

When Symbolic Analyses Mislead: Resolution Sensitivity in Neural Language Representations

Matthew Fried
Touro University

Abstract

Recent representation analyses have reported structure-like or invariant behavior even in randomly initialized transformers, motivating the interpretation that aspects of linguistic structure may be architectural rather than learned. We show that such conclusions can be artifacts of coarse symbolic discretization. We formalize symbolic analysis of neural representations as operating over discretized activation trajectories and identify symbolic resolution as a critical, previously uncontrolled variable. Varying resolution reveals a sharp collapse of apparent similarity in random models, while pretrained models maintain stable, operator-selective differentiation across all tested resolutions. We introduce resolution stability as a necessary validity criterion for symbolic and automata-based analyses, providing a falsifiable, architecture-agnostic diagnostic for auditing internal linguistic structure in neural language models.

1 Introduction

Neural language models achieve strong performance across a wide range of linguistic tasks, yet the internal organization of their representations remains only partially understood (Belinkov and Glass, 2019; Rogers et al., 2020). Behavioral evaluations reveal what models can do, but provide limited insight into how linguistic structure is encoded internally, how grammatical distinctions are preserved or collapsed, or how learning reshapes internal computation (Elazar et al., 2021).

To address this gap, a growing body of work applies symbolic, geometric, and automata-based analyses to hidden representations, reporting evidence of grammatical structure at specific layers or apparent invariance under perturbation. A common but largely implicit assumption underlying these analyses is that discretization or similarity-based abstraction preserves the qualitative conclusions of the analysis, that is, that observed structural pat-

terns reflect properties of the underlying representation rather than artifacts of the analysis resolution. Under this assumption, similar patterns observed in trained and untrained models are often interpreted as suggesting that aspects of linguistic structure may arise independently of task-driven learning.

This assumption has not been systematically tested. Prior work has reported nontrivial structure-like behavior in settings that minimize or omit task-driven learning, including algorithmic behavior in randomly initialized or partially trained transformers (Zhong and Andreas, 2024), interpretable latent features extracted from random transformers via sparse autoencoders (Heap et al., 2025), and findings that untrained architectures can partially approximate the representational geometry of trained models under certain similarity metrics (Schrimpf et al., 2021). While these works do not claim full grammatical invariance, they rely on discretized or similarity-based analyses that implicitly treat resolution as a benign implementation detail.

We show that this assumption is unwarranted. Conclusions drawn from symbolic analyses can be critically sensitive to resolution: coarse symbolic abstraction can inflate similarity and create the appearance of grammatical invariance even in randomly initialized models, while finer resolution reveals a sharp collapse to a null baseline. We formalize this effect by viewing internal activations as symbolic trajectory languages and demonstrate that valid claims of representational invariance must be stable under refinement of symbolic resolution. Our central result is that symbolic trajectory analysis exhibits a resolution-induced collapse in random representations, but not in pretrained ones. Pretrained models maintain stable, operator-selective differentiation across resolutions, reflecting a learning-induced breaking of discretization symmetry that we term *structural selectivity*. By introducing resolution stability as a necessary validity criterion, we identify a previously unrecog-

nized measurement confound and provide a falsifiable framework for internal analysis of linguistic representations.¹

2 Related Work

Structural probing and linguistic localization.

A substantial literature studies whether contextual word representations encode syntactic or semantic structure using supervised probes or auxiliary objectives, such as dependency depth, tree distance, or priming (Sinclair et al., 2022). Structurally informed pretraining further suggests that grammatical sensitivity is learned rather than architectural: incorporating parsing objectives during pretraining systematically reshapes internal representations (Vilares et al., 2020). While structural probes and layerwise analyses demonstrate that linguistic information is *decodable* from internal states (Hewitt and Manning, 2019; Tenney et al., 2019; Jawahar et al., 2019; Hewitt and Liang, 2019), probing evaluates recoverability rather than whether structure-preserving transformations induce equivalent internal trajectories, limiting its use for testing grammatical invariance.

Representation similarity and layerwise diagnostics. Representational similarity measures such as Centered Kernel Alignment (CKA) compare the geometry of neural representations across layers, models, and training dynamics (Kornblith et al., 2019). However, purely geometric similarity can be confounded by input structure and does not distinguish meaningful abstraction from collapse or artifact; deconfounded analyses show that naive similarity metrics can be misleading (Cui et al., 2022). Consequently, similarity alone cannot adjudicate whether observed alignment reflects genuine invariance or discretization-induced effects. We note that CKA and related continuous similarity measures are not themselves discretization-dependent; we discuss them because they illustrate a broader pattern in which abstraction choices made during analysis can drive conclusions, of which symbolic resolution is the discrete instance we study here.

Paraphrase consistency and robustness. Behavioral robustness evaluations test output stability under paraphrasing, perturbations, or shuffling (Taktasheva et al., 2021). Output consistency, however, is neither necessary nor sufficient for representational coherence, as identical predictions may arise from

divergent internal pathways. Our analysis targets internal dynamics directly, independent of output behavior.

Automata and finite-state abstractions of neural models. Prior work extracts automata from neural models to induce grammars or interpretable sequence abstractions (Weiss et al., 2018; Omlin and Giles, 1996). We though use automata as *diagnostic proxies* for internal representational dynamics, where linguistic operators define equivalence classes that a coherent representation should respect.

Syntactic perturbations and representational sensitivity. Controlled syntactic perturbations have been used to study hierarchical structure in pre-trained models (Ding et al., 2016). Unlike prior analyses based on continuous similarity or probing, our framework explicitly tests whether inferred structure is stable under symbolic abstraction and resolution changes.

Summary. Across probing, similarity analysis, robustness evaluation, and automata extraction, discretization is typically treated as a neutral preprocessing step. We show that this assumption is false: coarse symbolic resolutions can *create* apparent invariance even in randomly initialized models. As symbolic resolution is refined, random-init overlap collapses toward the null baseline, while pre-trained differentiation remains stable. By introducing resolution-controlled symbolic abstractions and null baselines, we provide a falsification criterion that prevents over-interpreting discretized signals as evidence of grammatical structure.

3 Resolution Dynamics and Structural Selectivity

We formalize symbolic analysis of neural language model representations and specify how it is instantiated empirically. Our goal is to characterize how discretization resolution governs symbolic similarity and to distinguish artifacts of coarse abstraction from learning-induced structure. Throughout, we use *phase transition* to denote a sharp qualitative change in measured symbolic similarity as a continuous resolution parameter is varied, rather than a thermodynamic notion; the term is used analogously in recent studies of deep learning dynamics (Anagnostopoulos et al., 2026; Aoyama and Wilcox, 2025).

Symbolic trajectories and resolution. Let f_θ be a neural language model with L layers, and let

¹Code and configuration files for all experiments are available at https://github.com/MatthewFried/ASC_EMNLP.

$\mathbf{h}_\ell(x) \in \mathbb{R}^d$ denote the hidden state at layer ℓ for input sentence x . A forward pass induces a continuous layerwise activation trajectory

$$\tau(x) = (\mathbf{h}_1(x), \dots, \mathbf{h}_L(x)).$$

Symbolic analysis proceeds by discretizing each layerwise state via a quantization map

$$Q_m : \mathbb{R}^d \rightarrow \Sigma_m,$$

where m controls the resolution of the finite alphabet Σ_m . The resulting symbolic trajectory

$$\hat{\tau}_m(x) = (Q_m(\mathbf{h}_1(x)), \dots, Q_m(\mathbf{h}_L(x)))$$

defines a word in a formal language $\mathcal{L}_m(f_\theta)$ induced by the model at resolution m . We associate each m with an effective symbolic entropy $H(m)$, measured empirically as the support size or entropy of Σ_m over observed trajectories.

Symbolic similarity and null baselines. Let o be a linguistic operator mapping sentences to sentences (e.g., subject–auxiliary inversion). Given paired inputs $(x, o(x))$, we define symbolic similarity

$$S_m(o) = \mathbb{E}[\text{sim}(\hat{\tau}_m(x), \hat{\tau}_m(o(x)))],$$

where sim may be BOW-F1, Jaccard similarity, or edit-based similarity over symbolic trajectories. We define the null baseline S_m^{null} as the expected similarity between randomly paired trajectories under the same resolution. And we identify the critical resolution for a fixed operator o and model f_θ , as

$$m_c(o, f_\theta) = \inf\{m : S_m(o) \leq S_m^{\text{null}} + \epsilon\},$$

for a small tolerance ϵ . Intuitively, m_c marks the point beyond which any apparent invariance collapses to the random baseline.

Resolution collapse in random representations.

We propose the following two arguments on resolution collapse and structural sensitivity. We emphasize that both Claims are empirical, operational statements grounded in observed high-dimensional discretization behavior; they are supported by the heuristic justifications below and validated experimentally in Section 5, and are not formal theorems.

Claim 1 (Resolution-Induced Collapse). *For randomly initialized networks with high-dimensional hidden states, there exists a finite critical resolution m_c such that for all operators o ,*

$$\lim_{m \rightarrow \infty} S_m(o) = S_m^{\text{null}},$$

and m_c is independent of operator semantics.

The justification for this claim is that at coarse resolution, discretization induces collisions that dominate similarity scores, yielding inflated overlap regardless of operator structure. As resolution increases, collision probability decays rapidly with symbolic entropy due to concentration of measure in high-dimensional random representations. Because random networks lack operator-dependent structure, all operators converge uniformly to the null baseline.

Claim 2 (Learning-Induced Selectivity). *For pre-trained language models, critical resolutions differ across operator classes. In particular, for tested syntactic operators o_{gram} ,*

$$\forall m, \quad S_m(o_{\text{gram}}) \leq S_m^{\text{null}} + \epsilon,$$

while identity transformations remain maximally similar.

This says that learning selectively breaks discretization-induced equivalence classes, stabilizing differentiation for linguistically meaningful transformations while preserving identity. We refer to this phenomenon as *structural selectivity*.

Thus, given a tokenized sentence $x = (x_1, \dots, x_T)$, we extract layerwise hidden states $H^\ell(x) \in \mathbb{R}^{T \times d}$ for $\ell = 1, \dots, L$. To obtain fixed-length trajectories, we apply a pooling operator to each layer, yielding

$$\tau(x) = (h^1(x), \dots, h^L(x)) \in \mathbb{R}^{L \times d},$$

where $h^\ell(x) \in \mathbb{R}^d$ is CLS pooling unless otherwise noted. Mean pooling yields qualitatively identical results. We then do an alphabetization wherein each pooled vector $h^\ell(x)$ is discretized via an alphabetizer α_ρ with resolution parameters $\rho = (m, b)$. We use a Top- m quantile alphabetizer: for each layer, the m coordinates with largest absolute magnitude are selected and assigned to one of b quantile bins based on background statistics. The resulting symbol encodes both index identity and bin assignment, and concatenating across layers yields a discrete trajectory $s(x) \in \Sigma^L$. Increasing m strictly increases symbolic entropy and reduces collision-induced similarity inflation.

We evaluate controlled linguistic operators by comparing $s(x)$ and $s(o(x))$. Identity serves as a mandatory sanity control and must yield perfect agreement. Subject–auxiliary inversion and passivization probe grammatical structure, while random shuffle and random pairing serve as structure-destroying negative controls that define the similarity floor.

Similarity metrics. Our primary metric is RAW symbolic similarity, by which we mean similarity computed directly on the discretized symbolic trajectories themselves, prior to (and independent of) any automaton compression. We report bag-of-words F1 over symbol multisets: given trajectories s and s' , let P and R denote the multiset precision and recall of shared symbols, i.e., $P = |s \cap s'|/|s'|$ and $R = |s \cap s'|/|s|$, where $|s \cap s'|$ counts symbols in the multiset intersection. Then

$$\text{BOWF1}(s, s') = \frac{2PR}{P + R},$$

and we also report strict positional match. All metrics induce nearly identical operator orderings.

We use automata only as secondary diagnostics, wherein discrete trajectories may optionally be compressed into DFAs using state-merging algorithms such as k -tails. Automata are not necessary for any of the main conclusions; they serve only to verify that the observed resolution sensitivity is not specific to a particular similarity metric or symbolic encoding. Because learned automata become sparse at high symbolic resolution, DFA-based metrics are treated as auxiliary and are not required for any core claim. Our primary validity criterion is resolution stability. Let ρ_2 denote a strictly finer symbolic resolution than ρ_1 . An invariance claim for an operator o is considered valid only if symbolic similarity remains stable under refinement and remains well separated from random baselines. Specifically, similarity for o must not collapse as resolution increases and must retain substantial lift above similarity observed under random pairing. Any effect that disappears under refinement is classified as a discretization artifact rather than evidence of learned structure.

4 Experimental Setup

Models. We evaluate both encoder-based and decoder-only transformer architectures: BERT-base-uncased, RoBERTa-base, and GPT-2 (small), each as a pretrained model and as a randomly initialized control with identical depth, hidden size, and attention configuration. Unless otherwise noted, main-text tables and figures report BERT-base-uncased; cross-model replication for RoBERTa and GPT-2 is reported in Section 5. All analyses operate on internal hidden-state trajectories and do not depend on task-specific heads or output logits. In particular, the standard HuggingFace warning that RoBERTa’s pooler weights are

newly initialized is irrelevant here, as trajectory extraction uses only hidden states and never the pooler head.

Data. Naturalistic sentences are drawn from the SST-2 training split of the GLUE benchmark (Socher et al., 2013; Wang et al., 2018), loaded via the HuggingFace datasets library; subsamples of size N are drawn with a fixed random seed. Grammatical alternations are instantiated as hand-constructed minimal-pair templates, e.g., “She can swim” $\leftrightarrow$ “Can she swim” (subject–auxiliary inversion) and “The cat chased the mouse” $\leftrightarrow$ “The mouse was chased by the cat” (passivization). Structure-destroying and noise operators (random shuffle, block shuffle, typographical noise) are applied programmatically to the sampled sentences. The full template lists and construction procedure are included in the released code.

Trajectory extraction and pooling. For an input sentence x , we extract layerwise hidden states $H_\ell(x) \in \mathbb{R}^{T \times d}$ and construct a fixed-length trajectory

$$\tau(x) = (h_1(x), \dots, h_L(x)) \in \mathbb{R}^{L \times d}$$

using CLS pooling unless otherwise stated (final-token pooling for GPT-2). Mean pooling yields qualitatively identical results.

Symbolic discretization. Hidden states are discretized using a Top- m quantile alphabetizer with $m \in \{2, 4, 8, 16\}$ and $b = 32$ bins unless otherwise noted. Increasing m strictly increases symbolic entropy and reduces collision-induced similarity. As alternatives to the Top- m quantile alphabetizer, we also evaluate simhash and k -means codebook discretizers (Section 5), which exhibit the same pretrained–random separation and confirm that the effect is not specific to a single encoding method.

Operators and controls. We evaluate controlled linguistic transformations including identity, subject–auxiliary inversion, passivization, paraphrase, synonym substitution, negation, argument swap, random shuffle, block shuffle, and typographical noise. Mandatory negative controls include random sentence pairing and randomly initialized models, which define the similarity floor.

Similarity metrics. As noted, our primary metric is RAW symbolic similarity computed directly on discrete trajectories using bag-of-words F1. Jaccard and normalized edit similarity are reported as secondary checks and induce nearly identical operator orderings (Spearman $\rho > 0.98$).

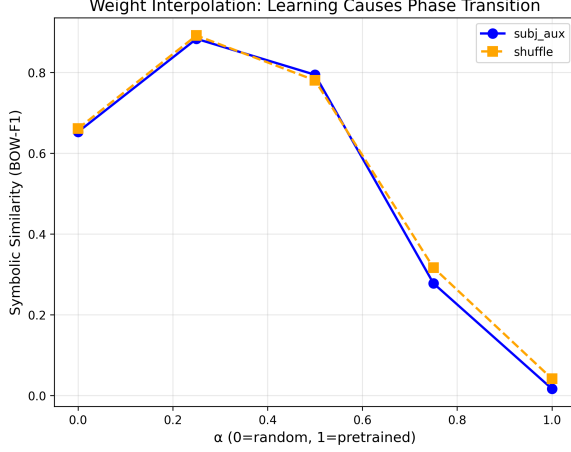


Figure 1: Linear weight interpolation $\theta_\alpha = (1 - \alpha)\theta_{\text{rand}} + \alpha\theta_{\text{pre}}$ between random-init ($\alpha = 0$) and pretrained ($\alpha = 1$) BERT-base. Symbolic similarity decreases overall from ≈ 0.65 to ≈ 0.02 , though not monotonically: the peak at $\alpha = 0.25$ reflects the loss barrier region. This demonstrates that learning (and not architectural properties) causes the differentiation effect.

Ablations and replication. We perform ablations over discretization resolution (m), automata horizon (k -tails $\in \{2, 3, 5, 7\}$), dataset size ($N \in \{500, 1000, 2000\}$), pooling strategy (CLS vs. mean), and model architecture (BERT, RoBERTa, GPT-2). Results are averaged over operator instances, with multi-seed statistics reported where applicable.

5 Numerical Analysis

We analyze symbolic trajectory similarity under controlled discretization, focusing on selectivity, stability, and falsification via negative controls. Unless otherwise stated, experiments use 1000 sentences, bins = 32, Top- $m = 8$, k -tails = 3, CLS pooling, and BERT-base-uncased.

Operator Taxonomy at Fixed Resolution. Table 1 summarizes operator behavior at a representative resolution. Identity yields perfect symbolic agreement for both pretrained and random-initialized models, validating the extraction and discretization pipeline. Across all non-identity operators, pretrained models exhibit near-zero similarity for grammatical alternations (subject-auxiliary inversion, passive) and for semantic and lexical perturbations. In contrast, randomly initialized models show nontrivial symbolic overlap across operator classes and fail to distinguish grammatical operators from structure-destroying controls such as random and block shuffle, indicating a lack of se-

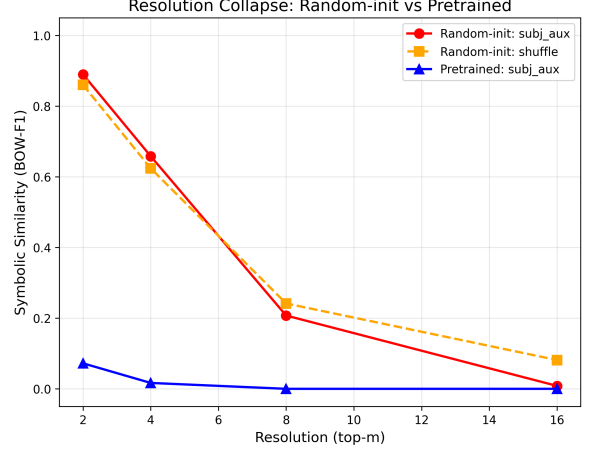


Figure 2: Resolution collapse in random-init vs. pretrained BERT-base. Random-init similarity for subject-auxiliary inversion (red) collapses from 0.89 to 0.01 as resolution increases, converging to the shuffle baseline (orange). Pretrained models (blue) remain near zero across all resolutions, demonstrating resolution-stable differentiation.

lectivity. Figure 1 confirms that this separation is induced by learning: linearly interpolating between random and pretrained weights produces a continuous transition in symbolic similarity.

Resolution Collapse and Selectivity. The central empirical result is shown in Table 2. As discretization resolution increases via Top- m , symbolic entropy rises monotonically. In randomly initialized models, subject-auxiliary similarity collapses sharply (0.876 at $m = 2$ to 0.013 at $m = 16$), converging to the random-shuffle baseline and revealing coarse-resolution invariance as a discretization artifact (Figure 2). In contrast, pretrained models exhibit resolution-stable behavior: similarity is already near zero at coarse resolution and remains so under refinement. This divergence yields a falsification criterion; that is, effects that collapse under refinement and fail to separate grammatical from non-grammatical operators cannot be attributed to learned structure.

Automata horizon and bin ablations. Tables 3 and 4 show that the qualitative contrast between pretrained and randomly initialized models is insensitive to automata horizon (k -tails) and discretization bin count. Across all tested settings, pretrained models maintain stable grammatical differentiation, while randomly initialized models remain non-selective relative to shuffle controls. These results indicate that the observed effects are driven by symbolic resolution (entropy), rather than automaton

Operator	Class	Pre-BOW	Pre-Jacc	Pre-Edit	Rand-BOW	Rand-Jacc	Spec
argument_swap	semantic	0.000	0.000	0.000	0.300	0.181	0.291
subj_aux	syntactic	0.000	0.000	0.000	0.208	0.121	0.199
random_shuffle	position	0.033	0.033	0.033	0.178	0.115	0.135
block_shuffle	position	0.333	0.333	0.333	0.453	0.401	0.110
typo	noise	0.000	0.000	0.000	0.067	0.036	0.058
paraphrase	paraphrase	0.000	0.000	0.000	0.067	0.038	0.057
synonym	lexical	0.000	0.000	0.000	0.044	0.024	0.035
passive	syntactic	0.000	0.000	0.000	0.030	0.016	0.021
identity	control	1.000	1.000	1.000	1.000	1.000	-0.009
negation	semantic	0.000	0.000	0.000	0.000	0.000	-0.009

Table 1: Operator taxonomy for BERT-base at bins=32, top- m =8, k =3, 1000 sentences. We report similarity (BOW-F1, Jaccard, EditSim). Lower values indicate stronger separation. Spec > 0 means the pretrained model separates (has lower similarity) more than random-init.

top- m	Pre-Ent	Rand-Ent	Pre SubjAux	Rand SubjAux	Rand Shuffle	SelectGap
2	13.55	5.14	0.072	0.876	0.851	0.025
4	13.55	7.50	0.017	0.652	0.614	0.037
8	13.55	10.01	0.000	0.213	0.260	-0.047
16	13.55	10.23	0.000	0.013	0.110	-0.097

Table 2: Resolution collapse under coarse symbolic encoding (BERT-base). We vary the discretization resolution via the top- m per-layer quantile alphabet; lower similarity indicates stronger separation. Random-init exhibits inflated similarity at low resolution that collapses under refinement, while pretrained remains near-zero across all resolutions. Entropy columns quantify effective trajectory symbol complexity.

Discretizer	Model	SubjAux	Shuffle	Gap (SubjAux–Shuffle)
simhash	pretrained	0.419	0.397	0.022
simhash	random-init	0.819	0.836	-0.017
kmeans_codebook	pretrained	0.450	0.458	-0.008
kmeans_codebook	random-init	0.844	0.833	0.011

Table 3: Discretizer robustness (BERT-base). SubjAux similarity (BOW-F1) is reported for pretrained vs. random-init under three discretizers. Random-init exhibits high similarity (non-selective collapse), while pretrained is substantially lower (greater differentiation). The effect persists beyond Top- m .

Interpretation: Across simhash and k-means codebooks, random-init yields high SubjAux similarity (≈ 0.82 – 0.84) while pretrained yields substantially lower similarity (≈ 0.42 – 0.45), demonstrating that the learned differentiation effect is not specific to Top- m encoding. Negative gap values in the random-init regime reflect noise-level fluctuations and do not indicate selectivity; both SubjAux and Shuffle remain uniformly high.

α	SubjAux	Passive	Shuffle	Null
0.00	0.633	0.327	0.617	0.153
0.25	0.936	0.753	0.933	0.640
0.50	0.814	0.533	0.800	0.255
0.75	0.386	0.203	0.403	0.058
1.00	0.017	0.000	0.042	0.000

Table 4: Weight interpolation results (BERT-base). We linearly interpolate between random-init and pretrained weights, $\theta_\alpha = (1 - \alpha)\theta_{\text{rand}} + \alpha\theta_{\text{pre}}$ with $\alpha \in [0, 1]$, and measure symbolic similarity. SubjAux similarity decreases sharply as $\alpha \rightarrow 1$, indicating that learning induces differentiation rather than architectural collapse.

Note: “Null” is the random-pairing baseline at each α . Identity remained 1.000 throughout (omitted for space).

depth or binning granularity.

Scaling with Dataset Size. We report (see Appendix) results for N from 500-7K, wherein pre-

trained models consistently maintain near-zero similarity for grammatical alternations across all dataset sizes. Random-initialized models exhibit variability in absolute overlap magnitude but continue to lack selectivity between grammatical and non-grammatical operators.

Multi-Seed Robustness and Effect Sizes. Multi-seed evaluation (as reported in the Appendix) confirms that the separation between pretrained and random-initialized regimes is not driven by seed variance. Across Top- $m \in 2, 4, 8, 16$, identity remains invariant (BOW-F1 = 1.0), while the tested syntactic operators exhibit large pretrained–random separation relative to standard deviations (often $|d| > 2$). Crucially, the random-init apparent invariance at coarse resolution collapses under refinement, converging toward the null baseline. Identity

remains invariant across all conditions, while the tested syntactic operators consistently show strong differentiation in pretrained models and resolution-dependent collapse in random-initialized ones.

Cross-Model Replication. Cross-model replication is consistent across both masked-LM encoder families and decoder-only architectures. At bins = 32, Top- m = 8, k = 3 (1000 sentences), both BERT-base and RoBERTa-base exhibit near-zero pretrained similarity on syntactic operators (SubjAux and Passive), while their randomly initialized counterparts show substantial overlap under the same discretization (e.g., BERT random-init SubjAux = 0.208 vs. pretrained = 0.000; RoBERTa random-init SubjAux = 0.042 vs. pretrained = 0.000). GPT-2 displays the same qualitative separation despite its decoder-only architecture: pretrained GPT-2 yields near-zero similarity for SubjAux and Passive across resolutions (SubjAux = 0.000 at Top- m = 8), while the random-initialized GPT-2 control exhibits elevated overlap that collapses sharply with increasing resolution (e.g., SubjAux Rand = 0.422 at Top- m = 8, dropping to 0.006 at Top- m = 16). Across all three model families, random-init models fail to distinguish grammatical operators from random shuffle, whereas pretrained models maintain resolution-stable differentiation. This consistency indicates that the observed effect is not tied to a specific pretraining objective or encoder-only architecture, but reflects a learning-induced restructuring of internal trajectory space. Notably, the effect does not rely on bidirectional context, masked-token objectives, or the presence of a dedicated classification token, as GPT-2 lacks all three yet exhibits the same resolution-dependent collapse in the random-init regime and stability under pretraining.

Training Trajectory Diagnostic (Fixed Discretizer). To isolate training-path effects independently of symbolic resolution, we evaluate symbolic similarity along the training trajectory while holding the discretizer fixed. Specifically, we track a model trained from random initialization at multiple checkpoints, using an alphabetizer calibrated on pretrained statistics throughout. We show (as reported in the Appendix) that symbolic similarity increases rapidly during training under this fixed reference frame. This behavior reflects a shift in symbolic alignment relative to the pretrained-calibrated alphabet. Note that this diagnostic fixes the discretizer to a pretrained-calibrated alphabet, whereas the phase-transition results vary symbolic

resolution. The two experiments therefore probe distinct axes (training-path dynamics vs. resolution stability) and are not expected to be monotone-consistent.

Summary. Across operators, resolutions, seeds, pooling strategies, automata horizons, dataset sizes, and model architectures, the same empirical pattern emerges. At coarse symbolic resolution, randomly initialized transformers produce elevated symbolic similarity across all operators, including grammatical alternations and structure-destroying controls. As resolution is refined, this inflated overlap collapses toward the random baseline, revealing coarse-resolution invariance as a discretization artifact. Pretrained models behave qualitatively differently, maintaining stable and selective differentiation of grammatical operators across all tested resolutions. These results establish resolution stability as a necessary empirical condition for interpreting symbolic similarity measurements.

6 Conclusion

We showed that symbolic analyses of neural language model representations are critically sensitive to discretization resolution, an assumption left untested in prior work. Varying resolution reveals a sharp phase transition in randomly initialized transformers: inflated similarity observed at coarse abstraction collapses under refinement to a null baseline, demonstrating that apparent grammatical invariance in untrained models is a measurement artifact rather than architectural structure.

Pretrained language models behave qualitatively differently. Across resolutions, architectures, and ablations, they maintain stable and selective differentiation of grammatical operators from identity and structure-destroying controls, indicating that learning reshapes internal trajectory structure rather than inducing uniform invariance.

Based on this contrast, we introduce *resolution stability* as a necessary falsifiability criterion for symbolic and automata-based analyses. Any claimed structural invariance that disappears under discretization refinement cannot be attributed to learned linguistic structure. Our results identify symbolic resolution as a central confound in internal representation analysis and provide an architecture-agnostic diagnostic for evaluating claims of grammatical structure in neural language models.

Limitations

This work analyzes internal neural representations using symbolic and automata-based diagnostics under controlled discretization. As a result, its conclusions apply specifically to analyses that rely on explicit symbolic abstraction of continuous activations and do not address all possible representational methodologies. Experiments focus on transformer encoder architectures and a fixed set of linguistic operators; broader linguistic coverage remains future work. While multiple resolutions, seeds, and three pretrained model families are evaluated, results may not directly generalize to other architectures, training objectives, or modalities. At high symbolic resolution and limited sample sizes, automata-based metrics can suffer from sparse trace coverage. Finally, the proposed resolution-stability criterion is diagnostic rather than explanatory and does not by itself characterize the mechanisms by which linguistic structure is learned.

AI Assistance Acknowledgment

The author(s) used AI-based tools to assist with code development, result aggregation, table generation, and editorial refinement of the manuscript. All experimental design, implementation decisions, analysis, interpretation of results, and final presentation choices were made by the author(s), who takes full responsibility for the content of the paper.

References

- Sokratis J. Anagnostopoulos, Juan Diego Toscano, Nikolaos Stergiopoulos, and George Em Karniadakis. 2026. Learning in PINNs: Phase transition, diffusion equilibrium, and generalization. *Neural Networks*.
- Tatsuya Aoyama and Ethan Wilcox. 2025. [Language models grow less humanlike beyond phase transition](#). In *Proceedings of the 63rd Annual Meeting of the ACL (Volume 1: Long Papers)*, pages 24938–24958. ACL.
- Yonatan Belinkov and James Glass. 2019. Analysis methods in neural language processing: A survey. *TACL*, 7:49–72.
- Tianyu Cui, Yogesh Kumar, Pekka Marttinen, and Samuel Kaski. 2022. Deconfounded representation similarity for comparison of neural networks. In *Advances in Neural Information Processing Systems*, volume 35.
- Nai Ding, Lucia Melloni, Hang Zhang, Xing Tian, and David Poeppel. 2016. Cortical tracking of hierarchical linguistic structures in connected speech. *Nature Neuroscience*, 19:158–164.
- Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. 2021. [Amnesic probing: Behavioral explanation with amnesic counterfactuals](#). *TACL*, 9:160–175.
- Thomas Heap, Tim Lawson, Lucy Farnik, and Laurence Aitchison. 2025. [Sparse autoencoders can interpret randomly initialized transformers](#). *arXiv preprint arXiv:2501.17727*.
- John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. In *EMNLP*.
- John Hewitt and Christopher Manning. 2019. A structural probe for finding syntax in word representations. In *NAACL*.
- Ganesh Jawahar, Benoit Sagot, and Djamé Seddah. 2019. What does bert learn about the structure of language? In *ACL*.
- Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. 2019. Similarity of neural network representations revisited. In *Proceedings of the 36th ICML*, pages 3519–3529. PMLR.
- Christian W. Omlin and C. Lee Giles. 1996. Extraction of rules from discrete-time recurrent neural networks. *Neural Networks*, 9(1):41–52.
- A. Rogers, O. Kovaleva, and A. Rumshisky. 2020. [A primer in bertology: What we know about how bert works](#). *TACL*, pages 842–866.
- Martin Schrimpf, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A Hosseini, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. 2021. [The neural architecture of language: Integrative modeling converges on predictive processing](#). *Proceedings of the National Academy of Sciences*, 118(45):e2105646118.
- A. Sinclair, J. Jumelet, W. Zuidema, and R. Fernández. 2022. [Structural persistence in language models: Priming as a window into abstract language representations](#). *TACL*, 10:1031–1050.
- Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. [Recursive deep models for semantic compositionality over a sentiment treebank](#). pages 1631–1642.
- Ekaterina Taktasheva, Vladislav Mikhailov, and Ekaterina Artemova. 2021. [Shaking syntactic trees on the sesame street: Multilingual probing with controllable perturbations](#). In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 191–210. ACL.
- Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. Bert rediscovers the classical nlp pipeline. In *ACL*.

- D. Vilares, M. Strzyz, A. Søgaard, and C. Gómez-Rodríguez. 2020. [Parsing as pretraining](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 34:9114–9121.
- Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. [GLUE: A multi-task benchmark and analysis platform for natural language understanding](#). In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium.
- Gail Weiss, Yoav Goldberg, and Eran Yahav. 2018. Extracting automata from recurrent neural networks using queries and counterexamples. In *ICML*.
- Ziqian Zhong and Jacob Andreas. 2024. [Algorithmic capabilities of random transformers](#). In *Advances in Neural Information Processing Systems 37 (NeurIPS)*.

A Appendix

This appendix reports supplementary analyses that support robustness and replication but are not required for establishing the core resolution-stability results in the main paper. These include scaling behavior with larger sample sizes, multi-seed statistics and effect sizes, and a training-trajectory diagnostic probing learning dynamics under a fixed discretization. All appendix results are consistent with the main findings and are provided to facilitate reproducibility and deeper inspection.

The complete experimental pipeline, including configuration files, minimal-pair templates, and evaluation code sufficient to reproduce all reported results, is available at https://github.com/MatthewFried/ASC_EMNLP.

N	Pre SubjAux	Rand SubjAux	Rand Shuffle
500	0.000	0.242	0.256
1000	0.000	0.214	0.206
2000	0.000	0.283	0.219
3000	0.000	0.219	0.223
4000	0.000	0.196	0.241
5000	0.006	0.206	0.237
6000	0.000	0.209	0.220
7000	0.003	0.187	0.243

Table 5: Scaling with dataset size (BERT-base, bins=32, Top- m =8). Pretrained grammatical similarity remains near zero across all N , while random-initialized similarity tracks shuffle, indicating stability with respect to sample size.

Operator	Top- m	Pretrained	Random-init	Cohen’s d
identity	2	1.000 $\pm$ 0.000	1.000 $\pm$ 0.000	0.00
identity	4	1.000 $\pm$ 0.000	1.000 $\pm$ 0.000	0.00
identity	8	1.000 $\pm$ 0.000	1.000 $\pm$ 0.000	0.00
identity	16	1.000 $\pm$ 0.000	1.000 $\pm$ 0.000	0.00
passive	2	0.037 $\pm$ 0.000	0.710 $\pm$ 0.063	-2.98
passive	4	0.000 $\pm$ 0.000	0.311 $\pm$ 0.087	-4.08
passive	8	0.000 $\pm$ 0.000	0.027 $\pm$ 0.021	-5.51
passive	16	0.000 $\pm$ 0.000	0.000 $\pm$ 0.000	-2.73
random shuffle	2	0.149 $\pm$ 0.056	0.851 $\pm$ 0.024	0.00
random shuffle	4	0.108 $\pm$ 0.058	0.614 $\pm$ 0.059	0.00
random shuffle	8	0.107 $\pm$ 0.057	0.260 $\pm$ 0.056	0.00
random shuffle	16	0.107 $\pm$ 0.057	0.110 $\pm$ 0.057	0.00
subj-aux	2	0.072 $\pm$ 0.000	0.876 $\pm$ 0.032	0.88
subj-aux	4	0.017 $\pm$ 0.000	0.652 $\pm$ 0.043	0.72
subj-aux	8	0.000 $\pm$ 0.000	0.213 $\pm$ 0.025	-1.09
subj-aux	16	0.000 $\pm$ 0.000	0.013 $\pm$ 0.008	-2.39

Table 6: Multi-seed robustness across resolutions (BERT-base; mean $\pm$ std, BOW-F1). Lower values indicate stronger separation. Cohen’s d compares pretrained vs. random-init overlap per operator at each Top- m .

Step	SubjAux	Shuffle	Gap (SubjAux–Shuffle)	Null
0	0.000	0.033	-0.033	0.000
100	0.611	0.578	0.033	0.283
500	0.825	0.836	-0.011	0.607
1000	0.811	0.794	0.017	0.585
2000	0.828	0.789	0.039	0.582

Table 7: Training trajectory diagnostic. Using a fixed alphabetizer fitted on the pretrained reference, we evaluate a model during training from random initialization. Similarity increases rapidly during training under this fixed discretization, indicating a strong shift in the induced symbolic overlap regime.

Note: This diagnostic fixes the discretizer to a pretrained-calibrated alphabet; unlike resolution-sweep experiments, it varies training dynamics rather than symbolic entropy.