^1 \to A_3^2\ A_3^0\ A_3^2$ [0.50]		$A_4^1 \to A_3^2\ A_3^0\ A_3^2$ [0.841]	
$A_4^1 \to A_3^2\ A_3^1\ A_3^1$ [0.50]		$A_4^1 \to A_3^2\ A_3^1\ A_3^1$ [0.159]	
$A_4^0 \to A_3^0\ A_3^1\ A_3^1$ [0.50]		$A_4^0 \to A_3^0\ A_3^1\ A_3^1$ [0.841]	
$A_4^0 \to A_3^2\ A_3^2\ A_3^2$ [0.50]		$A_4^0 \to A_3^2\ A_3^2\ A_3^2$ [0.159]	
$A_3^2 \to A_2^0\ A_2^2\ A_2^0$ [0.50]	Level 3	$A_3^2 \to A_2^0\ A_2^2\ A_2^0$ [0.841]	Level 3
$A_3^2 \to A_2^2\ A_2^2\ A_2^2$ [0.50]		$A_3^2 \to A_2^2\ A_2^2\ A_2^2$ [0.159]	
$A_3^1 \to A_2^1\ A_2^2\ A_2^2$ [0.50]		$A_3^1 \to A_2^1\ A_2^2\ A_2^2$ [0.841]	
$A_3^1 \to A_2^2\ A_2^1\ A_2^2$ [0.50]		$A_3^1 \to A_2^2\ A_2^1\ A_2^2$ [0.159]	
$A_3^0 \to A_2^0\ A_2^0\ A_2^2$ [0.50]		$A_3^0 \to A_2^0\ A_2^0\ A_2^2$ [0.841]	
$A_3^0 \to A_2^0\ A_2^1\ A_2^0$ [0.50]		$A_3^0 \to A_2^0\ A_2^1\ A_2^0$ [0.159]	
$A_2^2 \to A_1^2\ A_1^0\ A_1^0$ [0.50]	Level 2	$A_2^2 \to A_1^2\ A_1^0\ A_1^0$ [0.841]	Level 2
$A_2^2 \to A_1^0\ A_1^0\ A_1^1$ [0.50]		$A_2^2 \to A_1^0\ A_1^0\ A_1^1$ [0.159]	
$A_2^1 \to A_1^0\ A_1^0\ A_1^2$ [0.50]		$A_2^1 \to A_1^0\ A_1^0\ A_1^2$ [0.841]	
$A_2^1 \to A_1^0\ A_1^0\ A_1^1$ [0.50]		$A_2^1 \to A_1^0\ A_1^0\ A_1^1$ [0.159]	
$A_2^0 \to A_1^1\ A_1^1\ A_1^1$ [0.50]		$A_2^0 \to A_1^1\ A_1^1\ A_1^1$ [0.841]	
$A_2^0 \to A_1^1\ A_1^2\ A_1^2$ [0.50]		$A_2^0 \to A_1^1\ A_1^2\ A_1^2$ [0.159]	
$A_1^2 \to 2\ 1\ 1$ [0.50]	Level 1	$A_1^2 \to 2\ 1\ 1$ [0.841]	Level 1
$A_1^2 \to 7\ 1\ 6$ [0.50]		$A_1^2 \to 7\ 1\ 6$ [0.159]	
$A_1^1 \to 4\ 3\ 7$ [0.50]		$A_1^1 \to 4\ 3\ 7$ [0.841]	
$A_1^1 \to 1\ 3\ 9$ [0.50]		$A_1^1 \to 1\ 3\ 9$ [0.159]	
$A_1^0 \to 4\ 0\ 0$ [0.50]		$A_1^0 \to 4\ 0\ 0$ [0.841]	
$A_1^0 \to 5\ 4\ 5$ [0.50]		$A_1^0 \to 5\ 4\ 5$ [0.159]	

Figure 19: G_6^{uniform} (left) and G_6^{skewed} (right).

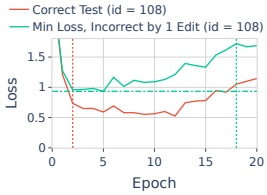
D RESULTS

D.1 LANGUAGE PROFICIENCY

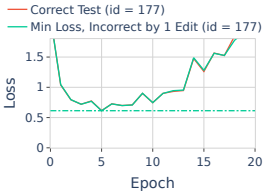


(a) G_1^{uniform} (b) G_2^{uniform} (c) G_3^{uniform}

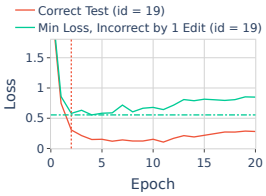
Figure 20: Generation loss of grammatically correct and incorrect strings across grammars (Table 2), where incorrect strings range from small edit distances to fully random strings. The loss gap between correct and incorrect strings increases with edit distance, showing why language proficiency must be measured by comparing strings that are close in edit distance (such as $k = 1$), yet belong to the opposite sides of the language boundary.



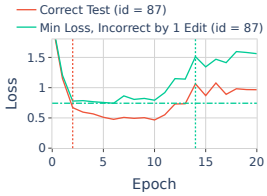
(a) G_1^{uniform}, learned



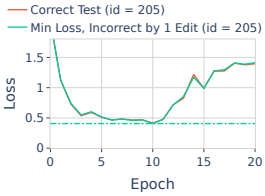
(b) G_1^{uniform}, never learned



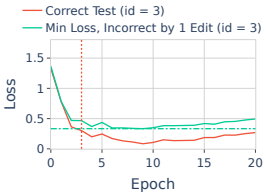
(c) G_1^{skewed}, learned



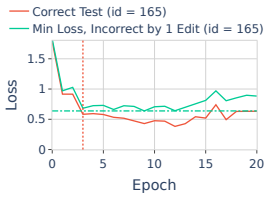
(d) G_2^{uniform}, learned



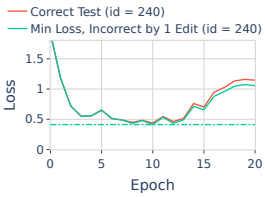
(e) G_2^{uniform}, never learned



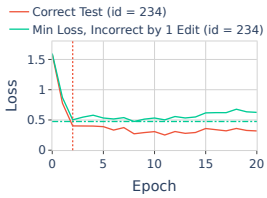
(f) G_2^{skewed}, learned



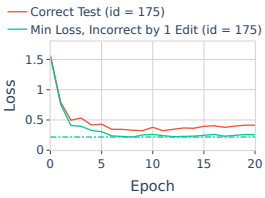
(g) G_3^{uniform}, learned



(h) G_3^{uniform}, never learned

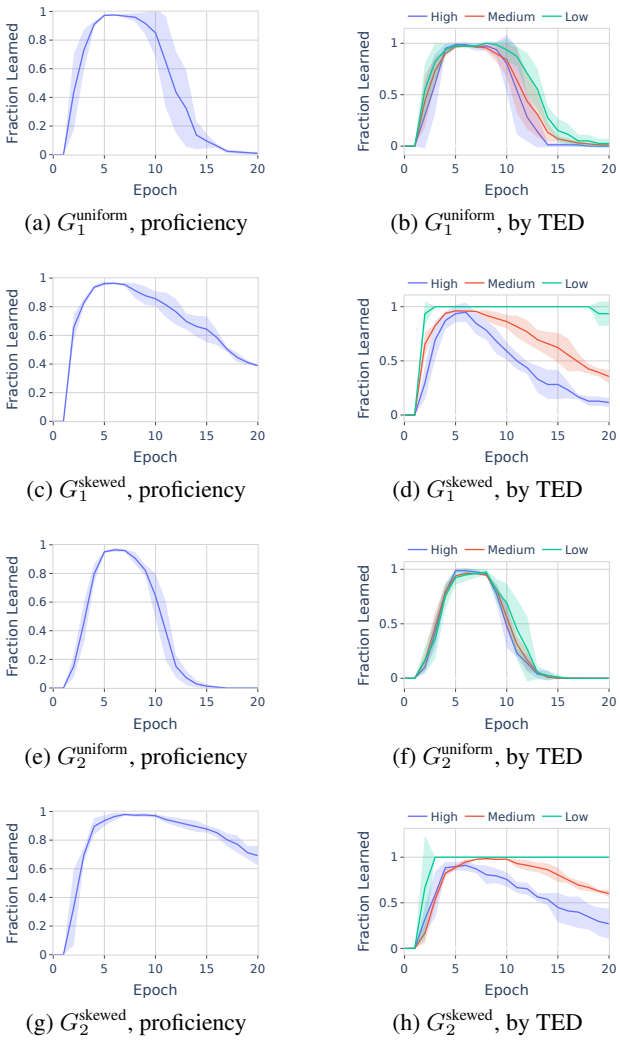


(i) G_3^{skewed}, learned



(j) G_3^{skewed}, never learned

Figure 21: Example strings that are learned, or never learned indifferent grammars, fixing the training seed to 1. For G_1^{skewed} and G_2^{skewed}, every string is learned at some point during training, so no never-learned example exists to show.



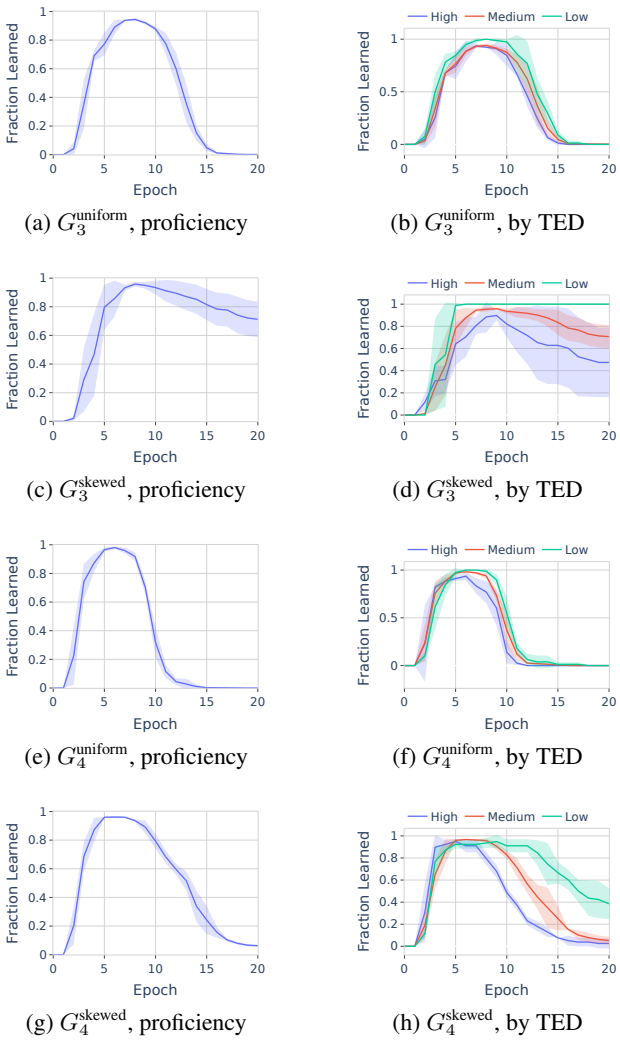
(a) G_1^{uniform}, proficiency (b) G_1^{uniform}, by TED

(c) G_1^{skewed}, proficiency (d) G_1^{skewed}, by TED

(e) G_2^{uniform}, proficiency (f) G_2^{uniform}, by TED

(g) G_2^{skewed}, proficiency (h) G_2^{skewed}, by TED

Figure 22: Language proficiency (fraction of strings learned over epochs) and its breakdown by tree-edit distance (TED) to the training set, for G_1 and G_2 (Table 2), aggregated over training seeds. The skew in rule probability between G_i^{uniform} and G_i^{skewed} is what produces the differentiated proficiency we observe across strings. In general, strings with lower edit distance to the training set are learned faster and retain their learned status longer, which is observed more effectively in a skewed language with variable distances.



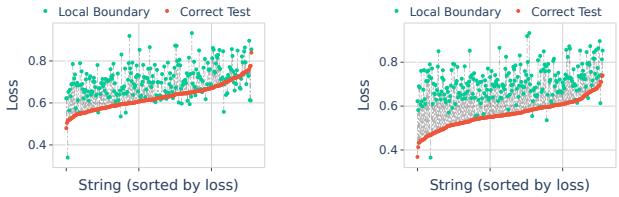
(a) G_3^{uniform}, proficiency (b) G_3^{uniform}, by TED

(c) G_3^{skewed}, proficiency (d) G_3^{skewed}, by TED

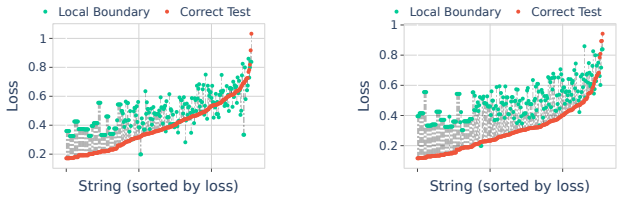
(e) G_4^{uniform}, proficiency (f) G_4^{uniform}, by TED

(g) G_4^{skewed}, proficiency (h) G_4^{skewed}, by TED

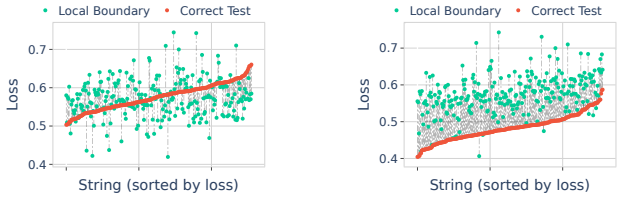
Figure 23: Language proficiency (fraction of strings learned over epochs) and its breakdown by tree-edit distance (TED) to the training set, for G_3 and G_4 (Table 2), aggregated over training seeds. The skew in rule probability between G_i^{uniform} and G_i^{skewed} is what produces the differentiated proficiency we observe across strings. In general, strings with lower edit distance to the training set are learned faster and retain their learned status longer, which is observed more effectively in a skewed language with variable distances.



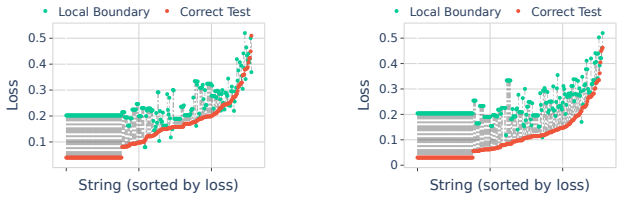
(a) G_1^{uniform}, optimal epoch / 2 (b) G_1^{uniform}, optimal epoch



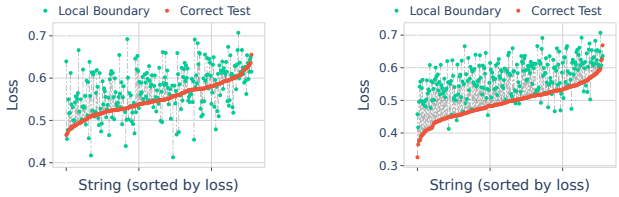
(c) G_1^{skewed}, optimal epoch / 2 (d) G_1^{skewed}, optimal epoch



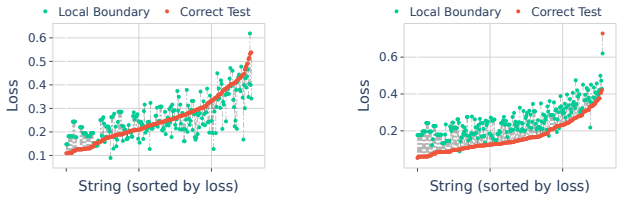
(e) G_2^{uniform}, optimal epoch / 2 (f) G_2^{uniform}, optimal epoch



(g) G_2^{skewed}, optimal epoch / 2 (h) G_2^{skewed}, optimal epoch



(i) G_3^{uniform}, optimal epoch / 2 (j) G_3^{uniform}, optimal epoch



(k) G_3^{skewed}, optimal epoch / 2 (l) G_3^{skewed}, optimal epoch

Figure 24: Loss of a string and its local boundary, at half the optimal epoch and at the optimal epoch, for each grammar (Table 2), fixing the training seed to 1.

D.2 PROFICIENCY TEST COMPARISONS

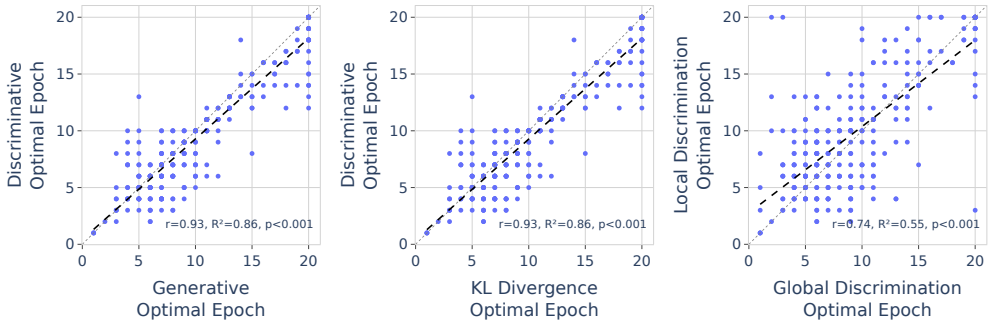
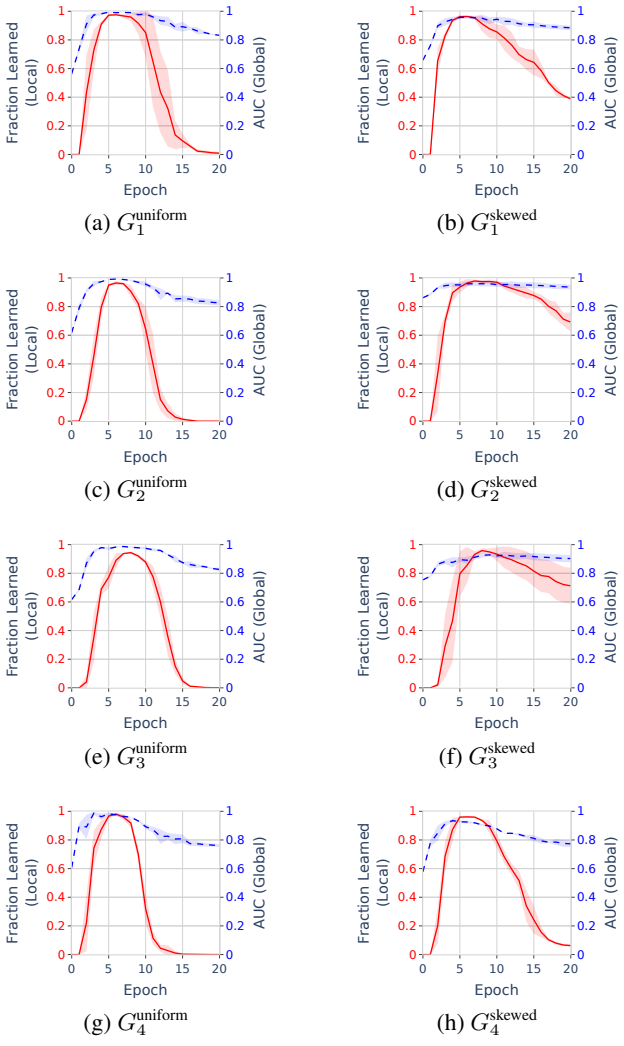


Figure 25: **Optimal local discrimination can be reached earlier than optimal generation.** Across 6 LLMs, 12 languages, and 4 training sizes, we compare the optimal learning epoch by discrimination vs. generation (left). Optimal epochs in both cases correlate strongly, yet a consistent trend emerges: discrimination reaches its optimum earlier than generation, especially when the optimal generation epoch is delayed. Practically, training toward the lowest generation loss (the standard early-stopping criterion) can continue well past the point at which discrimination-based proficiency has already peaked. *Put differently, training can stop earlier when the stopping criterion is discrimination-based rather than generation-based.*

When considering KL divergence instead of generation loss (middle), we observe a similar pattern. When comparing local and global discrimination (right), local discrimination does not consistently precede global discrimination: it lags behind global discrimination at earlier epochs and only catches up at later epochs.



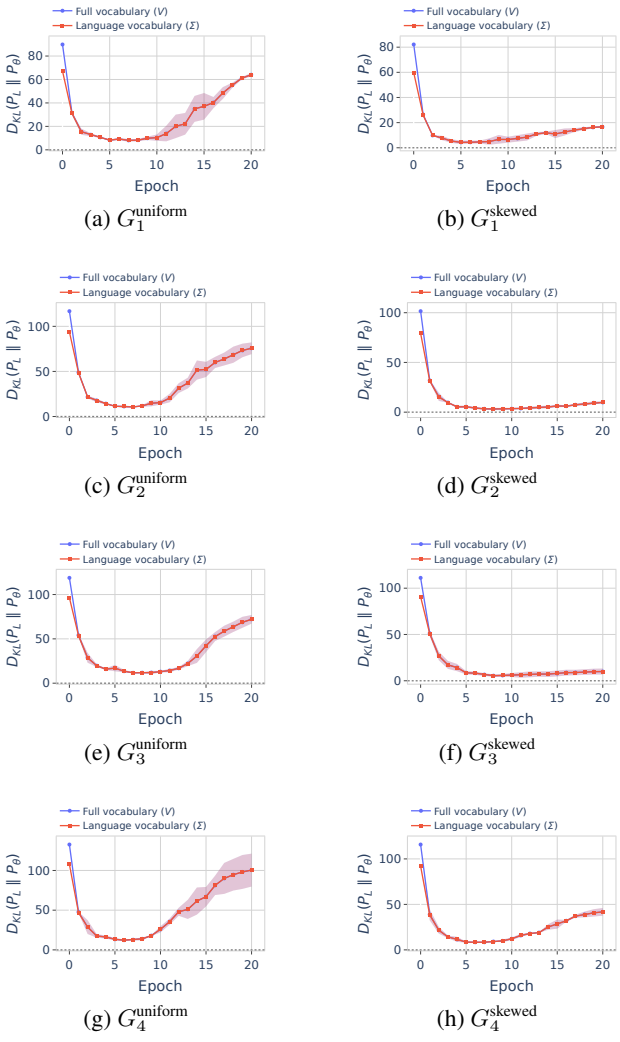
(a) G_1^{uniform} (b) G_1^{skewed}

(c) G_2^{uniform} (d) G_2^{skewed}

(e) G_3^{uniform} (f) G_3^{skewed}

(g) G_4^{uniform} (h) G_4^{skewed}

Figure 26: Language proficiency based on local discrimination and global discrimination (Ghosh et al., 2026), aggregated over training seeds. Global discrimination can potentially over-estimate learning relative to local discrimination.



(a) G_1^{uniform} (b) G_1^{skewed}

(c) G_2^{uniform} (d) G_2^{skewed}

(e) G_3^{uniform} (f) G_3^{skewed}

(g) G_4^{uniform} (h) G_4^{skewed}

Figure 27: KL divergence computed over the full LLM vocabulary ($\boldsymbol{V}$) vs. restricted to the language vocabulary ($\boldsymbol{\Sigma}$). The two diverge early in training, but converge by the final phase, once the LLM's probability mass concentrates on the language vocabulary.

D.3 Rule-Level Proficiency Analysis



(a) Grammatically correct



(b) Grammatically incorrect

Figure 28: A string from G_1^{uniform}, derived using rule $A11 \rightarrow 6\ 5$ (level 1), and its grammatically incorrect counterpart, obtained by editing this rule.

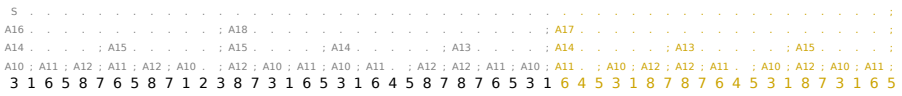


(a) Grammatically correct

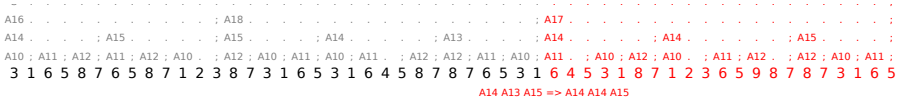


(b) Grammatically incorrect

Figure 29: A string from G_1^{uniform}, derived using rule $A15 \rightarrow A11\ A12\ A10$ (level 2), and its grammatically incorrect counterpart, obtained by editing this rule.



(a) Grammatically correct



(b) Grammatically incorrect

Figure 30: A string from G_1^{uniform}, derived using rule $A17 \rightarrow A14\ A13\ A15$ (level 3), and its grammatically incorrect counterpart, obtained by editing this rule.

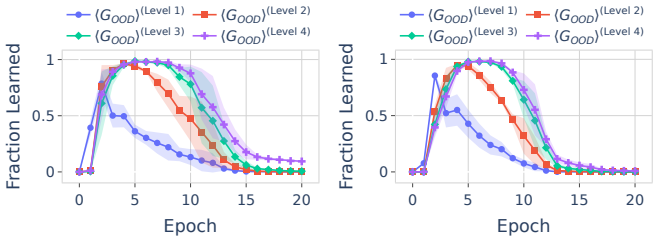


(a) Grammatically correct



(b) Grammatically incorrect

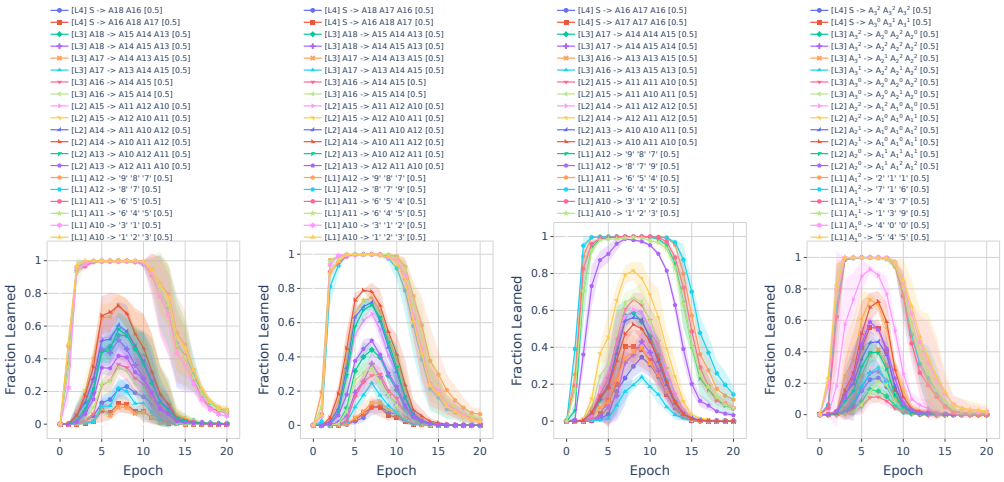
Figure 31: A string from G_1^{uniform}, derived using rule $S \to A18\ A16$ (level 4), and its grammatically incorrect counterpart, obtained by editing this rule.



(a) G_1^{uniform} (b) G_2^{uniform}

(c) G_3^{uniform} (d) G_4^{uniform}

Figure 32: Language proficiency across out-of-distribution languages induced by $\{G_{\text{OOD}}\}^{(\text{Level } \ell)}$, the family of grammars each editing one rule of G at level ℓ, across all four grammars (Table 2). Proficiency is better when editing higher-level rules, i.e., relaxing the global structure of strings in G. Variance is w.r.t. three seeds of the training dataset.

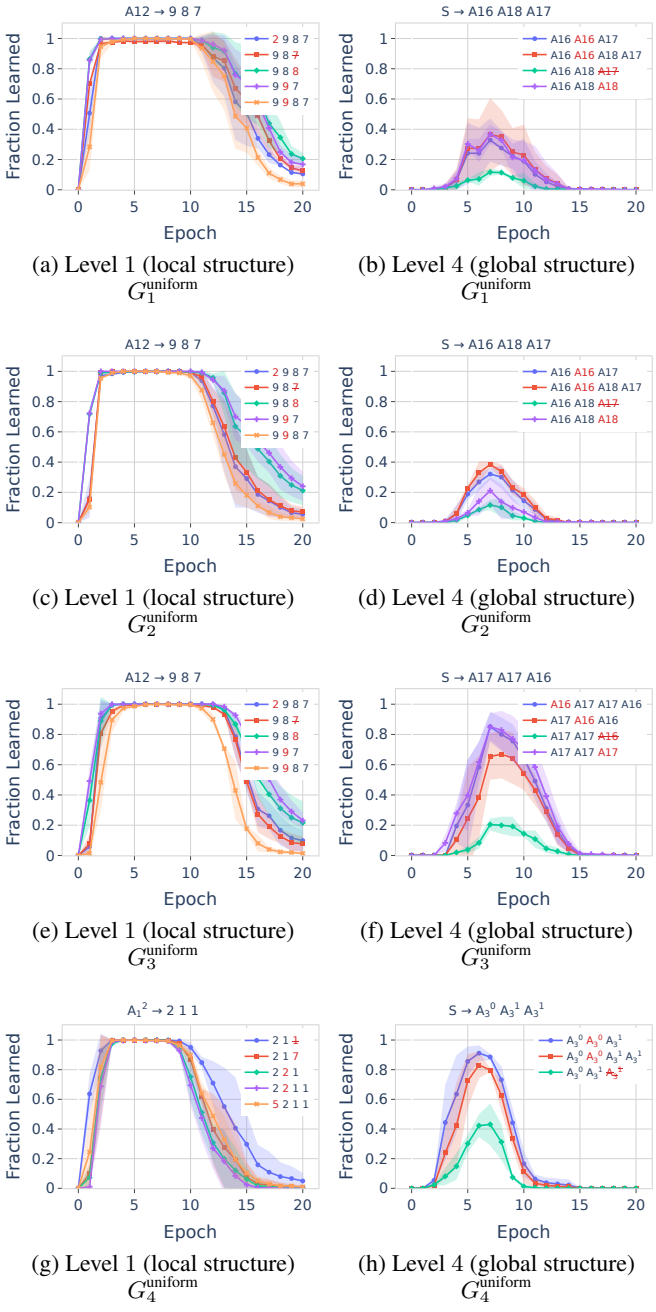


(a) G_1^{uniform} (b) G_2^{uniform} (c) G_3^{uniform} (d) G_4^{uniform}

Figure 33: **Rule Proficiency Measure.** Production rules in a formal grammar capture the structure of a language. To measure if a rule is learned, we focus on strings derived using the rule, edit the strings in positions where the rule is used, and apply the discriminative test defined in Section 4.

Formally, given a rule $R \in \boldsymbol{R}$, let L_R be the set of strings derived using R. Given a string $s \in L_R$, let $\mathcal{N}_k^R(s)$ be the set of neighboring grammatically incorrect strings obtained by editing R at most k times from s. As with string proficiency, we define its *rule-aware* local boundary as $\texttt{boundary-loss}_k^R(s) \triangleq \min_{s_{\text{incorrect}} \in \mathcal{N}_k^R(s)} \min_\theta \texttt{loss}_\theta(s_{\text{incorrect}})$. Now, s is learned in the context of applying R if it is a local minimum, i.e., $\texttt{loss}_\theta(s) < \texttt{boundary-loss}_k^R(s)$. The rule proficiency of R is then the expected fraction of strings in L_R that are learned, $\mathbb{E}_{s \sim P_{L_R}}\left[\mathbb{1}_\theta[s \text{ is learned}]\right]$, where P_{L_R} is the distribution over strings in P_L restricted and renormalized to L_R.

The figures show the rule proficiency of different rules.



(a) Level 1 (local structure) G_1^{uniform}

(b) Level 4 (global structure) G_1^{uniform}

(c) Level 1 (local structure) G_2^{uniform}

(d) Level 4 (global structure) G_2^{uniform}

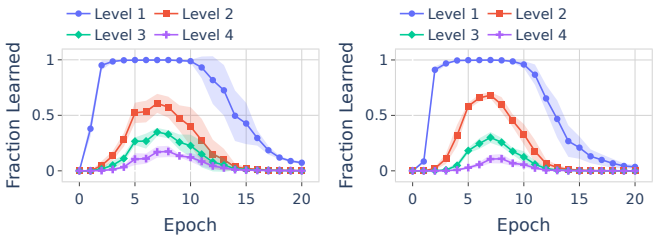
(e) Level 1 (local structure) G_3^{uniform}

(f) Level 4 (global structure) G_3^{uniform}

(g) Level 1 (local structure) G_4^{uniform}

(h) Level 4 (global structure) G_4^{uniform}

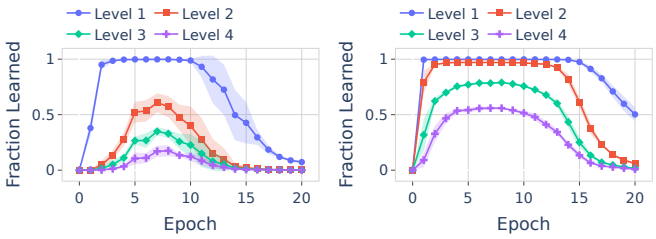
Figure 34: Language proficiency when we edit a rule (particularly, the RHS symbols) and attempt to discriminate strings generated using the edited rule, across all four grammars (Table 2). Red color denotes the edit, where a terminal or non-terminal in the RHS is edited within the same level.



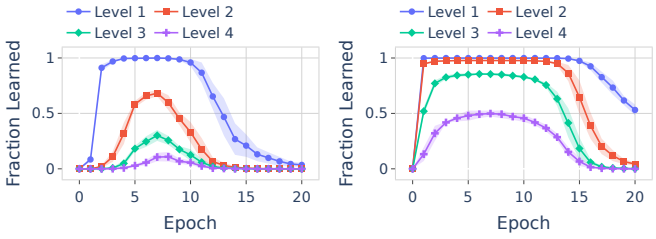
(a) G_1^{uniform} (b) G_2^{uniform}

(c) G_3^{uniform} (d) G_4^{uniform}

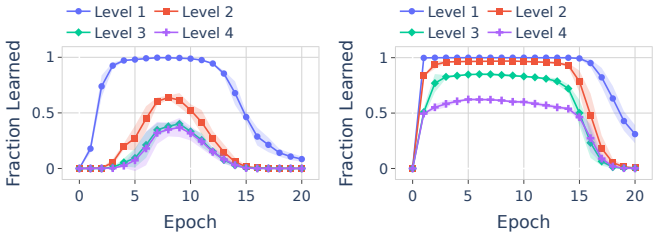
Figure 35: Rule proficiency across rules by levels: A lower level rule captures the local structure while a higher level rule captures the global structure.



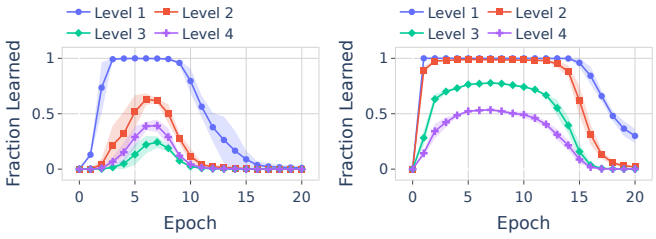
(a) G_1^{uniform}, $n = 64$ (b) G_1^{uniform}, $n = 1024$



(c) G_2^{uniform}, $n = 64$ (d) G_2^{uniform}, $n = 1024$

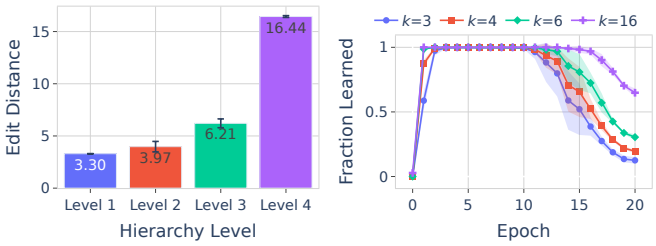


(e) G_3^{uniform}, $n = 64$ (f) G_3^{uniform}, $n = 1024$

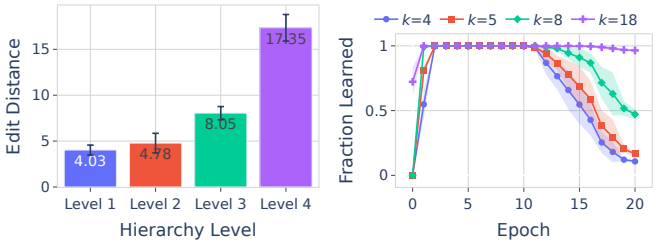


(g) G_4^{uniform}, $n = 64$ (h) G_4^{uniform}, $n = 1024$

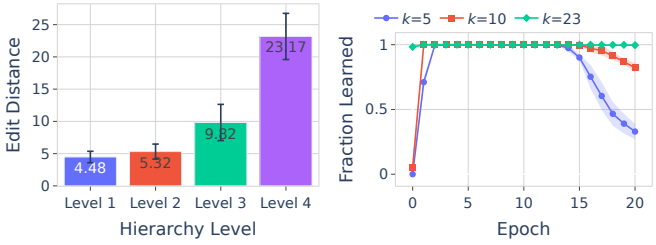
Figure 36: Impact of training size: rule proficiency improves with more training strings (n), across grammars and levels.



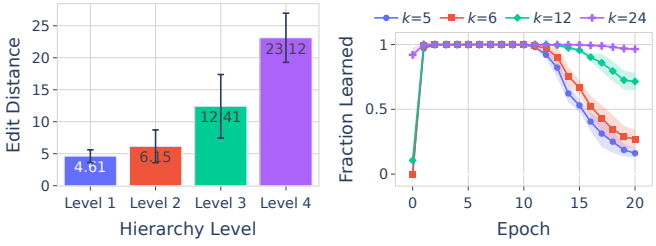
(a) Edits on resulting strings G_1^{uniform} (b) Edits at random positions G_1^{uniform}



(c) Edits on resulting strings G_2^{uniform} (d) Edits at random positions G_2^{uniform}

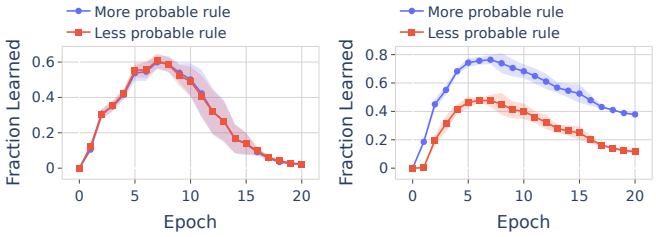


(e) Edits on resulting strings G_3^{uniform} (f) Edits at random positions G_3^{uniform}

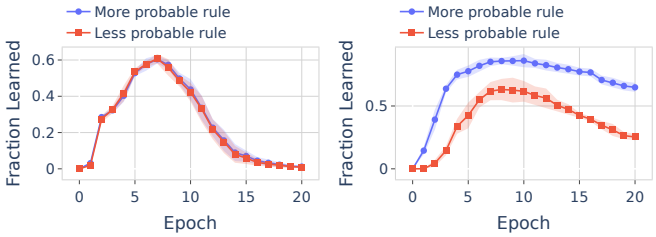


(g) Edits on resulting strings G_4^{uniform} (h) Edits at random positions G_4^{uniform}

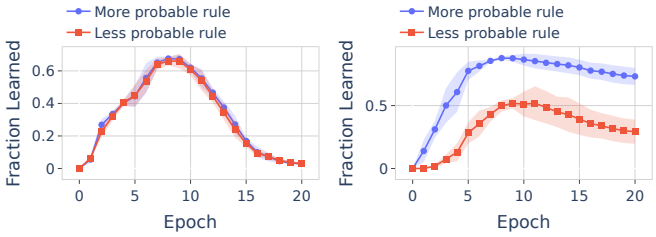
Figure 37: (a) Number of edits on resulting strings when equally editing rules at different levels. (b) When a matching number of edits is applied to strings at random positions instead, the fraction of learned strings increases with edit distance, the reverse of the trend in Figure 8 (main text). Shown across all four grammars (Table 2).



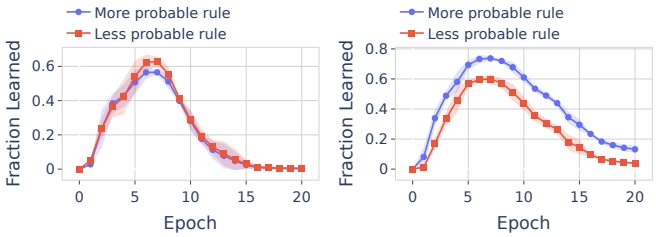
(a) G_1^{uniform} (b) G_1^{skewed}



(c) G_2^{uniform} (d) G_2^{skewed}



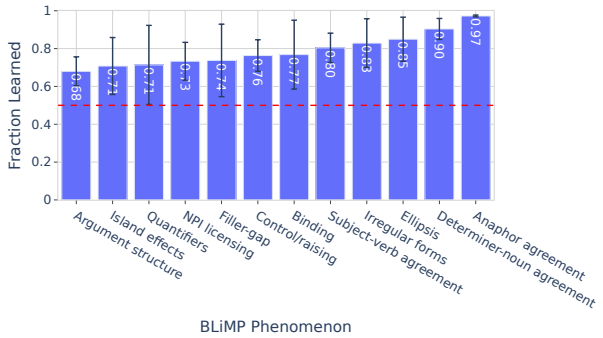
(e) G_3^{uniform} (f) G_3^{skewed}



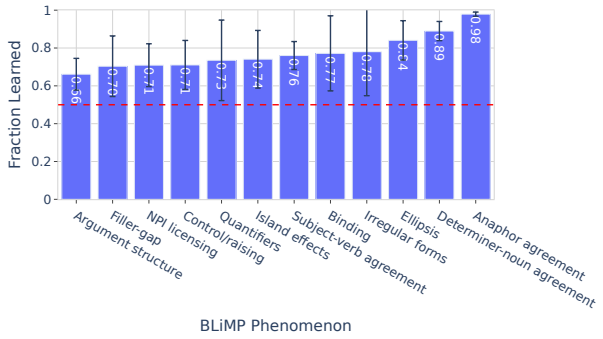
(g) G_4^{uniform} (h) G_4^{skewed}

Figure 38: Rule proficiency, ranked by rule probability. In the skewed language, where some rules are more likely than others, this produces a significant proficiency difference across rules.

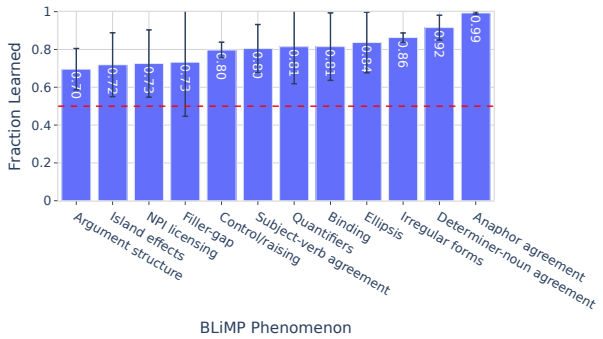
D.4 NATURAL LANGUAGE



(a) Qwen2.5-1.5B-Instruct



(b) Qwen2.5-7B-Instruct

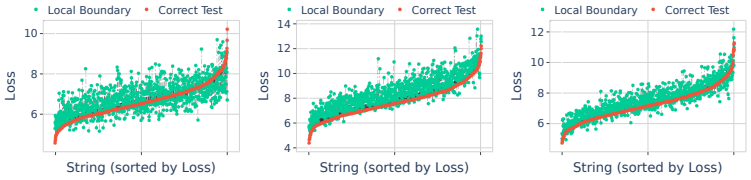


(c) Llama-3.1-8B-Instruct

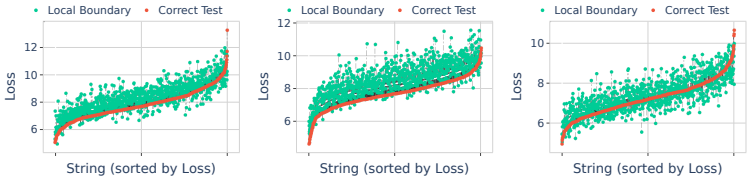
Figure 39: BLiMP (Warstadt et al., 2020) is a benchmark of linguistic minimal pairs for English: 67 datasets grouped into phenomena (e.g., anaphor agreement, subject-verb agreement), each contrasting a grammatical and an ungrammatical sentence that differ minimally. We show the discrimination rate by phenomenon, across models, aggregated over the datasets within each phenomenon; the spread visible here reflects variance across those datasets.



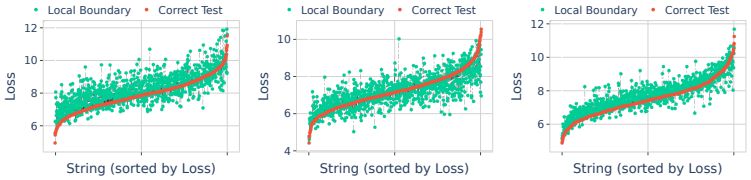
(a) Anaphor agreement (b) Argument structure (c) Binding (principle A)



(d) Control/raising (e) Determiner-noun agreement (f) Ellipsis

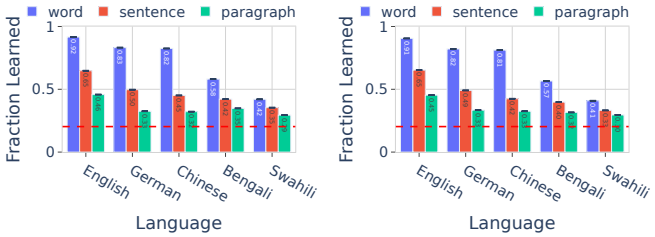


(g) Filler-gap (h) Irregular forms (i) Island effects



(j) NPI licensing (k) Quantifiers (l) Subject-verb agreement

Figure 40: Loss of grammatical vs. ungrammatical strings, one representative dataset per BLiMP phenomenon. In all figures, generation loss cannot reliably predict whether a string is learned or not based on local discrimination.



(a) Qwen2.5-1.5B (b) Qwen2.5-1.5B-Instruct

(c) Qwen2.5-7B-Instruct (d) Llama-3.1-8B-Instruct

Figure 41: Discrimination rate by word/sentence/paragraph edit in natural language, across models. Text is sampled from FineWeb-Edu (Lozhkov et al., 2024) (English) and FineWeb2 (Penedo et al., 2025) (German, Chinese, Bengali, Swahili). Analogous to Figure 8b in the main text, which shows only Qwen2.5-1.5B.

A related asymmetry has been observed by Khandelwal et al. (2018), who find that neural language models track nearby context precisely but represent more distant context only coarsely, ignoring word order beyond a few dozen tokens. Our test, however, fundamentally differs: rather than controlling context distance from the generated token, we randomly edit a word, sentence, or paragraph in the context and check whether the model assigns the edited, incoherent version lower probability than the original. Next, we discuss the data generation process in detail.

DATA PREPARATION: ENGLISH

We construct edits at three levels of granularity for the natural-language discrimination experiment shown in Figure 8b (main text) and Figure 41 (Appendix): **word**, **sentence**, and **paragraph**. At each level, we consider four operations: **replace**, **insert**, **delete**, and **swap**. Text is sampled from FineWeb-Edu (Lozhkov et al., 2024). Below, we show one example per operation for each level; red marks the edited span before the edit and green marks it after.

LEVEL: WORD

english · level=**word** · op=**replace** · location: paragraph 3, sentence 1, word 0

replace: "The" → "No."

BEFORE

Gender and Climate Change in India

Six Villages, Andhra Pradesh, India

The focus of the study are six villages with high indicators of poverty and livelihoods reliant on rainfed agriculture in the drought-prone districts of Anantapur and Mahabubnagar.

Two hundred and one farmers (106 women and 95 men) participated in the quantitative survey, while approximately 180 farmers participated in the focus group discussions. The men and women farmers have land holdings of 2.5 – 5 acres and are above 35 years old; the large majority of land owners are men. The study participants rely primarily on agriculture for their livelihoods and food security, but increasingly seasonal migration, loans, and government-funded programmes are adopted as livelihood strategies and to ensure food security. The major crops grown include paddy, ground nut, red gram, castor, jowar, sesame and green gram.

© E. L. F. Schipper

AFTER

Gender and Climate Change in India

Six Villages, Andhra Pradesh, India

The focus of the study are six villages with high indicators of poverty and livelihoods reliant on rainfed agriculture in the drought-prone districts of Anantapur and Mahabubnagar.

Two hundred and one farmers (106 women and 95 men) participated in the quantitative survey, while approximately 180 farmers participated in the focus group discussions. No. men and women farmers have land holdings of 2.5 – 5 acres and are above 35 years old; the large majority of land owners are men. The study participants rely primarily on agriculture for their livelihoods and food security, but increasingly seasonal migration, loans, and government-funded programmes are adopted as livelihood strategies and to ensure food security. The major crops grown include paddy, ground nut, red gram, castor, jowar, sesame and green gram.

© E. L. F. Schipper

english · level=**word** · op=**insert** · location: paragraph 3, sentence 2, word 31

insert: "willfully."

BEFORE

Gender and Climate Change in India

Six Villages, Andhra Pradesh, India

The focus of the study are six villages with high indicators of poverty and livelihoods reliant on rainfed agriculture in the drought-prone districts of Anantapur and Mahabubnagar.

Two hundred and one farmers (106 women and 95 men) participated in the quantitative survey, while approximately 180 farmers participated in the focus group discussions. The men and women farmers have land holdings of 2.5 – 5 acres and are above 35 years old; the large majority of land owners are men. The study participants rely primarily on agriculture for their livelihoods and food security, but increasingly seasonal migration, loans, and government-funded programmes are adopted as livelihood strategies and to ensure food security. The major crops grown include paddy, ground nut, red gram, castor, jowar, sesame and green gram.

© E. L. F. Schipper

AFTER

Gender and Climate Change in India

Six Villages, Andhra Pradesh, India

The focus of the study are six villages with high indicators of poverty and livelihoods reliant on rainfed agriculture in the drought-prone districts of Anantapur and Mahabubnagar.

Two hundred and one farmers (106 women and 95 men) participated in the quantitative survey, while approximately 180 farmers participated in the focus group discussions. The men and women farmers have land holdings of 2.5 – 5 acres and are above 35 years old; the large majority of land owners are men. The study participants rely primarily on agriculture for their livelihoods and food security, but increasingly seasonal migration, loans, and government-funded programmes are adopted as livelihood strategies and to ensure food security. willfully. The major crops grown include paddy, ground nut, red gram, castor, jowar, sesame and green gram.

© E. L. F. Schipper

english · level=**word** · op=**delete** · location: paragraph 1, sentence 0, word 4

delete: "India"

BEFORE

Gender and Climate Change in India

Six Villages, Andhra Pradesh, India

The focus of the study are six villages with high indicators of poverty and livelihoods reliant on rainfed agriculture in the drought-prone districts of Anantapur and Mahabubnagar.

Two hundred and one farmers (106 women and 95 men) participated in the quantitative survey, while approximately 180 farmers participated in the focus group discussions. The men and women farmers have land holdings of 2.5 – 5 acres and are above 35 years old; the large majority of land owners are men. The study participants rely primarily on agriculture for their livelihoods and food security, but increasingly seasonal migration, loans, and government-funded programmes are adopted as livelihood strategies and to ensure food security. The major crops grown include paddy, ground nut, red gram, castor, jowar, sesame and green gram.

© E. L. F. Schipper

AFTER

Gender and Climate Change in India

Six Villages, Andhra Pradesh,

The focus of the study are six villages with high indicators of poverty and livelihoods reliant on rainfed agriculture in the drought-prone districts of Anantapur and Mahabubnagar.

Two hundred and one farmers (106 women and 95 men) participated in the quantitative survey, while approximately 180 farmers participated in the focus group discussions. The men and women farmers have land holdings of 2.5 – 5 acres and are above 35 years old; the large majority of land owners are men. The study participants rely primarily on agriculture for their livelihoods and food security, but increasingly seasonal migration, loans, and government-funded programmes are adopted as livelihood strategies and to ensure food security. The major crops grown include paddy, ground nut, red gram, castor, jowar, sesame and green gram.

© E. L. F. Schipper

english · level=**word** · op=**swap** · location: paragraph 0, sentence 0, words 4 & 5

swap: "in" ↔ "India"

BEFORE

Gender and Climate Change in India

Six Villages, Andhra Pradesh, India

The focus of the study are six villages with high indicators of poverty and livelihoods reliant on rainfed agriculture in the drought-prone districts of Anantapur and Mahabubnagar.

Two hundred and one farmers (106 women and 95 men) participated in the quantitative survey, while approximately 180 farmers participated in the focus group discussions. The men and women farmers have land holdings of 2.5 – 5 acres and are above 35 years old; the large majority of land owners are men. The study participants rely primarily on agriculture for their livelihoods and food security, but increasingly seasonal migration, loans, and government-funded programmes are adopted as livelihood strategies and to ensure food security. The major crops grown include paddy, ground nut, red gram, castor, jowar, sesame and green gram.

© E. L. F. Schipper

AFTER

Gender and Climate Change India in

Six Villages, Andhra Pradesh, India

The focus of the study are six villages with high indicators of poverty and livelihoods reliant on rainfed agriculture in the drought-prone districts of Anantapur and Mahabubnagar.

Two hundred and one farmers (106 women and 95 men) participated in the quantitative survey, while approximately 180 farmers participated in the focus group discussions. The men and women farmers have land holdings of 2.5 – 5 acres and are above 35 years old; the large majority of land owners are men. The study participants rely primarily on agriculture for their livelihoods and food security, but increasingly seasonal migration, loans, and government-funded programmes are adopted as livelihood strategies and to ensure food security. The major crops grown include paddy, ground nut, red gram, castor, jowar, sesame and green gram.

© E. L. F. Schipper

LEVEL: SENTENCE

english · level=**sentence** · op=**replace** · location: paragraph 4, sentence 3

replace: "Schipper" → "- Physical Activity and Healthy Eating: The Perfect Combination for Weight Management"

BEFORE

Gender and Climate Change in India

Six Villages, Andhra Pradesh, India

The focus of the study are six villages with high indicators of poverty and livelihoods reliant on rainfed agriculture in the drought-prone districts of Anantapur and Mahabubnagar.

Two hundred and one farmers (106 women and 95 men) participated in the quantitative survey, while approximately 180 farmers participated in the focus group discussions. The men and women farmers have land holdings of 2.5 – 5 acres and are above 35 years old; the large majority of land owners are men. The study participants rely primarily on agriculture for their livelihoods and food security, but increasingly seasonal migration, loans, and government-funded programmes are adopted as livelihood strategies and to ensure food security. The major crops grown include paddy, ground nut, red gram, castor, jowar, sesame and green gram.

© E. L. F. Schipper

AFTER

Gender and Climate Change in India

Six Villages, Andhra Pradesh, India

The focus of the study are six villages with high indicators of poverty and livelihoods reliant on rainfed agriculture in the drought-prone districts of Anantapur and Mahabubnagar.

Two hundred and one farmers (106 women and 95 men) participated in the quantitative survey, while approximately 180 farmers participated in the focus group discussions. The men and women farmers have land holdings of 2.5 – 5 acres and are above 35 years old; the large majority of land owners are men. The study participants rely primarily on agriculture for their livelihoods and food security, but increasingly seasonal migration, loans, and government-funded programmes are adopted as livelihood strategies and to ensure food security. The major crops grown include paddy, ground nut, red gram, castor, jowar, sesame and green gram.

© E. L. F. - Physical Activity and Healthy Eating: The Perfect Combination for Weight Management

english · level=**sentence** · op=**insert** · location: paragraph 1, sentence 1

insert: "For example, higher gasoline prices-say $5.00 per gallon- would drive more people to hybrid and smaller more fuel efficient cars purchases than any environmental benefit factor."

BEFORE

Gender and Climate Change in India

Six Villages, Andhra Pradesh, India

The focus of the study are six villages with high indicators of poverty and livelihoods reliant on rainfed agriculture in the drought-prone districts of Anantapur and Mahabubnagar.

Two hundred and one farmers (106 women and 95 men) participated in the quantitative survey, while approximately 180 farmers participated in the focus group discussions. The men and women farmers have land holdings of 2.5 – 5 acres and are above 35 years old; the large majority of land owners are men. The study participants rely primarily on agriculture for their livelihoods and food security, but increasingly seasonal migration, loans, and government-funded programmes are adopted as livelihood strategies and to ensure food security. The major crops grown include paddy, ground nut, red gram, castor, jowar, sesame and green gram.

© E. L. F. Schipper

AFTER

Gender and Climate Change in India

Six Villages, Andhra Pradesh, India For example, higher gasoline prices-say $5.00 per gallon- would drive more people to hybrid and smaller more fuel efficient cars purchases than any environmental benefit factor.

The focus of the study are six villages with high indicators of poverty and livelihoods reliant on rainfed agriculture in the drought-prone districts of Anantapur and Mahabubnagar.

Two hundred and one farmers (106 women and 95 men) participated in the quantitative survey, while approximately 180 farmers participated in the focus group discussions. The men and women farmers have land holdings of 2.5 – 5 acres and are above 35 years old; the large majority of land owners are men. The study participants rely primarily on agriculture for their livelihoods and food security, but increasingly seasonal migration, loans, and government-funded programmes are adopted as livelihood strategies and to ensure food security. The major crops grown include paddy, ground nut, red gram, castor, jowar, sesame and green gram.

© E. L. F. Schipper

english · level=**sentence** · op=**delete** · location: paragraph 3, sentence 1

delete: "The men and women farmers have land holdings of 2.5 – 5 acres and are above 35 years old; the large majority of land owners are men."

BEFORE

Gender and Climate Change in India

Six Villages, Andhra Pradesh, India

The focus of the study are six villages with high indicators of poverty and livelihoods reliant on rainfed agriculture in the drought-prone districts of Anantapur and Mahabubnagar.

Two hundred and one farmers (106 women and 95 men) participated in the quantitative survey, while approximately 180 farmers participated in the focus group discussions. The men and women farmers have land holdings of 2.5 – 5 acres and are above 35 years old; the large majority of land owners are men. The study participants rely primarily on agriculture for their livelihoods and food security, but increasingly seasonal migration, loans, and government-funded programmes are adopted as livelihood strategies and to ensure food security. The major crops grown include paddy, ground nut, red gram, castor, jowar, sesame and green gram.

© E. L. F. Schipper

AFTER

Gender and Climate Change in India

Six Villages, Andhra Pradesh, India

The focus of the study are six villages with high indicators of poverty and livelihoods reliant on rainfed agriculture in the drought-prone districts of Anantapur and Mahabubnagar.

Two hundred and one farmers (106 women and 95 men) participated in the quantitative survey, while approximately 180 farmers participated in the focus group discussions. The study participants rely primarily on agriculture for their livelihoods and food security, but increasingly seasonal migration, loans, and government-funded programmes are adopted as livelihood strategies and to ensure food security. The major crops grown include paddy, ground nut, red gram, castor, jowar, sesame and green gram.

© E. L. F. Schipper

english · level=**sentence** · op=**swap** · location: paragraph.sentence 1.0 & 2.0

swap: "Six Villages, Andhra Pradesh, India" ↔ "The focus of the study are six villages with high indicators of poverty and livelihoods reliant on rainfed agriculture in the drought-prone districts of Anantapur and Mahabubnagar."

BEFORE

Gender and Climate Change in India

Six Villages, Andhra Pradesh, India

The focus of the study are six villages with high indicators of poverty and livelihoods reliant on rainfed agriculture in the drought-prone districts of Anantapur and Mahabubnagar.

Two hundred and one farmers (106 women and 95 men) participated in the quantitative survey, while approximately 180 farmers participated in the focus group discussions. The men and women farmers have land holdings of 2.5 – 5 acres and are above 35 years old; the large majority of land owners are men. The study participants rely primarily on agriculture for their livelihoods and food security, but

increasingly seasonal migration, loans, and government-funded programmes are adopted as livelihood strategies and to ensure food security. The major crops grown include paddy, ground nut, red gram, castor, jowar, sesame and green gram.

© E. L. F. Schipper

AFTER

Gender and Climate Change in India

The focus of the study are six villages with high indicators of poverty and livelihoods reliant on rainfed agriculture in the drought-prone districts of Anantapur and Mahabubnagar.

Six Villages, Andhra Pradesh, India

Two hundred and one farmers (106 women and 95 men) participated in the quantitative survey, while approximately 180 farmers participated in the focus group discussions. The men and women farmers have land holdings of 2.5 – 5 acres and are above 35 years old; the large majority of land owners are men. The study participants rely primarily on agriculture for their livelihoods and food security, but increasingly seasonal migration, loans, and government-funded programmes are adopted as livelihood strategies and to ensure food security. The major crops grown include paddy, ground nut, red gram, castor, jowar, sesame and green gram.

© E. L. F. Schipper

LEVEL: PARAGRAPH

english · level=**paragraph** · op=**replace** · location: paragraph 2

replace: "The focus of the study are six villages with high indicators of poverty and livelihoods reliant on rainfed agriculture in the drought-prone districts of Anantapur and Mahabubnagar." → "What Is Bad Breath Caused By?"

BEFORE

Gender and Climate Change in India

Six Villages, Andhra Pradesh, India

The focus of the study are six villages with high indicators of poverty and livelihoods reliant on rainfed agriculture in the drought-prone districts of Anantapur and Mahabubnagar.

Two hundred and one farmers (106 women and 95 men) participated in the quantitative survey, while approximately 180 farmers participated in the focus group discussions. The men and women farmers have land holdings of 2.5 – 5 acres and are above 35 years old; the large majority of land owners are men. The study participants rely primarily on agriculture for their livelihoods and food security, but increasingly seasonal migration, loans, and government-funded programmes are adopted as livelihood strategies and to ensure food security. The major crops grown include paddy, ground nut, red gram, castor, jowar, sesame and green gram.

© E. L. F. Schipper

AFTER

Gender and Climate Change in India

Six Villages, Andhra Pradesh, India

What Is Bad Breath Caused By?

Two hundred and one farmers (106 women and 95 men) participated in the quantitative survey, while approximately 180 farmers participated in the focus group discussions. The men and women farmers have land holdings of 2.5 – 5 acres and are above 35 years old; the large majority of land owners are men. The study participants rely primarily on agriculture for their livelihoods and food security, but increasingly seasonal migration, loans, and government-funded programmes are adopted as livelihood strategies and to ensure food security. The major crops grown include paddy, ground nut, red gram, castor, jowar, sesame and green gram.

© E. L. F. Schipper

english · level=**paragraph** · op=**insert** · location: paragraph 1

insert: "The new planet — named Kepler-22b — has key aspects it shares with Earth. It circles a star that could be the twin of our sun and at just about the same distance. The planet's year of 290 days is even close to ours. It likely has water and rock."

BEFORE

Gender and Climate Change in India

Six Villages, Andhra Pradesh, India

The focus of the study are six villages with high indicators of poverty and livelihoods reliant on rainfed agriculture in the drought-prone districts of Anantapur and Mahabubnagar.

Two hundred and one farmers (106 women and 95 men) participated in the quantitative survey, while approximately 180 farmers participated in the focus group discussions. The men and women farmers have land holdings of 2.5 – 5 acres and are above 35 years old; the large majority of land owners are men. The study participants rely primarily on agriculture for their livelihoods and food security, but increasingly seasonal migration, loans, and government-funded programmes are adopted as livelihood strategies and to ensure food security. The major crops grown include paddy, ground nut, red gram, castor, jowar, sesame and green gram.

© E. L. F. Schipper

AFTER

Gender and Climate Change in India

The new planet — named Kepler-22b — has key aspects it shares with Earth. It circles a star that could be the twin of our sun and at just about the same

distance. The planet's year of 290 days is even close to ours. It likely has water and rock.

Six Villages, Andhra Pradesh, India

The focus of the study are six villages with high indicators of poverty and livelihoods reliant on rainfed agriculture in the drought-prone districts of Anantapur and Mahabubnagar.

Two hundred and one farmers (106 women and 95 men) participated in the quantitative survey, while approximately 180 farmers participated in the focus group discussions. The men and women farmers have land holdings of 2.5 – 5 acres and are above 35 years old; the large majority of land owners are men. The study participants rely primarily on agriculture for their livelihoods and food security, but increasingly seasonal migration, loans, and government-funded programmes are adopted as livelihood strategies and to ensure food security. The major crops grown include paddy, ground nut, red gram, castor, jowar, sesame and green gram.

© E. L. F. Schipper

english · level=**paragraph** · op=**delete** · location: paragraph 3

delete: "Two hundred and one farmers (106 women and 95 men) participated in the quantitative survey, while approximately 180 farmers participated in the focus group discussions. The men and women farmers have land holdings of 2.5 – 5 acres and are above 35 years old; the large majority of land owners are men. The study participants rely primarily on agriculture for their livelihoods and food security, but increasingly seasonal migration, loans, and government-funded programmes are adopted as livelihood strategies and to ensure food security. The major crops grown include paddy, ground nut, red gram, castor, jowar, sesame and green gram."

BEFORE

Gender and Climate Change in India

Six Villages, Andhra Pradesh, India

The focus of the study are six villages with high indicators of poverty and livelihoods reliant on rainfed agriculture in the drought-prone districts of Anantapur and Mahabubnagar.

Two hundred and one farmers (106 women and 95 men) participated in the quantitative survey, while approximately 180 farmers participated in the focus group discussions. The men and women farmers have land holdings of 2.5 – 5 acres and are above 35 years old; the large majority of land owners are men. The study participants rely primarily on agriculture for their livelihoods and food security, but increasingly seasonal migration, loans, and government-funded programmes are adopted as livelihood strategies and to ensure food security. The major crops grown include paddy, ground nut, red gram, castor, jowar, sesame and green gram.

© E. L. F. Schipper

AFTER

Gender and Climate Change in India

Six Villages, Andhra Pradesh, India

The focus of the study are six villages with high indicators of poverty and livelihoods reliant on rainfed agriculture in the drought-prone districts of Anantapur and Mahabubnagar.

© E. L. F. Schipper

english · level=**paragraph** · op=**swap** · location: paragraphs 2 & 3

swap: "The focus of the study are six villages with high indicators of poverty and livelihoods reliant on rainfed agriculture in the drought-prone districts of Anantapur and Mahabubnagar." ↔ "Two hundred and one farmers (106 women and 95 men) participated in the quantitative survey, while approximately 180 farmers participated in the focus group discussions. The men and women farmers have land holdings of 2.5 – 5 acres and are above 35 years old; the large majority of land owners are men. The study participants rely primarily on agriculture for their livelihoods and food security, but increasingly seasonal migration, loans, and government-funded programmes are adopted as livelihood strategies and to ensure

food security. The major crops grown include paddy, ground nut, red gram, castor, jowar, sesame and green gram."

BEFORE

Gender and Climate Change in India

Six Villages, Andhra Pradesh, India

The focus of the study are six villages with high indicators of poverty and livelihoods reliant on rainfed agriculture in the drought-prone districts of Anantapur and Mahabubnagar.

Two hundred and one farmers (106 women and 95 men) participated in the quantitative survey, while approximately 180 farmers participated in the focus group discussions. The men and women farmers have land holdings of 2.5 – 5 acres and are above 35 years old; the large majority of land owners are men. The study participants rely primarily on agriculture for their livelihoods and food security, but increasingly seasonal migration, loans, and government-funded programmes are adopted as livelihood strategies and to ensure food security. The major crops grown include paddy, ground nut, red gram, castor, jowar, sesame and green gram.

© E. L. F. Schipper

AFTER

Gender and Climate Change in India

Six Villages, Andhra Pradesh, India

Two hundred and one farmers (106 women and 95 men) participated in the quantitative survey, while approximately 180 farmers participated in the focus group discussions. The men and women farmers have land holdings of 2.5 – 5 acres and are above 35 years old; the large majority of land owners are men. The study participants rely primarily on agriculture for their livelihoods and food security, but increasingly seasonal migration, loans, and government-funded programmes are adopted as livelihood strategies and to ensure food security. The major crops grown include paddy, ground nut, red gram, castor, jowar, sesame and green gram.

The focus of the study are six villages with high indicators of poverty and livelihoods reliant on rainfed agriculture in the drought-prone districts of Anantapur and Mahabubnagar.

© E. L. F. Schipper