

How Much Sonority Structure Do Phonotactic Language Models Learn from Distribution?

Jia Can Xu Mario Giulianelli

University College London

ucjvjxu@ucl.ac.uk m.giulianelli@ucl.ac.uk

Abstract

Sonority projection, the ability to rank untested consonant clusters by their sonority contour, has been interpreted as evidence for innate phonological knowledge. We test whether neural phonotactic language models can acquire sonority structure from English phoneme distributions through a controlled comparison of four input geometries: articulatory features, trainable embeddings, feature vectors randomly re-assigned to phonemes, and frozen random vectors. We train LSTMs and Transformers, probe whether sonority structure is linearly recoverable across representation spaces and throughout training, and evaluate phonotactic sensitivity using nonce-word violations and human acceptability judgements of unattested onsets. Sonority probes recover the consonant–vowel distinction from contextual phoneme representations under every input condition, whereas the hierarchy among consonant classes (obstruents, nasals, liquids, and glides) is most strongly recoverable with articulatory-feature input and less recoverable under the other conditions. Nevertheless, under all input geometry conditions, model surprisal discriminates phonotactic violations and correlates moderately with human judgements of unattested onsets. In summary, sonority structure and the tested phonotactic preferences can develop without sonority-aligned input representations: input geometry shapes how much fine-grained sonority structure is recoverable, but models with more recoverable structure, whether provided explicitly or learned during training, at the input or in the hidden states, show no consistent advantage on the phonotactic evaluations used here.

1 Introduction

Sonority describes the relative prominence of speech sounds and defines a hierarchical structure among them. Obstruents (e.g., /b/, /t/) are less sonorous than nasals (e.g., /m/, /n/), which are in turn less sonorous than liquids (/l/, /r/), glides (/w/, /j/), and vowels. Sonority plays an important role

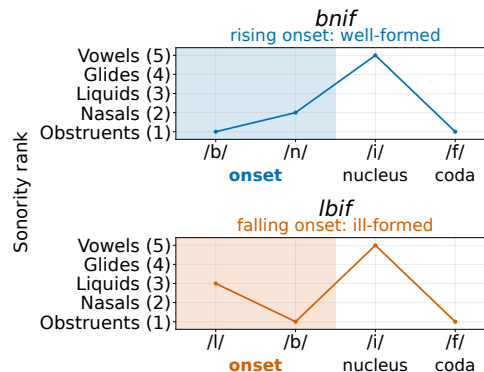


Figure 1: Schematic sonority contours for *bnif* and *lbif*. Both onset clusters are unattested in English, but listeners show a relative preference for the rising onset /bn-/ over the falling /lb-/ (Berent et al., 2007).

in theories of syllable structure. Syllables are typically organised around a nucleus, usually a vowel, preceded by an onset and followed by a coda. The Sonority Sequencing Principle states that sonority generally rises through the onset towards the nucleus and falls through the coda towards the syllable edge (Selkirk, 1984; Clements, 1990). Fig. 1 illustrates the contrast between a rising onset and a falling one. English speakers’ preferences reflect this sonority pattern even for onset clusters that do not occur in English. This generalisation of sonority-based preferences to unattested clusters is known as *sonority projection* (Daland et al., 2011). Because speakers cannot learn these clusters directly from English words, sonority projection has been taken as evidence for innate knowledge of the sonority hierarchy (Berent et al., 2007). This inference assumes that the hierarchy cannot itself be induced from the attested lexicon: that sonority structure must be given to the learner as prior phonological knowledge rather than learned from phoneme distributions alone.

Phonotactic language models make this assumption empirically testable: they can learn phonotactic regularities from distributions of phonemes in lexical data while allowing controlled variation in both learning mechanisms and

phonological representations. Comparing six families of models of phonotactic well-formedness, [Daland et al. \(2011\)](#) showed that lexical statistics alone can explain sonority projection, provided the model has phonological features and the ability to distinguish contexts such as onsets and codas. [Mayer and Nelson \(2020\)](#) subsequently showed that simple RNN phonotactic language models with randomly initialised, trainable embeddings achieved phonotactic performance comparable to models with phonological-feature representations. Their results suggest that sonority structure can be learned from lexical distributions rather than supplied explicitly, but leave open what form this structure takes within the network. In particular, it remains unclear whether sonority must be learned as a static, decontextualised property of phonemes or whether it can also arise in contextual phoneme representations, even when the input geometry remains unaligned with sonority. This paper therefore asks (i) whether, where, and when sonority structure becomes recoverable within the model and (ii) whether sonority-aligned representations improve phonotactic predictions.

We address these questions by training phonotactic language models under four input conditions. In addition to a feature-based model (FEAT) and a trainable-embedding model (EMB), we include two frozen-input controls that prevent the input layer from becoming sonority-aligned during training: the permuted-input condition (PERM), which uses the same articulatory feature vectors as FEAT but randomly reassigns them to phonemes; and the random-input condition (RAND), which uses random vectors. We test all four conditions in LSTMs and Transformers of varying depth, tracking the recoverability of sonority structure at several stages of processing, from the input representations to the fully contextual hidden states, and over the course of training. We then evaluate the models' phonotactic predictions using the violator-control nonce-word pairs of [Hayes and White \(2013\)](#) and the human acceptability judgements of [Daland et al. \(2011\)](#), analysing words with unattested onsets separately to assess sonority projection.

We find that the recoverability of sonority structure depends strongly on the granularity of the sonority distinctions considered. The coarse distinction between consonants and vowels is almost perfectly recoverable from contextual representations under every input condition, already within the first few epochs of training. In contrast, the

finer-grained hierarchy among consonant classes is strongly recoverable only when articulatory features are provided (in-context $\rho = 0.60\text{--}0.79$). Under the other input conditions it is essentially absent from the input layer and emerges only weakly in contextual representations ($\rho \approx 0.4\text{--}0.5$). However, despite these differences in representational structure, models trained under all input conditions distinguish violators from controls and correlate moderately, within a narrow band, with human judgements of words with unattested onsets. These results indicate that, on the phonotactic evaluations used here, sonority-aligned representations confer no consistent advantage in sonority projection, whether they are provided at the input or learned during training, in the embeddings or in the hidden states.

2 Background

2.1 Phonotactic Language Models

A phonotactic language model assigns probabilities to sequences of phonemes. The probability assigned to the phoneme sequence corresponding to a word can be interpreted as the model's estimate of how well the word conforms to the phonotactics of a language. An early probabilistic model of phonotactic acceptability was introduced by [Coleman and Pierrehumbert \(1997\)](#). In their model, words are parsed into prosodic constituents and their probabilities are estimated from lexical frequencies. The maximum-entropy model of [Hayes and Wilson \(2008\)](#) instead learns weighted phonotactic constraints from a lexicon. This model was later used by [Daland et al. \(2011\)](#) to show that sonority projection could be learned from lexical statistics alone, provided that phonological features were available to the model. [Dai et al. \(2023\)](#) further showed that continuous vectors derived from distributional statistics could replace discrete phonological features in a log-bilinear model of English sonority judgements.

Phonotactic language models parameterised by neural networks learn phoneme representations through next-phoneme prediction, and their input representations can be varied while the architecture and prediction task are held fixed. This makes it possible to ask how much phonological structure must be specified in advance for sonority structure to emerge from distributional learning. Previous work has found that explicitly encoded phonological features are not necessary for learning phono-

tactic patterns: using LSTMs, [Mirea and Bicknell \(2019\)](#) found that randomly initialised phoneme embeddings perform at least as well as embeddings initialised from phonological features, and [Mayer and Nelson \(2020\)](#) obtained similar results with simple recurrent networks.

2.2 Learnability and the Origins of Sonority Patterns

Sonority projection has been presented as evidence that phonological knowledge cannot be learned directly from experience. [Berent et al. \(2007\)](#) found that English speakers show graded preferences among unattested onset clusters and argued that, because these clusters are absent from the lexicon, the preferences cannot derive from direct experience with them and instead reflect universal phonological constraints. Neural language models offer a way to test such claims, since they make it possible to ask what a statistical learner acquires from distributional evidence alone, without built-in linguistic knowledge; in syntax, they have been used to test whether attested structures are learned more readily than unattested alternatives ([Kallini et al., 2024](#); [Ahuja et al., 2025](#); [Someya et al., 2025](#)). We ask the analogous question of sonority: what a learner extracts from an English-like lexicon when the sonority hierarchy is not given in advance.

Why that lexicon is sonority-shaped in the first place is a separate, diachronic question: typological regularities may reflect analytic biases in learners, channel biases in production and perception, and their interaction over historical transmission ([Moreton, 2008](#); [Blevins, 2004](#)). We set this question aside by holding the English lexical distribution fixed and varying only the phonological information available to the learner.

3 Training Phonotactic Language Models

We train LSTM and Transformer phonotactic language models using the standard self-supervised next-phoneme prediction objective. Across architectures, we vary the input representations to control whether sonority-related structure is provided explicitly, can be learned at the input layer, or must emerge in deeper representations.

3.1 Phoneme Sequence Data

We use the phoneme sequences compiled by [Daland et al. \(2011\)](#) from the CMU Pronouncing Dictionary of North American English ([Weide,](#)

1998). The dataset contains 18,233 lexical entries (orthography–pronunciation pairs), with sequence lengths ranging from 1 to 17 phonemes. Its vocabulary contains 24 consonants, 36 stress-marked vowels, and three special symbols (PAD, BOS, EOS), for a total of 63 symbols. Using a fixed random seed, we divide the entries into training, validation, and test sets in an 80/10/10 ratio. Training is over lexical entries rather than corpus frequencies: each orthography–pronunciation pair occurs once, so a phoneme sequence shared by several entries (homophones or spelling variants) occurs once for each of them.

3.2 Input Representations

All conditions represent each phoneme with a 21-dimensional vector, the number of articulatory features distributed with the [Daland et al. \(2011\)](#) materials for the stress-marked CMU inventory (excluding the syllable-dependent *rhyme* feature):

FEAT Frozen vectors of articulatory features, which provide a phoneme’s sonority structure directly at the input layer.

EMB Randomly initialised, trainable embeddings, which allow sonority structure to be learned.

PERM The same frozen vectors as in FEAT, but randomly reassigned to phonemes.

RAND Frozen random vectors with no designed linguistic structure.

For each seed, we generate a new reassignment for PERM and a new vector set for RAND, so the results do not depend on one particular permutation or random geometry. The initial vectors for EMB and RAND are sampled from $\mathcal{N}(0, \sigma^2)$, where $\sigma = 0.6541$ matches the element-wise standard deviation of the FEAT matrix and keeps input scale comparable across conditions. For a given seed, RAND freezes exactly the vectors with which EMB is initialised, so the two conditions differ only in whether the phoneme vectors are updated during training. In every condition, the PAD vector is fixed at zero and masked in the loss, the BOS vector is trainable, and EOS is predicted but never presented as input. These conditions distinguish sonority structure supplied at the input layer from structure learned in the inputs or hidden states.

3.3 Model Architectures

For each input condition, we train three architectures: a single-layer LSTM and decoder-only Transformers with two or four layers. The LSTM has 128 hidden units and a linear output layer, with no

dropout or weight tying. The Transformers have a model dimension of 128, four attention heads, a feed-forward dimension of 256, learned positional embeddings initialised to zero, and no dropout. Before the Transformer layers, a learned linear projection maps the 21-dimensional input vectors to the model dimension of 128, matching the size of the LSTM hidden states.

3.4 Model Training

For every combination of architecture and input condition, we train twenty models with different random seeds. Each model is trained for 40 epochs by minimising cross-entropy loss using the Adam optimiser, a learning rate of 10^{-3} , a batch size of 64, gradient clipping at 1.0, and no weight decay.

We save the untrained model and a checkpoint after each of the 40 epochs. Developmental analyses use the checkpoints at initialisation, epochs 1–10, and every fifth epoch thereafter; analyses of trained models use each seed’s checkpoint with the lowest validation loss. Best validation loss falls within a narrow range across input conditions: 2.13–2.20 for the LSTMs and 2.14–2.18 for the Transformers. Within the LSTMs, however, every PERM run has a higher best-validation loss than any run in the other three conditions (Tab. 2). This separation is absent in the Transformers, where the PERM loss ranges overlap substantially with those of EMB and RAND. The best-validation checkpoints fall at epochs 22–37 for the LSTMs, after which validation loss largely plateaus, and at epochs 7–17 for the Transformers, after which it increases.

4 Tracing the Representation of Sonority Structure

We study three representation spaces, at model initialisation and throughout training. The first is the **input representation** of each phoneme, which is frozen in FEAT, PERM, and RAND and learned in EMB. The second is the **BOS-conditioned hidden representation**, obtained by presenting each phoneme immediately after BOS and extracting its hidden state. This gives a minimal-context representation, intermediate between the static input geometry and fully contextualised representations. The third is the **in-context hidden representation**: for each phoneme type, we extract its hidden states from all occurrences in the training data and average them across occurrences, yielding one context-averaged vector per phoneme type.

4.1 Sonority Probes

We use linear probes to test whether sonority structure is recoverable from any of the three model representation spaces. Following Clements (1990), we assign each phoneme to one of five sonority ranks, from obstruents (1) to vowels (5). The full ranks are shown in Tab. 1, App. A.

We apply the probes to the initial model, the best-validation checkpoint, and the developmental checkpoints defined in §3.4. At each checkpoint, we fit a ridge regression probe ($\alpha = 1.0$) to predict sonority rank from the vector associated with each phoneme type. We evaluate the predictions using leave-one-out cross-validation and Spearman correlation; scores in the text are twenty-seed means. The maximum attainable correlation is 0.87 for FULL60, 0.85 for CV-BINARY, and 0.80 for CONSONANTS.¹ We therefore compare raw correlations across conditions within each probe, and only qualitatively across probes.

We use three probe target configurations to distinguish different components of sonority structure. The first probe, **FULL60**, covers all 60 phonemes and predicts ranks 1–5. Because 36 of these phonemes are vowels with the highest rank, its performance could be driven largely by the consonant–vowel distinction. We therefore also include **CV-BINARY**, which assigns a binary consonant–vowel target to all 60 phonemes, and **CONSONANTS**, which predicts ranks 1–4 for the 24 consonants (obstruent < nasal < liquid < glide).

Finally, because the representations are high-dimensional relative to the number of phoneme types, probes can score above zero even on arbitrary labels. Following the control-task logic of Hewitt and Liang (2019), we therefore fit shuffled-label controls, permuting the target labels across phoneme types twenty times, and require the sonority probes to exceed this baseline.

4.1.1 What Sonority Structure Is Recoverable?

We compare all three architectures at each model’s best-validation checkpoint (Fig. 2). At the input layer, the consonant–vowel distinction is strongly—and trivially—recoverable under FEAT in all architectures; under EMB, it is strongly recoverable only in the LSTM; while under PERM and RAND it is es-

¹Target sonority ranks contain ties because phonemes in the same sonority group share a rank. Therefore, when all predicted scores are distinct, even perfect ordering of the target groups yields $\rho < 1$.

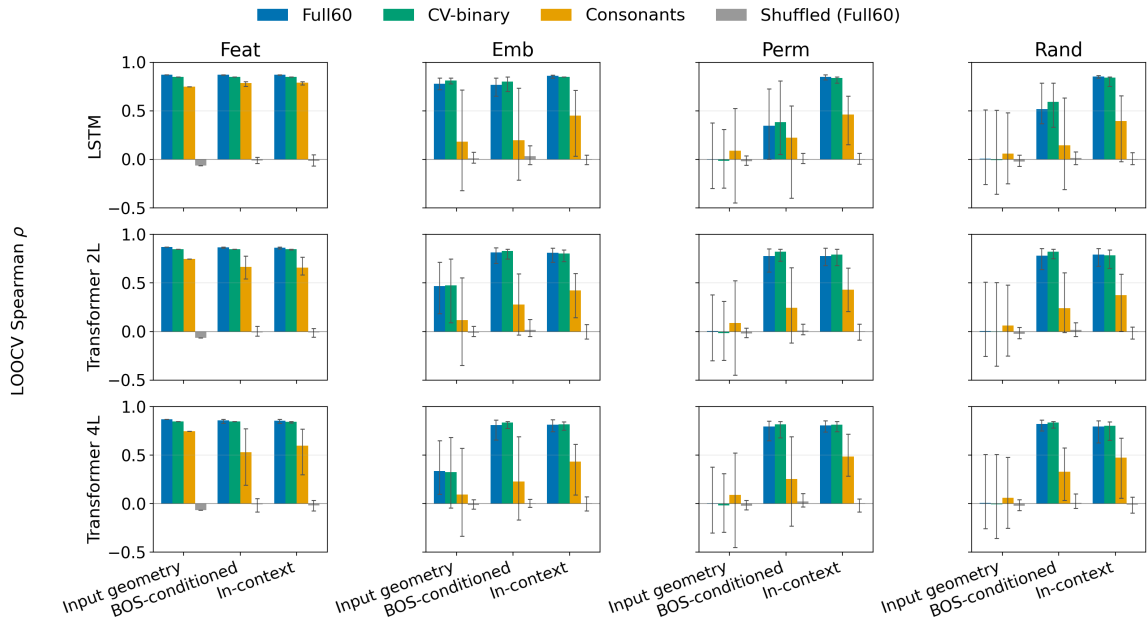


Figure 2: Sonority probe decomposition at each model’s best-validation checkpoint. Bars: twenty-seed mean leave-one-out Spearman correlations, with min–max seed ranges; grey: FULL60 shuffled-label control.

entially absent across architectures. With BOS conditioning, both FEAT and EMB yield consistently high CV-BINARY scores across architectures. For PERM and RAND, the architectures differ more substantially: the Transformers already reach scores of 0.82–0.84, close to the probe ceiling of 0.85, which FEAT attains, whereas the LSTM reaches only 0.38 under PERM and 0.59 under RAND. With full context, this architectural difference largely disappears: the consonant–vowel distinction is strongly recoverable across all architectures and input conditions ($\rho = 0.78$ –0.85), with the LSTM reaching about 0.84 and the Transformers 0.78–0.81 under both PERM and RAND.

The CV-BINARY probe closely tracks FULL60 across architectures, conditions, and representation spaces; the largest difference between their mean scores is 0.07, suggesting that the high FULL60 scores largely reflect the broad consonant–vowel distinction. We therefore turn to the CONSONANTS probe to examine the finer hierarchy among consonants. The consonant hierarchy is most clearly recoverable under FEAT in every architecture, trivially at the input layer but also from BOS-conditioned and in-context representations: in-context CONSONANTS scores reach 0.79 in the LSTM, 0.66 in the two-layer Transformer, and 0.60 in the four-layer Transformer (Tab. 9, App. I), indicating that sonority structure supplied explicitly at the input remains recoverable after contextual

processing. Under EMB, PERM, and RAND, some consonant hierarchy structure also becomes recoverable in context ($\rho = 0.37$ –0.49 across architectures). Its recovery under PERM and RAND, in particular, shows that a partial consonant hierarchy can emerge through contextual processing even without sonority-aligned input representations, although to a substantially lesser degree than the consonant–vowel distinction.

Per-phoneme ranking errors. For each model, we rank the predictions of the in-context CONSONANTS probe and compare them with the expected sonority ordering: obstruents at positions 1–17, nasals at 18–20, liquids at 21–22, and glides at 23–24. For each consonant, we measure how far its predicted position falls outside the range expected for its sonority class (positive for overranking, negative for underranking) and average this displacement across seeds. The most severe errors concern the nasals, the glide /j/, and the obstruent /h/. Among nasals, across the three non-FEAT conditions, /ŋ/ has a more negative mean displacement than /m/ in the LSTM, whereas the reverse holds in the Transformers. Among obstruents, /h/ shows a larger positive mean displacement in the Transformers, including under FEAT. Across phonemes, FEAT reduces the magnitude of mean displacement most clearly in the LSTM. Full per-phoneme results are reported in App. G (Fig. 10).

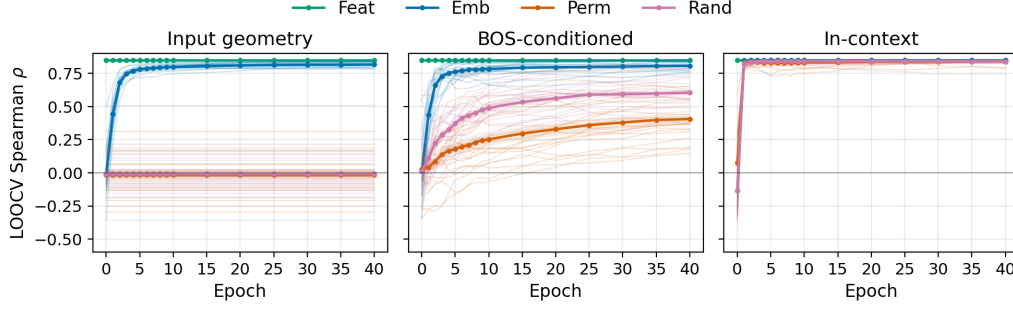
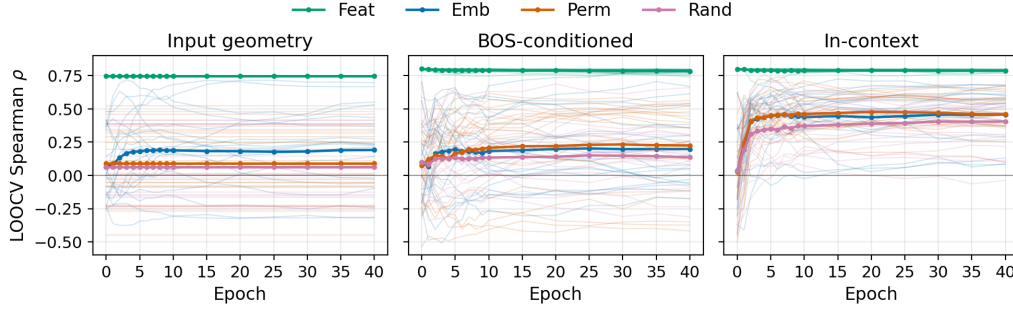
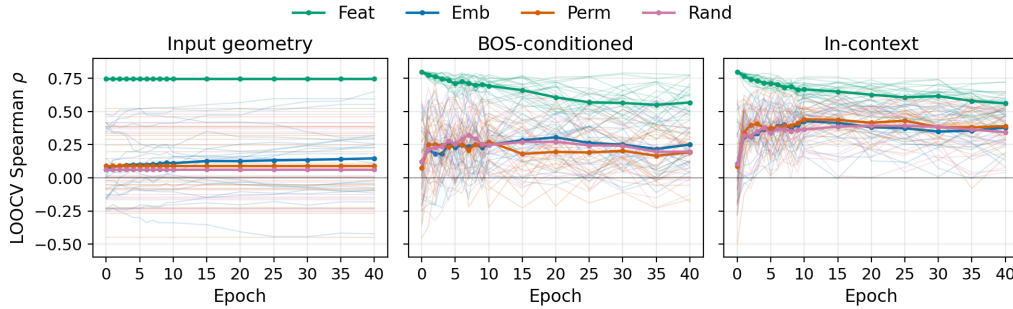


Figure 3: Consonant–vowel (CV-BINARY) recoverability across LSTM checkpoints. Thin lines: the twenty seeds, including initialisation; thick lines: their means; shading: ± 1 standard error of the mean; markers: probed checkpoints. Scores are leave-one-out Spearman correlations from the ridge regression probes.



(a) LSTM.



(b) Two-layer Transformer.

Figure 4: Consonant-hierarchy recoverability (CONSONANTS) across checkpoints. Thin lines: the twenty seeds; thick lines: their means; scores as in Fig. 3. Four-layer Transformer in Fig. 14 (App. I).

4.1.2 How Does Sonority Structure Develop During Training?

The consonant–vowel distinction becomes recoverable from in-context states early in training—by epoch 1 in the LSTM—and remains so under every architecture and input condition (Fig. 3; see Fig. 13 for Transformers). Under the minimal context provided by BOS-conditioned states, however, the architectures differ from the first epoch. Under PERM and RAND, the Transformers reach CV-BINARY scores of 0.64–0.80 after one epoch, compared with 0.04 and 0.11 in the LSTM, so the Transformer advantage under minimal context observed at the best-validation checkpoint (§4.1.1) is

already present early in training. Within the LSTM, PERM lags behind RAND under minimal context throughout training (0.41 against 0.61 at epoch 40), consistent with PERM’s misleading feature similarities being harder to overcome than RAND’s random geometry; with full context, no such gap appears, and both conditions score 0.84 from the first epoch onwards.

The consonant hierarchy follows a different trajectory. In the LSTM, in-context recoverability rises early and then plateaus (Fig. 4a). In the Transformers, the trajectories are less stable over training. Under FEAT, recoverability declines steadily from initialisation, from 0.80 to 0.56 in the two-layer model and 0.46 in the four-layer model by epoch

40 (the best-validation checkpoints, at epochs 7–17, still show 0.66 and 0.60; §4.1.1). Under EMB, PERM, and RAND, it peaks between epochs 5 and 20, then declines clearly in the four-layer Transformer while largely plateauing in the two-layer model (Figs. 4b and 14).

4.2 Similarity Between Representation Spaces

We use representational similarity analysis (Kriegeskorte et al., 2008) to compare input and hidden geometries for the 24 consonants at the best-validation checkpoint. In the LSTM, input-to-in-context similarity is highest under FEAT (0.72) and lowest under PERM (0.43), with EMB (0.54) and RAND (0.52) in between; see App. F for the full RSA results. A plausible explanation is that the feature structure in FEAT is compatible with distinctions learned from the training data and can therefore be retained, whereas PERM imposes a misleading feature similarity structure that the model must overcome. EMB and RAND pattern alike: both start from the same random geometry (§3.2), and EMB learns little consonant-hierarchy structure at the input layer (CONSONANTS 0.18; Tab. 3). In the Transformers, by contrast, input-to-hidden similarities are far lower (0.03–0.45, against 0.43–0.72 in the LSTM; App. F), so their hidden geometries are largely decoupled from the input geometry, feature-based or not. RSA alone cannot establish whether retaining or changing representation structure helps sonority projection, which we turn to in the next section.

5 Evaluating Phonotactic Generalisation

We evaluate the models’ phonotactic predictions in three ways: their predicted sonority contours on held-out words; their sensitivity to nonce-word violations; and their alignment with human judgments of unattested onsets, i.e., sonority projection. Throughout, we compare them with a bigram baseline with Witten–Bell smoothing (Witten and Bell, 1991) trained on the same data.

5.1 Predicted Sonority Contours

We first ask whether the models’ next-phoneme predictions reproduce the distribution of sonority changes observed in held-out words from the lexical test split. At each held-out position, we group the next-phoneme probability mass by the sonority-rank change it would induce relative to the current phoneme (*fall 4* to *rise 4*, *level*, or word *end*), then

average these distributions over positions with the same syllabic role and compare them with the corresponding empirical proportions. Models are trained on unsyllabified phoneme sequences; syllabic labels are used only to group positions for analysis. Results are shown in Fig. 5. The predicted distributions closely match the empirical proportions in every syllabic role except final codas, where the word end is not predictable from the preceding context.

At onset positions, the models favour sonority rises. For the first consonant of an onset, *rise 4* is the largest category; a rise of four ranks can only lead from an obstruent to a vowel, showing a strong expectation of vowels. After vowels, *fall 4* is the largest category, indicating a preference for low-sonority obstruents. These patterns hold across input conditions and architectures (Transformers in Fig. 8, App. E).

5.2 Sensitivity to Phonotactic Violations

Our second evaluation of phonotactic generalisation uses 40 violator–control nonce-word pairs from Hayes and White (2013): each violator breaches one target phonotactic constraint, and its matched control is otherwise as similar as possible while violating none. Twenty pairs involve natural constraints—phonetically and/or typologically motivated, several penalising sonority-sequencing violations—and twenty involve unnatural ones. The unnatural constraints were induced from an English lexicon by the Hayes–Wilson learner (Hayes and Wilson, 2008) and selected to have few or no lexical exceptions and weights comparable to the natural set, despite at most weak phonetic or typological motivation; under the Hayes–Wilson grammar, natural and unnatural violations are thus penalised with similar strength. Hayes and White call such constraints *accidentally true*: true of the lexicon but not part of speakers’ phonotactic knowledge.

For each model and violator–control nonce-word pair, we compute the difference in mean per-token surprisal $\Delta_{\text{surprisal}}$ between the violator and its control, with positive values indicating that the model assigns lower probability to the violator. For natural and unnatural constraints separately, we report the mean $\Delta_{\text{surprisal}}$ and discrimination accuracy, defined as the percentage of pairs for which the violator has higher surprisal than its control.

Results are summarised in Fig. 6 (numerical values in Tab. 4, App. D). Across all input conditions

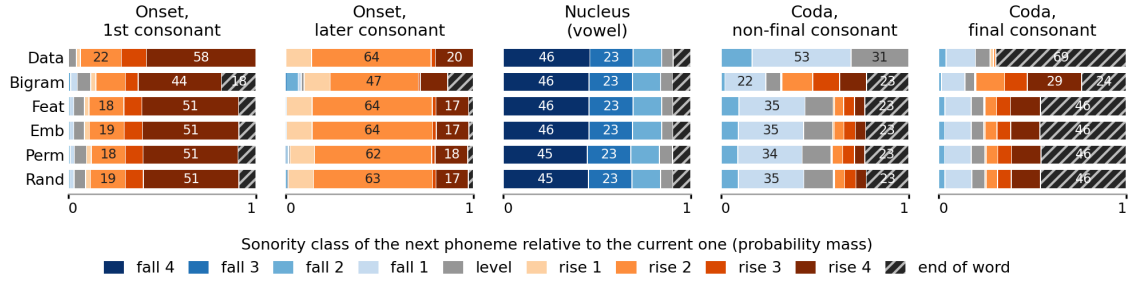


Figure 5: Predicted sonority contours on held-out words. Bars show next-phoneme probability mass grouped by sonority-rank change relative to the current phoneme and averaged across held-out contexts (LSTM; twenty seeds per condition; Transformers in Fig. 8, App. E). The Data row gives the corresponding empirical proportions.

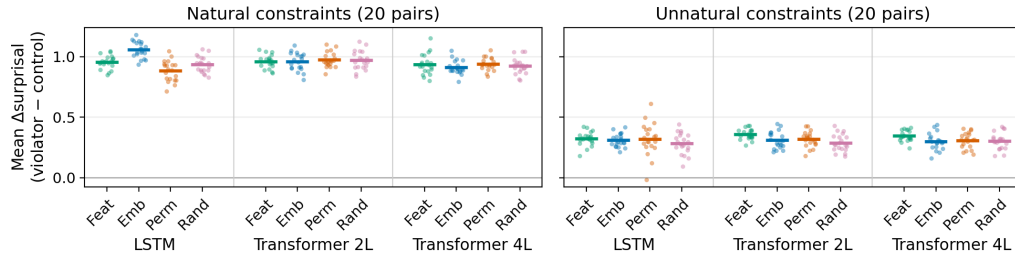


Figure 6: Violator-control discrimination: mean Δ surprisal (violation minus control) over the natural and the unnatural pairs, for every model (points; twenty seeds per condition) with condition means (bars), by architecture.

and architectures, violators receive higher surprisal than their matched controls on average. Discrimination is consistently stronger for natural constraints, where mean Δ surprisal ranges from 0.88 to 1.06, and models assign higher surprisal to the violator in 93–95% of pairs. For unnatural constraints, mean Δ surprisal ranges from 0.28 to 0.36, with discrimination accuracies of 71–80%. No input condition or architecture exhibits a consistent advantage across the two constraint groups. Like the Hayes–Wilson learner, the neural models capture these accidentally true generalisations, though more weakly than the natural ones. In the human experiment, violations of the unnatural constraints had at best a weak effect on ratings (Hayes and White, 2013, §5.2.5), a divergence that Hayes and White take as evidence for a naturalness bias in human learning.

Among the pairs that penalise sonority-sequencing violations, we single out the four whose constraints apply in codas, the only nonce-word test of coda sonority in our evaluations (the Daland items test onsets): three violators have rising coda clusters and one a plateau, against falling or level clusters in their controls (Tab. 7, App. D); all input conditions assign these violators higher surprisal on average. Discrimination trajectories across training are reported in App. H.

5.3 Alignment with Human Preferences

We test whether the models reproduce human preferences for nonce words, using Daland et al.’s dataset; see Tab. 6 (App. D) for onset and tail inventories. The onsets fall into three groups: Attested (common in English), Marginal (rare or mainly found in loanwords), and Unattested (absent from English). Our primary analysis concerns the Unattested group, whose onsets are also absent from our training data, since pooled correlations may largely reflect a shared preference for attested over unattested onsets.

We use negative mean per-token surprisal as the model’s acceptability rating and correlate it with human judgements using Spearman’s ρ . For the human judgements, we use two-alternative forced-choice scores because they preserve acceptability differences among unattested onsets that receive near-floor Likert ratings (see Daland et al., 2011, §2.6.3). Results are shown in Fig. 7; numerical correlation scores in Tab. 5 (App. D).

The bigram baseline correlates positively with human preferences when all items are pooled ($\rho = 0.53$), but negatively within the Unattested group ($\rho = -0.13$). By contrast, the neural models retain positive correlations within the Unattested group across all architectures and input conditions ($\rho = 0.43$ – 0.60). Within the Unattested group, FEAT has

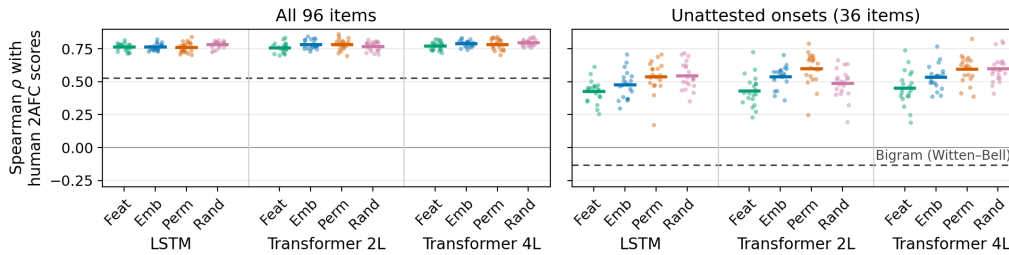


Figure 7: Spearman correlation between model acceptability and human two-alternative forced-choice scores, for all 96 items (left) and the 36 unattested-onset items (right); layout as in Fig. 6. Dashed line: bigram baseline.

lower mean correlations with human preferences than PERM and RAND in all three architectures, despite its higher consonant-hierarchy probe scores.

6 Discussion and Conclusion

Our results show that both sonority structure in internal representations and probabilistic phonotactic preferences can develop without sonority-aligned input representations. Earlier neural models of phonotactics with trainable embeddings left open whether their success depended on learning a sonority-aligned input geometry (Mayer and Nelson, 2020). Our frozen PERM and RAND conditions show that it need not: their input vectors cannot change during training, yet sonority distinctions become recoverable from contextual hidden states and model scores correlate with human preferences for unattested onsets. The EMB–RAND comparison allows us to assess more directly what is gained by learning the input representations themselves, since the two conditions start from the same random vectors and differ only in whether those vectors can be updated during training. Across architectures, allowing the input vectors to update yields only modest differences, both in consonant-hierarchy recoverability from hidden states and in behaviour, and neither difference consistently favours EMB.

When we probe separately for the consonant–vowel distinction and for the finer hierarchy among consonants, a clear asymmetry emerges. The consonant–vowel distinction is strongly recoverable from contextual representations under every input condition, whereas the consonant hierarchy is recovered much less consistently and is clearest under FEAT; in the Transformers it even declines over training, most clearly under FEAT and in the four-layer model. Consonant–vowel alternation is frequent, which may support the emergence of this distinction in model representations, whereas organising consonants along a fine-grained sonority hierarchy may not be necessary for predicting in-

dividual consonants. Despite its higher consonant-hierarchy probe scores, FEAT has lower mean correlations with human preferences for unattested onsets than PERM and RAND in all three architectures (§5.3).

Our developmental analyses provide complementary evidence that recoverability and behaviour need not track one another. In the Transformers, consonant-hierarchy recoverability from hidden states declines from initialisation under FEAT and peaks early under the other conditions, while violator–control discrimination improves over the same epochs and remains high at the best-validation checkpoint; in the LSTM, recoverability plateaus within the first epochs, whereas discrimination continues to improve for several more. This may be because phoneme-specific positional and co-occurrence patterns support phonotactic predictions without requiring a clearly ordered consonant hierarchy. More generally, however, the probes only measure whether sonority structure can be linearly recovered across phoneme types; they do not establish which information the models use for prediction.

Although all input conditions discriminate natural violations more strongly than unnatural ones, the models still penalise violations of unnatural constraints that hold in the lexicon but had at best a weak effect on human ratings (§5.2). The models’ success on unattested onsets therefore does not fully explain which lexical regularities human learners rely on to generalise.

Finally, learning sonority structure from an English-like lexicon and explaining why that lexicon is sonority-shaped are separate problems. Our experiments establish that the tested preferences can be acquired without a pre-specified sonority hierarchy in the inputs, but leave both the origins of the distribution and the presence of a sonority-specific bias in human learners open.

Limitations

Several limitations qualify these results. English phonotactics are strongly correlated with sonority, so the present findings need cross-linguistic replication, particularly in languages with different cluster restrictions.

Our models learn from a pronunciation lexicon, following established phonotactic modelling practice (Hayes and Wilson, 2008; Daland et al., 2011). Each training entry is presented once per epoch, without weighting by corpus token frequency. This choice is supported by Hayes and Wilson (2008, p. 395, fn. 10; p. 402), who found slightly better predictions of English onset judgements using type rather than token frequencies. The lexicon does not, however, reproduce human learners' vocabulary or exposure history. Whether the findings extend to vocabulary drawn from child-directed speech, and how they depend on token-frequency weighting, remain open questions.

Because the data are split at the entry level, some pronunciations occur in more than one split. Excluding validation entries whose pronunciations occurred in training left the best-validation epoch unchanged for 84.2% of models. For the remaining models, selecting a different checkpoint reduced loss on the filtered validation set by a median of 0.0004 and a maximum of 0.0021. The Hayes–White and Daland evaluations use separate nonce-word materials whose complete phoneme sequences are absent from training, so these evaluations do not involve this overlap. However, in the lexical test split used for the sonority-contour analysis, 4.1% of entries share a pronunciation with a training entry, which may affect the aggregate contours.

The remaining limitations concern the resolution of our measurements. Checkpoints were saved once per epoch, but much of the learning happens within the first epoch: the consonant–vowel distinction is at ceiling and the consonant hierarchy already partly recoverable at the epoch-1 checkpoint, so we cannot resolve their relative order of emergence. The within-group nonce-word correlations are computed over small item sets and vary considerably across seeds. Only four Hayes and White pairs target the sonority-related coda constraints considered here (Tab. 7), limiting the coverage of this part of the evaluation. Finally, the behavioural evaluations do not identify the contribution of a particular internal representation; targeted contrasts

and interventions would be needed to establish the behavioural role of the fine-grained distinctions probed here.

Acknowledgments

We thank James White for sharing the experimental materials used in both of our behavioural evaluations (Daland et al., 2011; Hayes and White, 2013), and for clarifying the interpretation of the Hayes and White (2013) findings. We also thank the anonymous reviewers for their comments, which helped improve the paper.

References

- Kabir Ahuja, Vidhisha Balachandran, Madhur Panwar, Tianxing He, Noah A. Smith, Navin Goyal, and Yulia Tsvetkov. 2025. [Learning syntax without planting trees: Understanding hierarchical generalization in transformers](#). *Transactions of the Association for Computational Linguistics*, 13:121–141.
- Iris Berent, Donca Steriade, Tracy Lennertz, and Vered Vaknin. 2007. [What we know about what we have never heard: Evidence from perceptual illusions](#). *Cognition*, 104(3):591–630.
- Juliette Blevins. 2004. *Evolutionary Phonology: The Emergence of Sound Patterns*. Cambridge University Press, Cambridge.
- G. N. Clements. 1990. [The role of the sonority cycle in core syllabification](#). In John Kingston and Mary E. Beckman, editors, *Papers in Laboratory Phonology*, pages 283–333. Cambridge University Press, Cambridge.
- John Coleman and Janet Pierrehumbert. 1997. [Stochastic phonological grammars and acceptability](#). In *Computational Phonology: Third Meeting of the ACL Special Interest Group in Computational Phonology*.
- Huteng Dai, Connor Mayer, and Richard Futrell. 2023. [Rethinking representations: A log-bilinear model of phonotactics](#). In *Proceedings of the Society for Computation in Linguistics 2023*, pages 259–268, Amherst, MA. Association for Computational Linguistics.
- Robert Daland, Bruce Hayes, James White, Marc Garellek, Andrea Davis, and Ingrid Norrmann. 2011. [Explaining sonority projection effects](#). *Phonology*, 28(2):197–234.
- Bruce Hayes and James White. 2013. [Phonological naturalness and phonotactic learning](#). *Linguistic Inquiry*, 44(1):45–75.
- Bruce Hayes and Colin Wilson. 2008. [A maximum entropy model of phonotactics and phonotactic learning](#). *Linguistic Inquiry*, 39(3):379–440.

- John Hewitt and Percy Liang. 2019. [Designing and interpreting probes with control tasks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2733–2743, Hong Kong, China. Association for Computational Linguistics.
- Julie Kallini, Isabel Papadimitriou, Richard Futrell, Kyle Mahowald, and Christopher Potts. 2024. [Mission: Impossible language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14691–14714, Bangkok, Thailand. Association for Computational Linguistics.
- Nikolaus Kriegeskorte, Marieke Mur, and Peter A. Bandettini. 2008. [Representational similarity analysis – connecting the branches of systems neuroscience](#). *Frontiers in Systems Neuroscience*, 2:4.
- Connor Mayer and Max Nelson. 2020. [Phonotactic learning with neural language models](#). In *Proceedings of the Society for Computation in Linguistics 2020*, pages 291–301, New York, New York. Association for Computational Linguistics.
- Nicole Mirea and Klinton Bicknell. 2019. [Using LSTMs to assess the obligatoriness of phonological distinctive features for phonotactic learning](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1595–1605, Florence, Italy. Association for Computational Linguistics.
- Elliott Moreton. 2008. [Analytic bias and phonological typology](#). *Phonology*, 25(1):83–127.
- Elisabeth O. Selkirk. 1984. On the major class features and syllable theory. In Mark Aronoff and Richard T. Oehrle, editors, *Language Sound Structure: Studies in Phonology Presented to Morris Halle by His Teacher and Students*, pages 107–136. MIT Press, Cambridge, MA.
- Taiga Someya, Anej Svete, Brian DuSell, Timothy J. O’Donnell, Mario Giulianelli, and Ryan Cotterell. 2025. [Information locality as an inductive bias for neural language models](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 27995–28013, Vienna, Austria. Association for Computational Linguistics.
- R. L. Weide. 1998. The CMU pronouncing dictionary. <http://www.speech.cs.cmu.edu/cgi-bin/cmudict>.
- Ian H. Witten and Timothy C. Bell. 1991. [The zero-frequency problem: Estimating the probabilities of novel events in adaptive text compression](#). *IEEE Transactions on Information Theory*, 37(4):1085–1094.

A Sonority Ranks

Tab. 1 lists the five sonority ranks and the CMU phonemes assigned to each.

B Validation Loss

Tab. 2 summarises the best-validation losses reported in §3.4.

C Endpoint Probe Values

Tab. 3 gives the numerical values for the LSTM row of Fig. 2.

D Phonotactic Evaluation: Full Results

Tab. 4 gives the violator–control discrimination values plotted in Fig. 6; Tab. 7 lists the 40 pairs with the constraint each violator breaches; Tab. 5 gives the correlations with the human two-alternative forced-choice scores, pooled and by attestedness group.

E Predicted Sonority Contours: Transformers

Fig. 8 repeats Fig. 5 for the two Transformers. Their distributions differ from the LSTM’s by less than 0.02 in the probability mass of any category, in every syllabic role and input condition.

F Similarity Between Representation Spaces

Representational similarity is computed between the representation spaces at the best-validation checkpoint: the pairwise correlation distances between phonemes are computed within each space, and two spaces are compared by the Spearman correlation of their distances. The analysis is restricted to the 24 consonants, since the consonant–vowel block structure otherwise dominates.

Results. Fig. 9 and Tab. 8 show the results. In the LSTM, input-to-BOS similarity is 0.59–0.70 in all four conditions, lowest for PERM. The lower value for PERM indicates weaker correspondence between the pairwise distance rankings in the two spaces. One possible interpretation is that PERM imposes structured but misleading similarities by assigning articulatory feature vectors to the wrong phonemes, and the model may need to undo these relationships in its hidden representations to learn distributionally useful distinctions. By contrast, RAND imposes no feature-based grouping to undo,

although its random geometry may still need to be transformed. The lower input-to-BOS similarity for PERM is consistent with this hypothesis, but RSA alone does not establish that such correction occurs or is necessary for prediction. The substantial input-to-BOS similarity also holds for RAND: random vectors also define a pattern of pairwise similarities, and the BOS-conditioned space resembles this phonologically arbitrary pattern to a similar degree as the feature-based pattern. All conditions therefore show substantial correspondence between input and BOS-conditioned geometry, PERM somewhat less. Second, input-to-in-context similarity is higher for FEAT (0.72) than for PERM (0.43), EMB (0.54) or RAND (0.52). These differences are consistent with a relationship between input alignment and contextual geometry, but the input conditions vary multiple representational properties, so the comparison does not isolate sonority alignment as the cause. The probes provide a complementary but restricted measure of linear generalisation across phoneme types; they do not identify how sonority-related information is represented more generally or whether it is used for prediction. Establishing a specific mechanism would require controls that isolate the relevant input properties or targeted interventions on the representations. In the Transformers, all input-to-hidden similarities are much lower (0.03–0.45), while BOS-to-in-context similarity remains at 0.57–0.64: their hidden geometries are largely decoupled from the input geometry, feature-based or not.

G Which Phonemes Do the Probes Misrank?

Fig. 10 shows the twenty-seed mean displacements defined in §4.1.1 for every consonant, architecture, and input condition. Two properties of the measure matter for interpretation. First, positive and negative displacements can cancel across seeds, so a near-zero mean does not necessarily imply consistently accurate ranking; the figure’s dots mark directional consistency across seeds, not statistical significance. Second, the measure is asymmetric across classes: nonzero displacements can only be positive for obstruents and negative for glides, and the unequal block sizes give classes different probabilities of containing a randomly assigned position. We therefore focus on differences between phonemes within the same sonority class.

Beyond the patterns reported in §4.1.1, /s/ shows

positive mean displacements without feature input, and /f/ is more prominently displaced in the Transformers than in the LSTM. Among nasals, /n/ has negative mean displacement in the non-feature LSTMs but positive mean displacement in the Transformers, distinguishing it from /m/ and /ŋ/. Within liquids, /l/ has lower mean displacement than /r/ across all architectures and input conditions; within glides, /j/ has more negative mean displacement than /w/ across all Transformer conditions, whereas their ordering varies across input conditions in the LSTM. In the Transformers, substantial mean displacements remain even under FEAT, notably for /h/ and /j/.

H Violator–control Discrimination Across Checkpoints

The violator–control discrimination of §5.2 was computed on the developmental checkpoints of §3.4 (the untrained model, every epoch up to 10, and every fifth epoch thereafter), for the same twenty seeds, every condition, and every architecture. The resulting trajectories are shown in Fig. 11. Mean Δ surprisal rises early and then grows more slowly overall, with occasional late decreases. The conditions show broadly similar developmental trends, with differences in magnitude, particularly for the four-pair sonority subset. Mean Δ surprisal averaged over epochs 35 and 40 is higher than over epochs 5–10 for every architecture, input condition, and pair group. Because the Transformers overfit after their best-validation epochs (§3.4), this late growth partly reflects a sharpening of the output distribution rather than better discrimination: for natural pairs, the proportion with $\Delta > 0$ peaks at 94–95% between epochs 5 and 15 and falls to 89–94% by epoch 40, whereas in the LSTM it plateaus at 92–95% from epoch 10.

I Supplementary Probe Results

Figs. 12 to 14 give the complete developmental comparison: for each probe, the three architectures (rows) and the three representation spaces (columns), with all twenty seeds at every checkpoint. Tab. 9 gives the CONSONANTS endpoint values across architectures.

Rank	Class	CMU phonemes
1	Obstruents	P B T D K G CH JH F V TH DH S Z SH ZH HH
2	Nasals	M N NG
3	Liquids	L R
4	Glides	W Y
5	Vowels	AA AE AH AO AW AY EH ER EY IH IY OW OY UH UW (any stress variant)

Table 1: Sonority ranks and corresponding phonemes.

Condition	LSTM	Transformer 2L	Transformer 4L
FEAT	2.148 [2.139, 2.155]	2.153 [2.145, 2.161]	2.149 [2.140, 2.160]
EMB	2.136 [2.129, 2.142]	2.167 [2.159, 2.176]	2.161 [2.152, 2.168]
PERM	2.187 [2.173, 2.199]	2.171 [2.160, 2.176]	2.162 [2.153, 2.172]
RAND	2.156 [2.147, 2.168]	2.170 [2.156, 2.181]	2.161 [2.152, 2.171]

Table 2: Best-validation cross-entropy (nats per predicted token, including EOS), shown as mean [min, max] across twenty seeds. Each seed contributes its minimum loss over 40 epochs on the original validation split.

Space	Probe	FEAT	EMB	PERM	RAND
Input	FULL60	0.870	0.779	0.004	0.007
Input	CV-BINARY	0.849	0.813	-0.017	-0.007
Input	CONSONANTS	0.747	0.183	0.089	0.061
BOS hidden	FULL60	0.871	0.767	0.347	0.519
BOS hidden	CV-BINARY	0.849	0.801	0.385	0.592
BOS hidden	CONSONANTS	0.787	0.198	0.225	0.144
In-context	FULL60	0.871	0.859	0.850	0.853
In-context	CV-BINARY	0.849	0.849	0.838	0.840
In-context	CONSONANTS	0.788	0.450	0.462	0.395

Table 3: Probe decomposition at the best-validation checkpoint (LSTM, twenty-seed means). Shuffled-label controls for FULL60 and CONSONANTS are near zero and are omitted for readability; see the LSTM row of Fig. 2 for seed ranges and the FULL60 shuffled baseline.

Constraint	Cond.	LSTM		Transformer 2L		Transformer 4L	
		Δ	%	Δ	%	Δ	%
Natural	FEAT	0.957	94	0.961	95	0.938	94
Natural	EMB	1.060	94	0.960	94	0.913	93
Natural	PERM	0.884	94	0.976	94	0.941	95
Natural	RAND	0.937	95	0.972	95	0.925	94
Unnatural	FEAT	0.323	74	0.359	80	0.345	79
Unnatural	EMB	0.310	72	0.312	74	0.301	76
Unnatural	PERM	0.321	74	0.319	76	0.309	75
Unnatural	RAND	0.285	71	0.290	75	0.303	73

Table 4: Violator-control discrimination across architectures: mean Δ surprisal (violator minus control) and percentage of pairs in which the violator is assigned higher surprisal (twenty-seed means).

Scope	Cond.	LSTM	Transformer 2L	Transformer 4L
Pooled (96)	FEAT	0.764	0.759	0.774
Pooled (96)	EMB	0.764	0.784	0.790
Pooled (96)	PERM	0.761	0.785	0.786
Pooled (96)	RAND	0.782	0.769	0.797
Pooled (96)	Bigram		0.530	
Unattested (36)	FEAT	0.426	0.432	0.454
Unattested (36)	EMB	0.478	0.540	0.535
Unattested (36)	PERM	0.541	0.600	0.595
Unattested (36)	RAND	0.546	0.489	0.599
Unattested (36)	Bigram		-0.135	
Marginal (24)	FEAT	0.364	0.155	0.226
Marginal (24)	EMB	0.188	0.132	0.205
Marginal (24)	PERM	0.289	0.152	0.160
Marginal (24)	RAND	0.233	0.143	0.151
Marginal (24)	Bigram		0.077	
Attested (36)	FEAT	0.364	0.354	0.364
Attested (36)	EMB	0.453	0.340	0.421
Attested (36)	PERM	0.470	0.368	0.419
Attested (36)	RAND	0.410	0.341	0.406
Attested (36)	Bigram		0.357	

Table 5: Spearman correlation between model acceptability and human two-alternative forced-choice scores, pooled and by attestedness group (twenty-seed means). The bigram rows are a single Witten–Bell bigram model trained on the same training split.

Onsets		
Attested (18)	Marginal (12)	Unattested (18)
tw tr sw fr pr pl kw kr kl gr gl fr fl dr br bl sn sm	gw fl vw fw fn fm vl bw dw fw vr θw	pw zr mr tl dn km fn ml nl dg pk lm ln rl lt rn rd rg
Tails (6)		
-atuf	-ibid	-asip
-εpid	-igif	-εzig

Table 6: Onset and tail inventories used to construct the 96 nonce words of [Daland et al. \(2011\)](#). Each of the 48 onsets was combined with two of the six tails; for example, dr occurs with -εpid and -igif.

Pair (violator/control)	Naturalness	Constraint violated	Son.-seq.
<i>bikf/bimf</i>	natural	*[+DORSAL, +rhyme][+LABIAL, +rhyme]	
<i>bwell/brell</i>	natural	*[+LABIAL, -rhyme][+LABIAL, -rhyme]	
<i>canifl/canift</i>	natural	*[-sonorant, +rhyme][+sonorant, +rhyme]	yes
<i>cnopel/clope</i>	natural	*[-continuant, -rhyme][+nasal, -rhyme]	
<i>cping/sping</i>	natural	*[-continuant, -rhyme][-continuant, -rhyme]	
<i>ctice/stice</i>	natural	*[-continuant, -rhyme][-continuant, -rhyme]	
<i>esger/ezger</i>	natural	*[-voice][+voice]	
<i>hlup/plup</i>	natural	*[-consonantal, -rhyme][+consonantal, -rhyme]	
<i>hmit/smit</i>	natural	*[-consonantal, -rhyme][+consonantal, -rhyme]	
<i>jouyl/jout</i>	natural	*[+high, +rhyme]	
<i>kipl/kilp</i>	natural	*[-sonorant, +rhyme][+sonorant, +rhyme]	yes
<i>pneck/sneck</i>	natural	*[-continuant, -rhyme][+nasal, -rhyme]	
<i>pwickon/twickon</i>	natural	*[+LABIAL, -rhyme][+LABIAL, -rhyme]	
<i>rufl/ruft</i>	natural	*[+LABIAL, +rhyme][+DORSAL, +rhyme]	
<i>sadekp/sadect</i>	natural	*[+DORSAL, +rhyme][+LABIAL, +rhyme]	
<i>shapenr/shapent</i>	natural	*[+consonantal, +rhyme][-consonantal, +rhyme]	yes
<i>tighw/tibe</i>	natural	*[+high, +rhyme]	
<i>tilr/tilse</i>	natural	*[+consonantal, +rhyme][-consonantal, +rhyme]	yes
<i>trefkl/treft</i>	natural	*[+LABIAL, +rhyme][+DORSAL, +rhyme]	
<i>trocdall/troctal</i>	natural	*[-voice][+voice]	
<i>boitcher/boisser</i>	unnatural	*[+round, -back][-anterior]	
<i>foushert/fousert</i>	unnatural	*[+diphthong][+continuant, -anterior]	
<i>hethker/hethler</i>	unnatural	*[-strident][-sonorant]	
<i>ishtyl/ishmy</i>	unnatural	*[+consonantal, -anterior][-sonorant]	
<i>luhallem/laihallem</i>	unnatural	*[+round, +high][-consonantal, -sonorant]	
<i>metchter/metchner</i>	unnatural	*[+consonantal, -anterior][-sonorant]	
<i>muthpy/muspy</i>	unnatural	*[-strident][-sonorant]	
<i>noiran/nyron</i>	unnatural	*[+round, -back][-anterior]	
<i>oidloit</i>	unnatural	*[+word_boundary][+diphthong, +round][+voice]	
<i>ooker/locker</i>	unnatural	*[+word_boundary][-diphthong, +round, +high]	
<i>ouziel/oussie</i>	unnatural	*[+word_boundary][+diphthong, +round][+voice]	
<i>pothol/pothy</i>	unnatural	*[-strident][-stress, +round]	
<i>pyshon/pyson</i>	unnatural	*[+diphthong][+continuant, -anterior]	
<i>taithol/taithy</i>	unnatural	*[-strident][-stress, +round]	
<i>tuheim/towheim</i>	unnatural	*[+round, +high][-consonantal, -sonorant]	
<i>utrum/otrum</i>	unnatural	*[+word_boundary][-diphthong, +round, +high]	
<i>youse/yoss</i>	unnatural	*[-back][+diphthong]	
<i>youtlyut</i>	unnatural	*[-back][+diphthong]	
<i>zhad/zhar</i>	unnatural	*[+continuant, +voice, -anterior][+stress][-sonorant]	
<i>zhepl/zhem</i>	unnatural	*[+continuant, +voice, -anterior][+stress][-sonorant]	

Table 7: The 40 Hayes and White (2013) violator–control pairs with the constraint each violator breaches, as given in the Daland et al. (2011) materials. Hayes and White (their §4.1) classify five natural constraints as sonority-based, three of them onset constraints (/h/ followed by a consonant, stop–stop, stop–nasal); we mark only the two coda constraints. The four pairs marked as sonority-sequencing are thus selected by their violated feature constraint (*[-sonorant, +rhyme][+sonorant, +rhyme] or *[+consonantal, +rhyme][-consonantal, +rhyme]); this constraint-based criterion defines the set used in §5.2. On our sonority scale, three violators have rising coda clusters and one, *tilr* (T IH1 L R), has a plateau: L and R both have rank 3, although the original feature constraint distinguishes them. The matched controls have falling or level coda clusters.

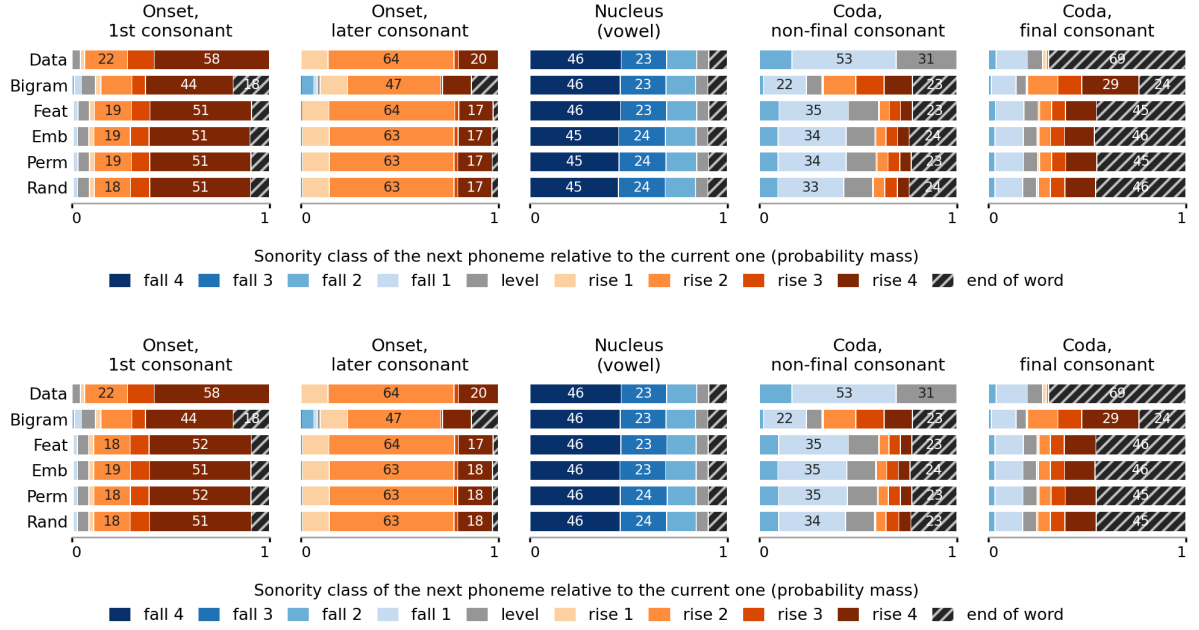


Figure 8: Predicted sonority contours for the two-layer (top) and four-layer (bottom) Transformers; layout as in Fig. 5.

Model	Pair	FEAT	EMB	PERM	RAND
LSTM	input $\leftrightarrow$ BOS hidden	0.703	0.672	0.586	0.680
LSTM	input $\leftrightarrow$ in-context	0.715	0.536	0.432	0.516
LSTM	BOS $\leftrightarrow$ in-context	0.754	0.655	0.669	0.651
Transformer 2L	input $\leftrightarrow$ BOS hidden	0.355	0.181	0.097	0.133
Transformer 2L	input $\leftrightarrow$ in-context	0.449	0.156	0.046	0.100
Transformer 2L	BOS $\leftrightarrow$ in-context	0.601	0.616	0.642	0.620
Transformer 4L	input $\leftrightarrow$ BOS hidden	0.300	0.123	0.046	0.088
Transformer 4L	input $\leftrightarrow$ in-context	0.440	0.120	0.031	0.071
Transformer 4L	BOS $\leftrightarrow$ in-context	0.569	0.599	0.630	0.593

Table 8: Representational similarity (Spearman correlation of pairwise correlation-distance matrices, 24 consonants; twenty-seed means) for all three architectures.

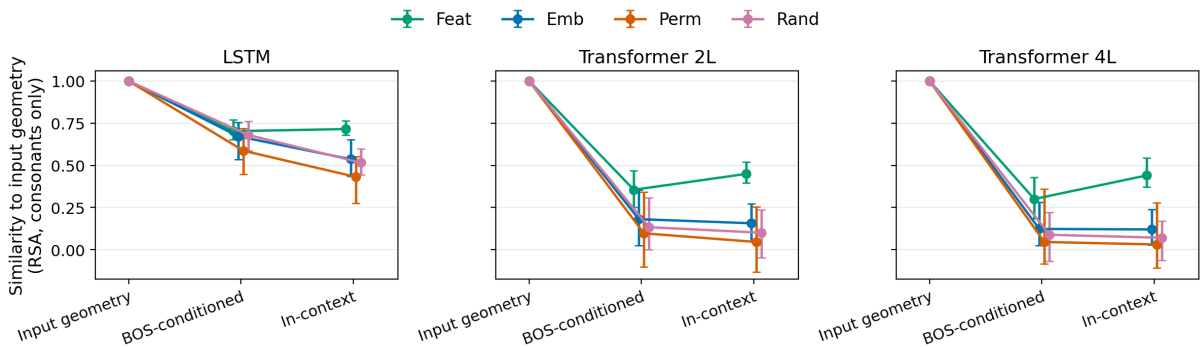


Figure 9: Similarity of each representation space to the input geometry (RSA over the 24 consonants, best-validation checkpoint; twenty-seed means with min-max seed ranges), for the LSTM and the two Transformers. The value at the input is 1 by definition. In the LSTM, the BOS-conditioned spaces retain substantial similarity to their inputs in every condition, PERM somewhat less, and in-context similarity is higher for FEAT than for the other conditions; in the Transformers, only FEAT retains appreciable similarity to the input.

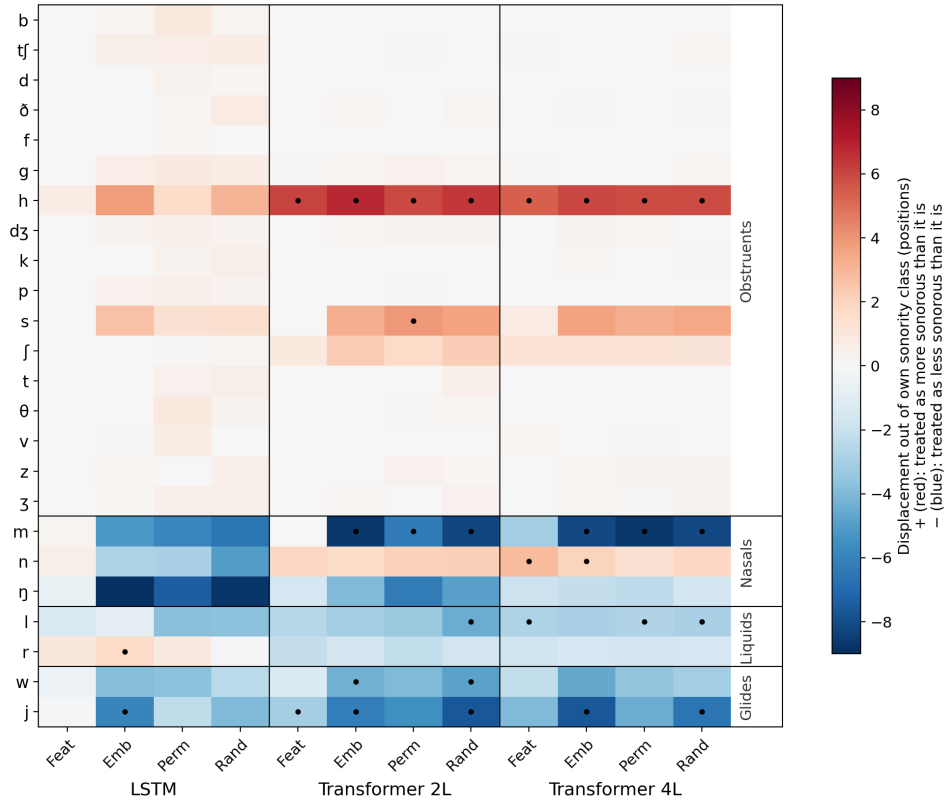


Figure 10: Per-phoneme errors of the CONSONANTS probe on the in-context representations at the best-validation checkpoint (means over twenty seeds). Each cell gives a phoneme’s mean displacement out of the block of ranking positions occupied by its sonority class: red (positive) indicates overranking on average, blue (negative) indicates underranking on average. White indicates a mean signed displacement near zero, which can reflect either rankings within the correct class block or cancellation of positive and negative displacements across seeds. Dots mark cells in which at least 90% of seeds are displaced in the same direction.

Model	FEAT	EMB	PERM	RAND
LSTM	0.788 [0.767, 0.802]	0.450 [0.033, 0.711]	0.462 [0.152, 0.651]	0.395 [−0.024, 0.657]
Transformer 2L	0.659 [0.584, 0.766]	0.422 [0.143, 0.599]	0.432 [0.208, 0.656]	0.374 [0.002, 0.591]
Transformer 4L	0.598 [0.298, 0.768]	0.432 [0.089, 0.613]	0.487 [0.284, 0.718]	0.476 [0.057, 0.675]

Table 9: Consonants-only probe at the best-validation checkpoint (in-context hidden states; twenty-seed means with min–max seed ranges in brackets). FEAT has the highest mean score in each architecture; the ordering of EMB, PERM and RAND is not stable across seeds or depths.

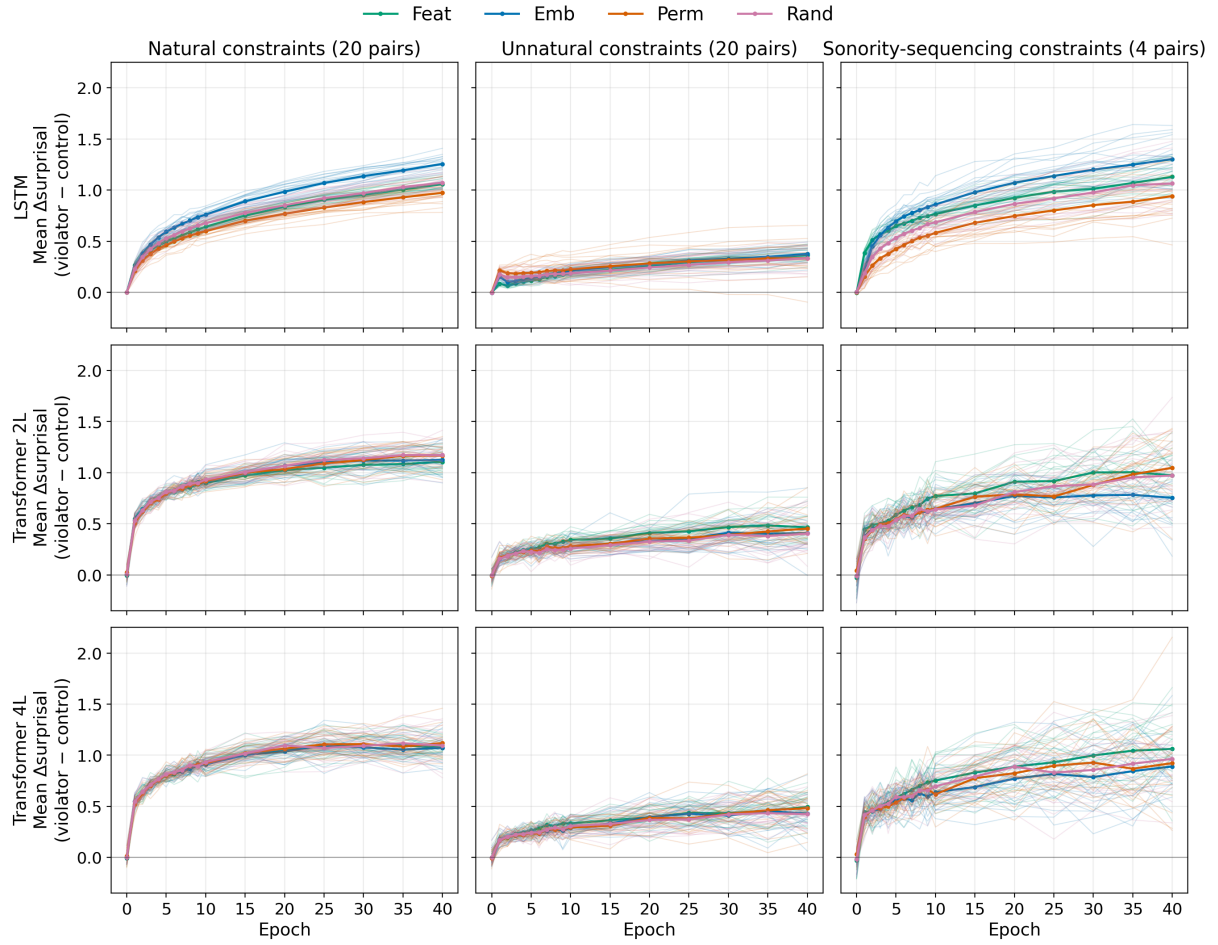


Figure 11: Hayes and White mean Δ surprisal (violator minus control) across training for the natural, unnatural, and sonority-sequencing pairs (columns) in each architecture (rows). Thin lines: individual seeds; thick lines: twenty-seed means; markers: probed checkpoints.

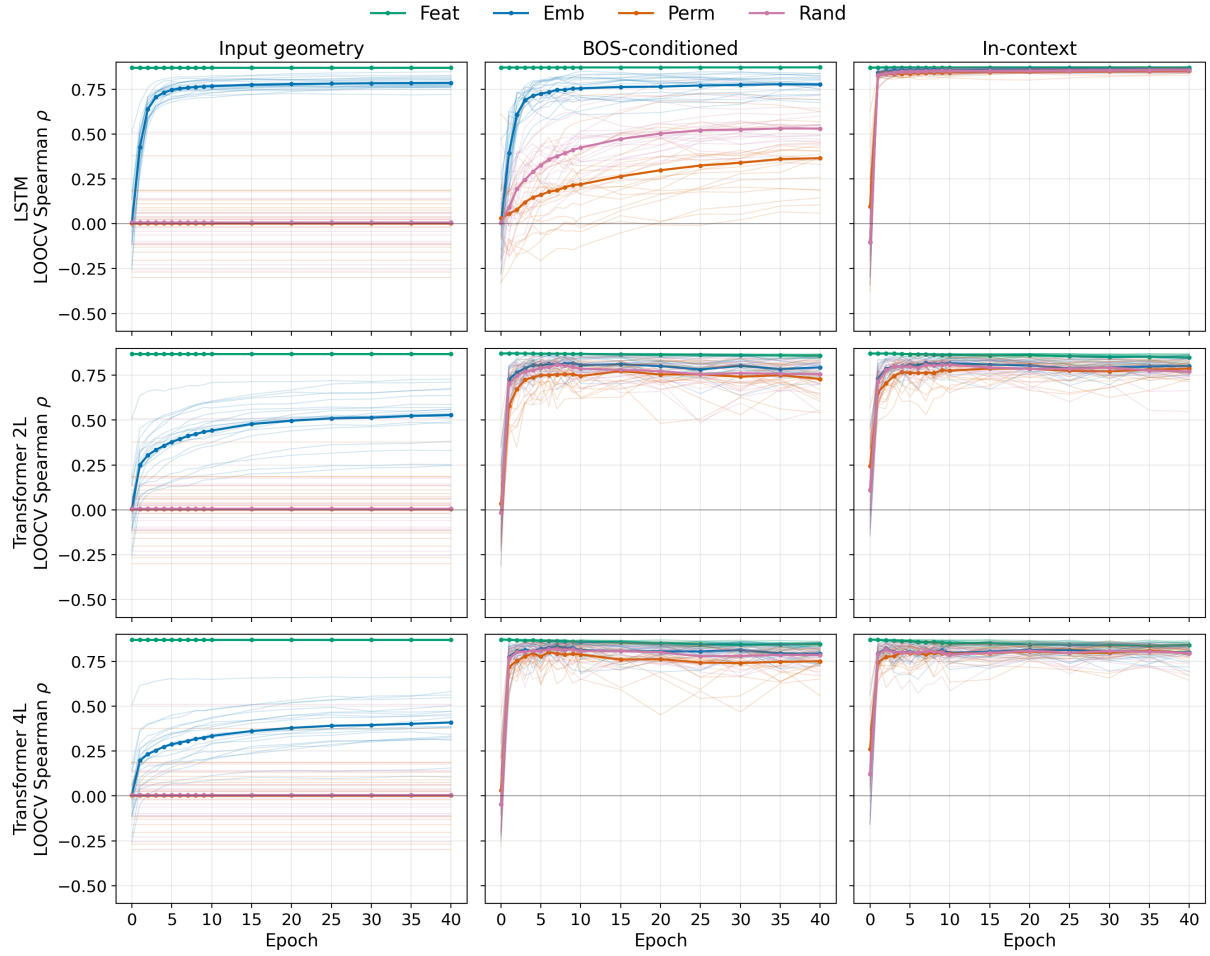


Figure 12: Full-inventory (FULL60) recoverability across training for the LSTM (top), two-layer Transformer (middle), and four-layer Transformer (bottom), in the input, BOS-conditioned, and in-context spaces (columns). Thin lines show the twenty training seeds at every checkpoint, including initialisation; thick lines show their means. Scores are leave-one-out Spearman correlations from ridge regression probes.

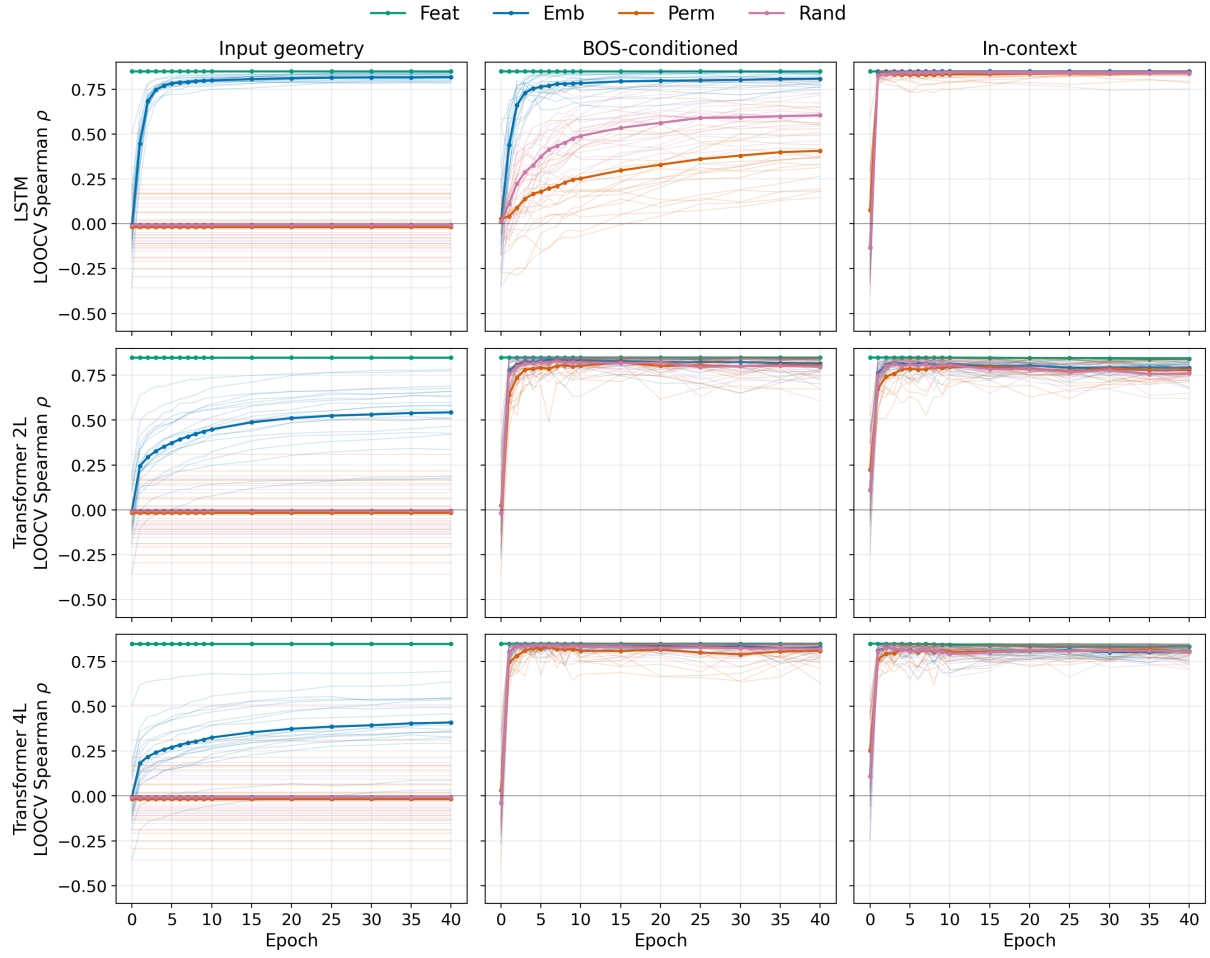


Figure 13: Consonant–vowel (CV-BINARY) recoverability across training; layout as in Fig. 12. In the BOS-conditioned space, the Transformers separate consonants from vowels under PERM and RAND from the first epoch, whereas the LSTM does so only slowly.

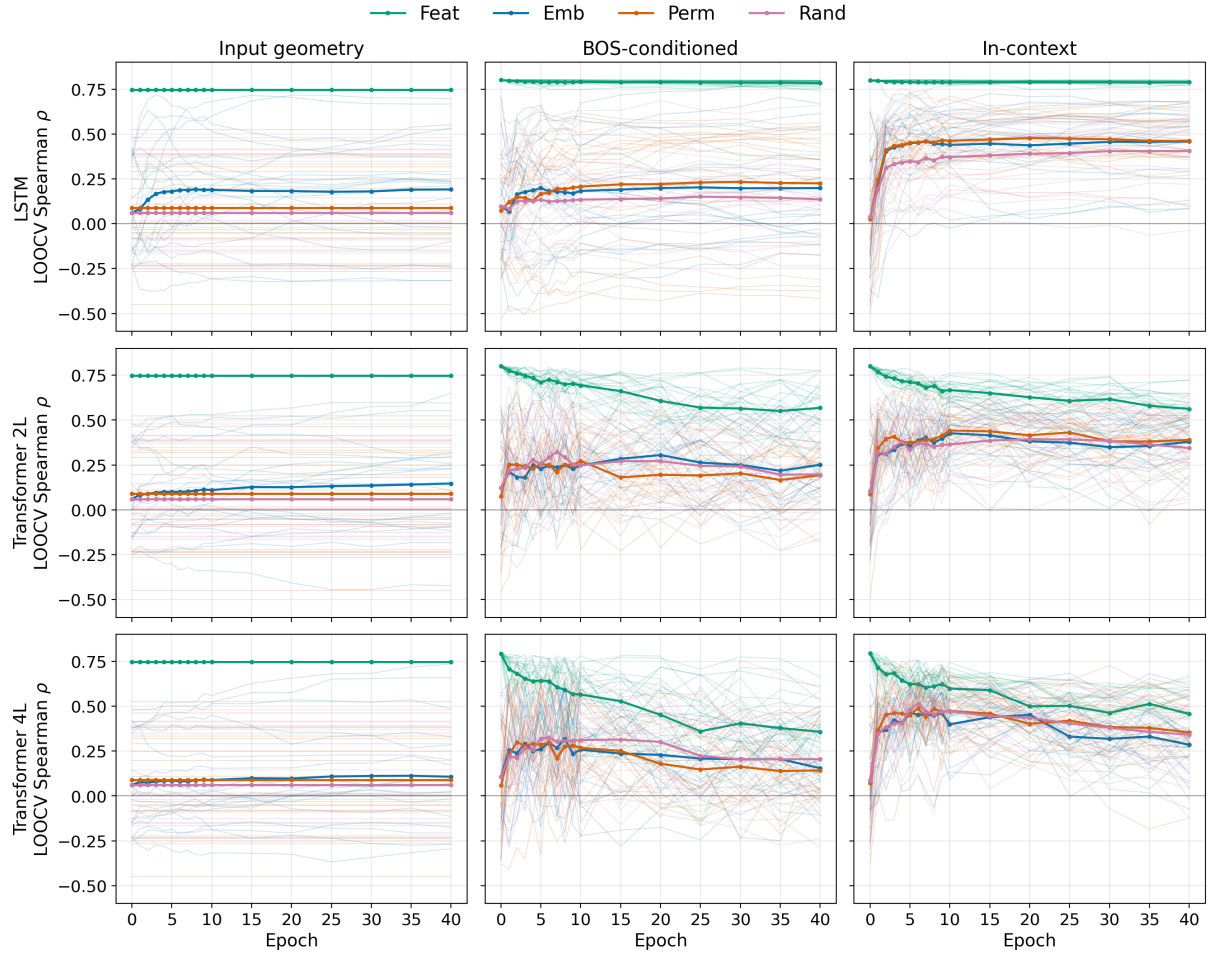


Figure 14: Consonant-hierarchy (CONSONANTS) recoverability across training; layout as in Fig. 12. The feature-input advantage persists in every architecture, while the later decline in context appears under FEAT at two layers and under all input conditions at four.