Latent Readiness Before Generalization: Tracking Pre-Transition States in Compositional Language Learning

Anonymous ACL submission

Abstract

Neural sequence models can fit training data before they generalize to held-out compositional structures. We ask whether out-of-distribution gains appear only when task-relevant structure is first acquired, or whether internal preparation accumulates earlier and becomes behaviorally visible later. We call this pre-transition phenomenon **latent readiness**: a theoretical construct referring to measurable internal preparation that rises before corresponding gains appear in out-of-distribution behavior. We study this idea in two structured language-learning settings: COGS, which exhibits delayed compositional generalization, and CoNLL–SIGMORPHON 2017 Task 1 morphological reinflection, which is learned more gradually in our English and Turkish settings. On COGS, readiness indicators positively predict later out-of-distribution improvement in all three seeds, with the clearest case showing a readiness lead of 7k steps before sustained 50% generalization. In contrast, the SIGMORPHON conditions show negative predictive correlations and no comparable transition-like pattern. These results support a regime-sensitive view of latent readiness: in our experiments, readiness is most informative in the delayed, transition-like COGS setting rather than in the more gradual morphology conditions studied here.

1 Introduction

Neural sequence models do not always generalize as soon as they fit their training data. On structured language tasks, a model can achieve high in-distribution performance while still failing on held-out compositional or structural cases (Lake and Baroni, 2018; Kim and Linzen, 2020). This gap matters because fitting observed examples and acquiring reusable structure are not the same process (Lake and Baroni, 2018; Kim and Linzen, 2020; Wu et al., 2023).

Prior work has shown delayed generalization, abrupt performance gains, and transition-like learning behavior in neural networks (Power et al., 2022; Liu et al., 2022; Belkin et al., 2019; Nakkiran et al., 2021). These results suggest that **visible generalization** may lag behind internal change. Yet most analyses focus on *when* performance improves, not on *what state* the model is in before improvement becomes visible. It therefore remains unclear whether out-of-distribution generalization appears only when relevant structure is first acquired, or whether internal preparation accumulates earlier and is expressed later.

We address this question through **latent readiness**: a theoretical construct referring to a measurable **pre-transition state** in which task-relevant internal preparation improves before corresponding gains appear in out-of-distribution behavior. Under this view, generalization onset need not mark the first appearance of relevant organization. It may instead mark the point at which accumulated preparation becomes behaviorally accessible.

We study latent readiness in two settings. First, we use COGS, a benchmark for compositional generalization in semantic interpretation (Kim and Linzen, 2020). COGS is useful here because it separates strong in-distribution performance from harder held-out generalization cases, allowing us to test whether internal preparation improves while exact-match generalization remains low (Kim and Linzen, 2020; Wu et al., 2023). Second, we contrast this setting with morphological reinflection using CoNLL–SIGMORPHON 2017 Task 1 (Cotterell et al., 2017). Morphological reinflection provides a useful contrast because it is also a structured sequence transduction problem, but one in which progress may be expressed more continuously in output behavior. This makes it a good test of whether latent readiness is a broad feature of structured learning or is most informative when behavior passes through a delayed, transition-like

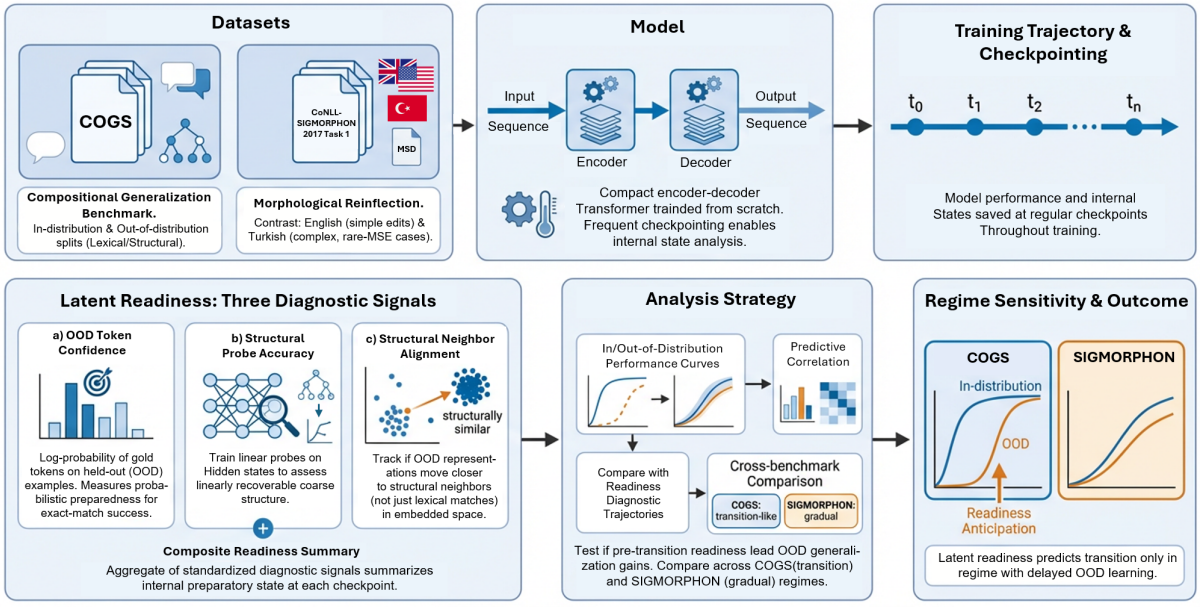

Figure 1: Overview of the latent readiness framework. We compare two structured language-learning settings, COGS and CoNLL–SIGMORPHON 2017 Task 1, using compact encoder–decoder Transformers trained from scratch with frequent checkpointing. At each checkpoint, we extract three diagnostic indicators—OOD token confidence, structural probe accuracy, and structural neighbor alignment—and test whether these indicators rise before visible out-of-distribution generalization. We then compare predictive readiness across a delayed, transition-like regime (COGS) and more gradual morphology settings (SIGMORPHON).

bottleneck. We study English and Turkish. Our goal is not one-to-one task equivalence, but a controlled contrast between a delayed, transition-like compositional regime and a more gradual structured learning regime.

To operationalize latent readiness, we use three indicators: token-level confidence on out-of-distribution targets, structural probes on hidden representations (Conneau et al., 2018; Hewitt and Manning, 2019; Tenney et al., 2019; Belinkov, 2022), and structural neighbor alignment in representation space. We then ask three questions. Does out-of-distribution performance show delayed or transition-like onset relative to in-distribution performance? Do readiness indicators rise before that onset becomes behaviorally visible? And is latent readiness broadly informative across structured learning settings, or mainly in particular regimes?

Contributions. This paper makes three contributions. First, we introduce **latent readiness**, a framework for analyzing internal preparation before observable out-of-distribution generalization. Second, we operationalize this construct with token-level confidence, structural probing, and representation-space neighbor alignment, while treating the resulting composite score as a practical proxy rather than a universal definition. Third, we show empirically that COGS exhibits delayed and asymmetric visible generalization together with positive predictive readiness, whereas the CoNLL–SIGMORPHON 2017 English and Turkish conditions studied here do not show the same pattern. Taken together, these results support a regime-sensitive interpretation: latent readiness is most informative in the delayed, transition-like setting we observe in COGS.

2 Related Work

2.1 Delayed Generalization and Transition-Like Learning

A growing literature studies cases in which neural networks fit training data well before they generalize to held-out examples (Power et al., 2022; Liu et al., 2022; Belkin et al., 2019; Nakkiran et al., 2021; Raghu et al., 2017; Morcos et al., 2018). These studies show that learning trajectories can be highly non-uniform, with long plateaus followed by abrupt gains (Power et al., 2022; Liu et al., 2022). Work on representation dynamics further suggests that important internal changes can precede or accompany later behavioral improvements (Raghu et al., 2017; Morcos et al., 2018; Kornblith et al., 2019). This motivates our premise that visible generalization may understate internal progress before a visible transition.

Our work shifts attention from delayed generalization itself to the state that precedes it. Rather than asking only whether generalization is delayed, we ask whether the pre-transition period forms a distinct and measurable state in which internal preparation accumulates before behavior changes.

2.2 Compositional Generalization in Language

Compositional generalization benchmarks test whether models can extend learned structure to unseen combinations (Lake and Baroni, 2018; Keysers et al., 2020; Kim and Linzen, 2020; Hupkes et al., 2023). These benchmarks are useful because strong in-distribution accuracy can coexist with brittle out-of-distribution behavior, making it possible to separate surface success from structural generalization (Lake and Baroni, 2018; Keysers et al., 2020; Kim and Linzen, 2020). Prior work has used datasets such as SCAN, CFQ, and COGS to diagnose failures of systematic or compositional generalization in sequence models and semantic parsers (Lake and Baroni, 2018; Loula et al., 2018; Keysers et al., 2020; Oren et al., 2020; Shaw et al., 2021). COGS is especially useful because it isolates held-out lexical and structural cases within a semantic interpretation task (Kim and Linzen, 2020), while later work has highlighted how benchmark design and logical-form choices affect conclusions in this area (Wu et al., 2023; Li et al., 2023).

We build on this line of work but change the target of analysis. Existing studies typically ask whether a model generalizes compositionally, which architectures help, or how benchmark design shapes the result (Oren et al., 2020; Shaw et al., 2021; Wu et al., 2023; Li et al., 2023). We instead ask what internal changes precede visible generalization, and whether a model may already be moving toward more task-relevant organization before its exact-match performance improves.

2.3 Representation Analysis and Behavioral Emergence

Many studies analyze hidden representations to understand what neural models encode during learning. Probing methods test whether specific properties are recoverable from internal states, while representation-space analyses examine how examples cluster, align, or compare geometrically (Conneau et al., 2018; Hewitt and Manning, 2019; Tenney et al., 2019; Belinkov, 2022; Raghu et al., 2017; Kornblith et al., 2019). These methods are useful because they can reveal internal structure even when output behavior does not change (Hewitt and Manning, 2019; Belinkov, 2022; Kornblith et al., 2019).

We draw on this methodological tradition but use it dynamically rather than statically. Instead of treating probes or representation geometry as fixed diagnostics, we track whether behavior-adjacent and representation-level indicators rise before observable generalization. This temporal perspective is central to latent readiness.

3 Methodology

3.1 Datasets

We study latent readiness in two structured language-learning settings: COGS and CoNLL–SIGMORPHON 2017 Task 1 morphological reinflection.

COGS. COGS is a benchmark for compositional generalization in semantic interpretation (Kim and Linzen, 2020). Each example pairs a natural language input with a target logical form. The dataset provides standard in-distribution splits and a dedicated generalization split containing held-out compositional patterns (Kim and Linzen, 2020). We use train, development, and test as in-distribution evaluation, and the generalization split as the main out-of-distribution target. For analysis, we further divide the generalization split into lexical and structural subsets. Lexical cases require extending familiar compositional behavior to held-out lexical items or substitutions. Structural cases require extending to held-out compositional configurations. We report exact-match accuracy on the full generalization split and on both subsets.

CoNLL–SIGMORPHON 2017 Task 1. As a contrastive benchmark family, we study universal morphological reinflection using English and Turkish from CoNLL–SIGMORPHON 2017 Task 1 (Cotterell et al., 2017). Each example consists of a lemma, a target MSD (morphosyntactic description), and a target inflected form. The model takes the lemma and target MSD as input and generates the target form. We evaluate both medium-resource and low-resource conditions, following the shared-task setup (Cotterell et al., 2017). Because the two languages differ in difficulty structure, we use language-specific analytical groupings rather than one shared decomposition. For both languages, we analyze edit-distance buckets defined by Lev-

enshtein distance between lemma and target form. For Turkish, we also analyze a hard subset defined by rare MSD patterns. This subset is meaningful in Turkish but not in English, where high-edit and rare-MSD cases are largely absent in the conditions we study. In the main text, we therefore focus on Turkish as the clearest contrastive morphology case, while using English as a supplementary comparison.

These settings are not fully matched in the sense of identical subset construction or task form. We use them instead as a contrastive family for learning-regime comparison: COGS provides a delayed, transition-like case, whereas the morphology settings provide more gradual trajectories under the conditions studied here. Dataset-specific preprocessing details, subset definitions, and split statistics are in Appendix A.

3.2 Model

For both tasks, we use small encoder–decoder Transformers trained from scratch. Our goal is not to maximize benchmark performance through scale or pretraining. Instead, we use compact settings in which learning dynamics can be observed throughout training. This supports dense checkpointing, repeated hidden-state extraction, and representation analysis across the full trajectory.

The model is trained with teacher forcing using token-level cross-entropy loss. We evaluate task performance at regular checkpoints and save hidden representations and prediction statistics for readiness analysis. Full architecture details, optimization hyperparameters, training schedules, and implementation notes are deferred to Appendix B.

3.3 Transition-Oriented Learning Analysis

Our goal is to study not only whether generalization occurs, but also what happens before it does. We therefore analyze the model at multiple checkpoints throughout training rather than relying only on final performance. At each checkpoint, we record in-distribution performance, out-of-distribution performance, and a set of readiness indicators.

For COGS, we define a transition point as the first checkpoint at which out-of-distribution exact-match accuracy crosses a fixed threshold and stays above that threshold for multiple subsequent checkpoints. For the morphology contrast, we apply the same procedure to the Turkish hard subset, but the resulting trajectories are much more gradual. We use these threshold-based transition points as operational markers rather than as a standalone definition of transition-like learning, and interpret them jointly with full trajectory shape, ID–OOD lag, subset asymmetry, and readiness lead-time patterns. Threshold sensitivity analyses are reported in Appendix E.

3.4 Operationalizing Latent Readiness

We define **latent readiness** as a theoretical construct referring to a measurable **pre-transition state** in which task-relevant internal preparation improves before such improvement is fully reflected in out-of-distribution performance. Because this construct is not directly observable, we approximate it using three complementary indicators that capture different aspects of the same underlying process. Token confidence provides a behavior-adjacent view of movement toward correct outputs, probes test whether task-relevant distinctions become more recoverable from hidden states, and neighbor alignment tracks whether representation geometry reorganizes around structure rather than superficial similarity. Together, these indicators allow us to ask whether readiness appears across multiple levels of analysis rather than in only one metric family.

OOD Token Confidence. Our first indicator measures token-level confidence on out-of-distribution targets. For each held-out example, we compute the teacher-forced log-probability of the gold target tokens and average across positions. This captures whether the model assigns more probability mass to correct outputs before sequence-level exact match becomes reliable.

Structural Probe Accuracy. Our second indicator measures the recoverability of structural information from hidden representations. At each checkpoint, we extract sentence-level or sequence-level encoder representations and train lightweight linear probes to predict coarse structural attributes. If probe accuracy improves before held-out performance does, this suggests that the model is organizing task-relevant distinctions before it can consistently express them behaviorally. We interpret this signal cautiously: higher probe accuracy indicates greater linear recoverability, not causal use of structure for prediction (Hewitt and Manning, 2019; Belinkov, 2022).

Structural Neighbor Alignment. Our third indicator measures changes in representation-space

geometry. For each held-out example, we identify nearest neighbors among training examples in encoder representation space and test whether those neighbors increasingly reflect structural similarity rather than superficial lexical overlap. This is a geometric diagnostic of reorganization rather than a mechanistic explanation.

Composite Readiness Score. To summarize these indicators, we compute a composite readiness score by standardizing each component across checkpoints and averaging them with equal weight. We use this as a practical proxy for the latent construct rather than as a universal measure of readiness. Additional details on feature extraction, probe targets, and normalization are provided in Appendix C.

3.5 Analysis Strategy

Our analysis centers on the temporal relationship between latent readiness and visible generalization. First, we compare the trajectories of in-distribution and out-of-distribution performance to determine whether a delayed generalization regime exists. Second, we track each readiness indicator across the same checkpoints and test whether these indicators rise before out-of-distribution performance does. Third, we test whether readiness at checkpoint t predicts improvements in held-out performance over the following checkpoints. Finally, we compare COGS and the SIGMORPHON conditions studied here to determine whether latent readiness is broadly informative across structured learning settings or instead becomes most informative in delayed, transition-like regimes.

4 Results

We report results from three seeds on both benchmark families. The results develop in four steps. First, COGS shows delayed and asymmetric visible generalization. Second, a pre-transition readiness pattern appears before that transition. Third, this pattern is predictively useful on COGS. Fourth, the same pattern does not appear in the SIGMORPHON conditions studied here.

4.1 COGS exhibits delayed and asymmetric visible generalization

Table 1 summarizes the main COGS results. In-distribution performance is near-saturated in all three seeds (0.998 max dev exact match), but out-of-distribution generalization remains much lower,

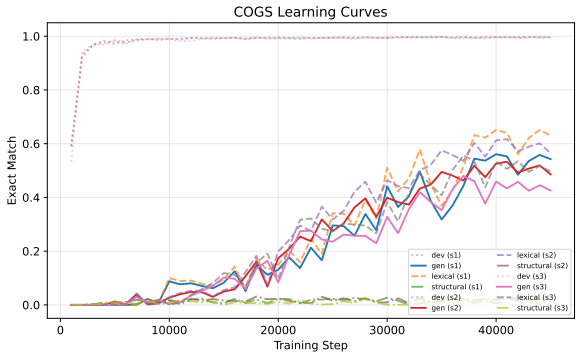

Figure 2: Learning trajectories on COGS. In-distribution exact match rises earlier than out-of-distribution generalization, while structural generalization remains the most delayed.

peaking between 0.480 and 0.561. The gap becomes sharper when the generalization split is decomposed into lexical and structural subsets. Exact-match accuracy on lexical generalization reaches 0.557–0.651, whereas structural generalization remains near floor at 0.024–0.031.

Figure 2 shows the underlying trajectory. In-distribution exact match rises quickly, whereas generalization improves much later. The delayed pattern is also asymmetric: lexical and structural generalization evolve at very different rates. Some out-of-distribution behavior improves while structural generalization remains largely blocked.

This establishes the empirical setting in which latent readiness is meaningful: COGS shows a long lag between in-distribution success and held-out success, together with a strong lexical–structural asymmetry.

4.2 A pre-transition readiness pattern appears before visible generalization on COGS

We next ask whether readiness indicators rise before behavioral generalization. Across all three seeds, readiness has a positive predictive correlation with future out-of-distribution improvement, with the clearest evidence in Seed 1. Table 2 reports the key quantities. Predictive correlations are positive in every seed, ranging from $+0.09$ to $+0.58$. In Seed 1, readiness crosses its 75th percentile about 7k steps before sustained 50% generalization.

Figure 3 shows the composite readiness score together with out-of-distribution exact match for Seed 1, the clearest transition case. Readiness rises earlier and more smoothly than generalization. Seed-specific versions for all three seeds are provided in Appendix H.

Table 1: Main COGS results across three seeds. In-distribution performance saturates quickly, but out-of-distribution generalization remains much lower. Lexical generalization rises substantially, whereas structural generalization remains near floor.

Seed	Max Dev EM	Max GEN EM	Max Lexical	Max Structural	Transition@30%	Transition@50%
1	0.998	0.561	0.651	0.027	30k	38k
2	0.998	0.533	0.617	0.031	26k	–
3	0.998	0.480	0.557	0.024	32k	–

Table 2: Predictive readiness results on COGS. All three seeds show positive predictive correlations between readiness and future held-out improvement, with the strongest effect in Seed 1.

Seed	Max GEN EM	Pred. Corr.	Lead Time @50%	Transition @50%
1	0.561	+0.58	+7k	38k
2	0.533	+0.29	–	–
3	0.480	+0.09	–	–

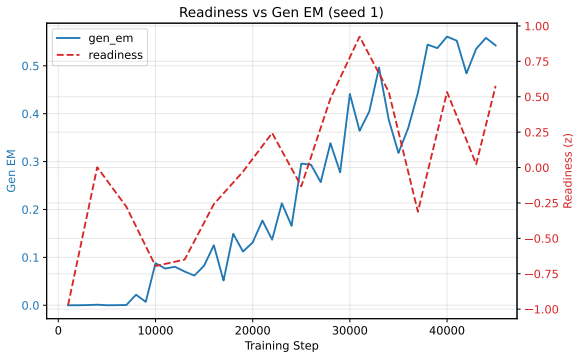

Figure 3: Composite readiness score and out-of-distribution exact match over training on COGS (Seed 1). Readiness rises before generalization becomes behaviorally visible.

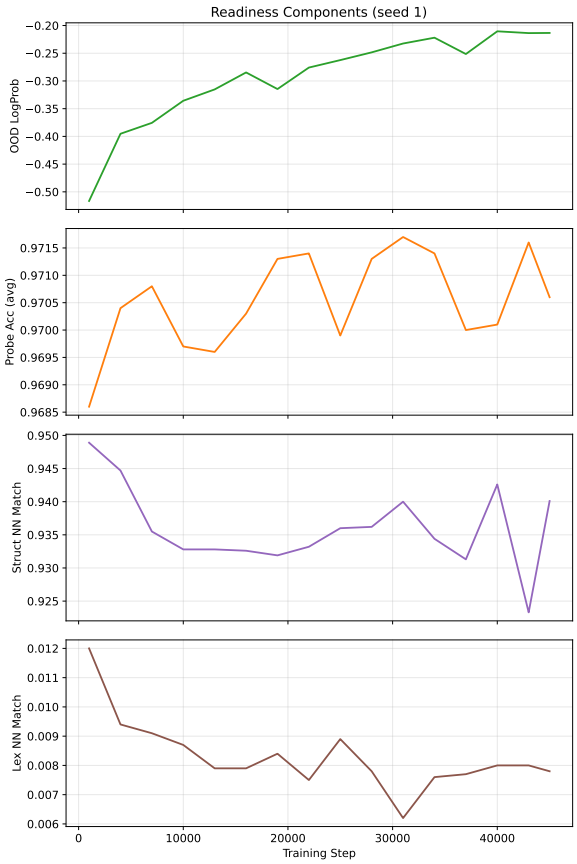

Figure 4: The components of latent readiness on COGS (Seed 1): out-of-distribution token confidence, structural probe accuracy, and structural neighbor alignment.

Figure 4 decomposes readiness into its three component indicators for Seed 1. In our current runs, OOD token confidence is the strongest predictive component, whereas the probe and neighbor indicators provide weaker supportive evidence. This means readiness is not purely a hidden-state construct, but neither is it reducible to exact-match performance alone: the appendix shows that checkpoints with similar current performance can still differ in future gains as a function of readiness, and that structural neighbor alignment diverges from lexical neighbor alignment (Figure 14).

Taken together, these results suggest that on COGS, differences in internal preparation become measurable before held-out gains are fully visible at the sequence level.

4.3 Turkish SIGMORPHON reinflection is learned gradually and does not exhibit the same readiness-before-generalization pattern

We now turn to the main contrastive morphology setting. Table 3 reports Turkish SIGMORPHON reinflection results on the hard subset defined by rare MSD patterns. The picture changes sharply. Hard-subset accuracy reaches 0.642–0.703 across seeds, and the transition-to-50% point occurs very early, around 1.2k–1.5k steps. More importantly, predictive correlations are strongly negative, ranging from -0.93 to -0.96.

Figure 5 shows the Turkish medium-resource

Table 3: Turkish SIGMORPHON reinflection results on the hard subset defined by rare MSD patterns. Negative lead times are omitted because readiness lags rather than precedes performance in this setting. Unlike COGS, the task is learned gradually and does not show a readiness-before-transition pattern.

Seed	Max Rare MSD Acc.	Transition @50%	Lead Time @50%	Pred. Corr.
1	0.642	1.5k	–	-0.95
2	0.662	1.2k	–	-0.96
3	0.703	1.3k	–	-0.93

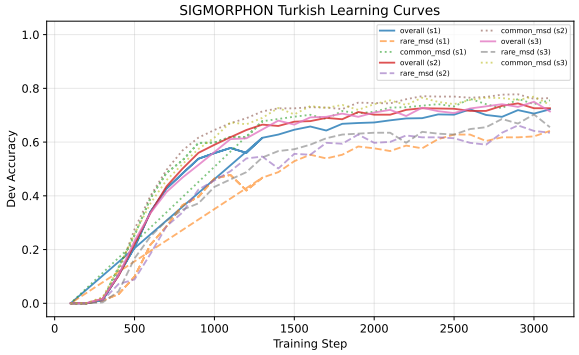

Figure 5: Turkish medium-resource SIGMORPHON reinflection learning curves. The hard subset is learned gradually rather than through a delayed transition.

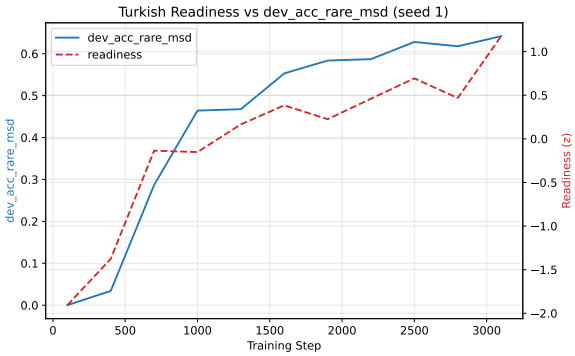

Figure 6: Readiness versus hard-subset accuracy for Turkish medium-resource SIGMORPHON reinflection (Seed 1). Unlike COGS, readiness does not lead later improvement.

learning curve, and Figure 6 shows the corresponding readiness-versus-accuracy trajectory for Seed 1. Instead of a long pre-transition interval, the task improves smoothly and crosses the 50% threshold almost immediately. The readiness signal does not lead later improvement; it lags or co-moves with performance.

This contrast rules out a simple progress-based interpretation of latent readiness. If readiness simply tracked generic training progress, it should predict improvement across both benchmark families. Instead, in Turkish reinflection, the indicators lag or co-move with performance rather than preceding it.

4.4 Latent readiness is most informative in the transition-like regime studied here

The cross-dataset comparison is summarized in Table 4. COGS exhibits three properties at once: a long delay between in-distribution and held-out performance, strong asymmetry between easier and harder forms of generalization, and positive predictive readiness. The SIGMORPHON conditions studied here do not show the same pattern: learning is smoother, the transition threshold is reached early on the Turkish hard subset, and readiness is not predictive.

Figure 7 in Appendix G extends this comparison with low-resource English and Turkish morphology settings. Across these additional conditions, the broader qualitative pattern is unchanged: the morphology settings considered here behave more gradually than COGS, and the strongest readiness-before-transition effect remains concentrated in the transition-like compositional setting.

This is the paper's main empirical result. The evidence does not support latent readiness as a uniformly expressed feature of structured learning. Instead, it supports a narrower claim: in the settings studied here, latent readiness is most informative when learning passes through a delayed, transition-like bottleneck.

5 Discussion

Our framework is motivated by a simple possibility: a model may fail to generalize not because it lacks relevant structure entirely, but because that structure has not yet become coordinated strongly enough to support correct sequence-level outputs. The present results support this interpretation in COGS, but they also refine it in an important way.

In COGS, we find evidence for a measurable **pre-transition state** in which internal preparation improves before **visible generalization** does. This suggests that generalization onset need not coincide with the first appearance of relevant organization. At least in this setting, the model appears to accumulate internal preparation before it can express that preparation behaviorally. This reframing is especially useful because COGS shows strong lexical–structural asymmetry: some held-out be-

Table 4: Cross-dataset comparison. Readiness is predictive on COGS, which exhibits delayed and transition-like generalization, but not in the SIGMORPHON English and Turkish conditions studied here, which are learned more gradually.

Dataset	Easy/Hard Asymmetry	Transition-Like Onset	Predictive Readiness	Dominant Signal
COGS	Strong (lexical $\gg$ structural)	Yes	Yes	OOD token confidence
SIGMORPHON (English/Turkish)	Much weaker	No / gradual	No	–

havior improves while structural generalization remains blocked, so a single out-of-distribution score hides important internal differences.

At the same time, latent readiness is evidence, not a complete mechanism. Token-level confidence reflects whether the model is moving probabilistically toward correct outputs. Structural probes test whether hidden states increasingly encode task-relevant distinctions. Structural neighbor alignment measures whether representation-space geometry is reorganizing around structure rather than lexical overlap. Together, these indicators provide a richer picture than any one of them alone, but they do not establish a complete causal account of generalization. A probe may recover structure without proving that the model uses it for prediction, and neighbor alignment may reveal geometric regularity without identifying the mechanism that produces it.

This caution matters especially because our strongest predictive component is behavior-adjacent. OOD token confidence is informative precisely because it provides a dense precursor to exact-match success, but for that reason it should not be read as pure evidence of hidden-state organization by itself. Our interpretation instead rests on the joint pattern: behavior-adjacent improvement together with weaker but aligned representational signals.

The SIGMORPHON English and Turkish conditions studied here do not recover the same readiness-before-generalization pattern. This helps separate latent readiness from generic training progress. If readiness were simply another name for optimization progress, we would expect it to behave similarly across settings. Yet that is not what we observe. In the settings studied here, latent readiness is most informative where learning shows delayed, transition-like onset and least informative where learning is more gradual. One plausible explanation is that COGS contains a longer interval in which partial structural preparation can accumulate before exact-match behavior becomes stable, whereas in morphology similar improvements may be expressed more directly and continuously in output behavior. We offer this as an interpretation rather than a demonstrated mechanism.

More broadly, these results support a temporal view of language learning dynamics. Internal organization and external behavior need not change at the same pace. In some structured tasks, the model may first become more prepared and only later become consistently correct. In others, learning may be gradual enough that no such pre-transition regime becomes distinguishable. The key question is therefore not simply whether readiness exists, but when it becomes predictively useful.

6 Conclusion

We proposed **latent readiness** to characterize what happens before generalization becomes behaviorally visible. Rather than treating out-of-distribution success as the first point at which relevant structure appears, our framework asks whether internal preparation accumulates earlier and can be measured before behavior changes.

The results support three conclusions. First, COGS exhibits delayed and asymmetric visible generalization. Second, within that setting, we find evidence for a measurable pre-transition state in which internal preparation improves before held-out behavior does. Third, we do not recover the same pattern in the SIGMORPHON English and Turkish settings studied here. Taken together, these findings support a regime-sensitive interpretation of latent readiness rather than a universal one.

Limitations

This work studies latent readiness in deliberately controlled settings. Our experiments use small encoder–decoder Transformers trained from scratch, which allow dense trajectory analysis but limit the scope of our conclusions. Even if we observe pre-transition readiness in COGS, this does not imply that the same dynamics hold in larger

pretrained language models or in more open-ended language tasks.

A second limitation is that our readiness measures are operational rather than universal. Token-level confidence, structural probes, and neighbor alignment are useful indicators in our setup, but they may not exhaust the forms of internal preparation relevant to other tasks or architectures. Our composite readiness score should therefore be interpreted as a practical proxy for a latent construct in these experiments, not as a task-independent definition of readiness.

A third limitation is that the present analysis is trajectory-based and does not establish causality by itself. Even if readiness indicators precede generalization onset in COGS, this does not prove that they are the mechanism responsible for later behavioral success. Stronger causal claims would require intervention-based studies that directly manipulate training conditions or representational structure.

A fourth limitation is that our cross-dataset contrast remains limited. While the SIGMORPHON results show that the same readiness-before-generalization pattern is not recovered in the settings considered here, English and Turkish do not span the full space of morphological learning regimes. More broadly, our SIGMORPHON analysis is not fully symmetric across languages: English serves mainly as a supplementary comparison based on edit-distance difficulty, whereas Turkish receives a more detailed hard-subset analysis based on rare MSD patterns. A fuller account of latent readiness will require matched subset constructions and parallel analyses across additional datasets, languages, architectures, and learning regimes.

Ethical Considerations

This work studies learning dynamics on publicly available benchmark datasets using small encoder–decoder Transformers trained from scratch. We do not collect new human-subject data, deploy user-facing systems, or study settings involving personal or sensitive information. The main ethical consideration is therefore the interpretation and downstream use of our findings: latent readiness is an analytical construct for understanding model behavior, not a guarantee of robust generalization or safe deployment. We encourage future work to examine whether similar readiness signals remain reliable in larger pretrained systems and in higher-stakes applications.

References

Guillaume Alain and Yoshua Bengio. 2016. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*.

Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*.

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. 2019. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single \textdollar\!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sandra Kübler, David Yarowsky, Jason Eisner, and Mans Hulden. 2017. CoNLL-SIGMORPHON 2017 shared task: Universal morphological reinflection in 52 languages. In *Proceedings of the CoNLL SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection*.

Aslı Göksel and Celia Kerslake. 2005. *Turkish: A Comprehensive Grammar*. Routledge.

Martin Haspelmath. 2002. *Understanding Morphology*. Arnold.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of NAACL-HLT*.

Dieuwke Hupkes, Verna Dankers, Mathijs Mul, Elia Bruni, and 1 others. 2023. A taxonomy and review of generalization research in NLP. *Nature Machine Intelligence*.

Daniel Keysers, Nathanael Schärli, Nathan Scales, Hylke Buisman, Daniel Furrer, Sergii Kashubin, Nikola Momchev, Danila Sinopalnikov, Lukasz Stafiniak, Tibor Tihon, Dmitriy Tsarkov, Xiao Wang, Marc van Zee, and Olivier Bousquet. 2020. Measuring compositional generalization: A comprehensive method on realistic data. In *International Conference on Learning Representations*.

Najoung Kim and Tal Linzen. 2020. COGS: A compositional generalization challenge based on semantic interpretation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. 2019. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*.

Brenden M. Lake and Marco Baroni. 2018. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*.

Vladimir I. Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8):707–710.

Bingzhi Li, Lucia Donatelli, Alexander Koller, Tal Linzen, Yuekun Yao, and Najoung Kim. 2023. SLOG: A structural generalization benchmark for semantic parsing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

Ziming Liu, Ouail Kitouni, Niklas Nolte, Eric J. Michaud, Max Tegmark, and Mike Williams. 2022. Towards understanding grokking: An effective theory of representation learning. In *Advances in Neural Information Processing Systems*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

João Loula, Marco Baroni, and Brenden Lake. 2018. Rearranging the familiar: Testing compositional generalization in recurrent networks. In *Proceedings of the Workshop on Generalization in the Age of Deep Learning*.

Ari S. Morcos, Maithra Raghu, and Samy Bengio. 2018. Insights on representational similarity in neural networks with canonical correlation. In *Advances in Neural Information Processing Systems*.

Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. 2021. Deep double descent: Where bigger models and more data hurt. *Journal of Statistical Mechanics: Theory and Experiment*.

Inbar Oren, Jonathan Herzig, Nitish Gupta, Matt Gardner, and Jonathan Berant. 2020. Improving compositional generalization in semantic parsing. In *Findings of the Association for Computational Linguistics: EMNLP 2020*.

Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. 2022. Grokking: Generalization beyond overfitting on small algorithmic datasets. *arXiv preprint arXiv:2201.02177*.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. 2017. SVCCA: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Advances in Neural Information Processing Systems*.

Peter Shaw, Ming-Wei Chang, Panupong Pasupat, and Kristina Toutanova. 2021. Compositional generalization and natural language variation: Can a semantic parsing approach handle both? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R. Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R. Bowman, Dipanjan Das, and Ellie Pavlick. 2019. What do you learn from context? probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*.

Zhengxuan Wu, Christopher D. Manning, and Christopher Potts. 2023. ReCOGS: How incidental details of a logical form overshadow an evaluation of semantic interpretation. *Transactions of the Association for Computational Linguistics*, 11.

A Dataset Details

COGS. We use the standard splits of COGS (Kim and Linzen, 2020): 24,155 training examples, 3,000 development (dev) examples, 3,000 test examples, and 21,000 generalization (gen) examples. The task is semantic parsing: each input is an English sentence and the output is its logical-form representation. The development and test splits share the same distribution as training and serve to track in-distribution (ID) performance. The generalization split contains 21 systematically constructed cases that require compositional generalization to novel structures or lexical items not seen during training.

We partition the 21 generalization cases into two analytical groupings:

- *Lexical* (18 cases): the primary novelty lies in held-out lexical items appearing in otherwise familiar compositional patterns (e.g., a novel verb used as a subject or object of a known construction).
- *Structural* (3 cases): successful prediction requires generalization to held-out compositional configurations—specifically, CP recursion (doubly embedded "that" clauses), PP recursion (nested prepositional phrases), and object-PP-to-subject-PP transfer.

This distinction is analytical rather than taxonomic: it provides a coarse separation between generalization dominated by lexical novelty and generalization dominated by structural novelty. All general-

ization results are reported as exact-match (EM) accuracy over the corresponding subsets.

CoNLL–SIGMORPHON 2017 Task 1. We use English and Turkish from the CoNLL–SIGMORPHON 2017 Shared Task on universal morphological reinflection (Cotterell et al., 2017). The task is character-level transduction: given a lemma and a target morphosyntactic description (MSD), the model must generate the correctly inflected surface form. We evaluate both medium-resource (1,000 training examples) and low-resource (100 training examples) conditions. Development sets contain 1,000 examples for both languages.

Language choice. Turkish and English are chosen as two contrastive settings within the same benchmark family rather than as an attempt at exhaustive typological coverage. Turkish is an agglutinative language with rich suffixal morphology, where sequences of suffixes can mark case, number, possession, tense, and related grammatical categories, often yielding surface forms that differ substantially from the lemma. English, by comparison, has a less extensive inflectional system and places much weaker demands on suffix stacking and morphophonological alternation (Göksel and Kerslake, 2005; Haspelmath, 2002). This design allows us to test whether the readiness pattern observed in COGS persists in a distinct family of structured language-learning tasks while also probing the effect of morphological complexity.

A.1 Task-Specific Evaluation Groupings

We define several task-specific evaluation groupings to distinguish different kinds of generalization difficulty across datasets.

COGS generalization subsets. For COGS, we evaluate on the full generalization split and additionally on lexical and structural subsets as defined above. Because structural generalization cases (CP recursion, PP recursion, object-PP-to-subject-PP) require the model to produce compositional structures never seen during training, they represent a qualitatively harder generalization challenge. In all three seeds, structural accuracy remains near zero ($\leq$0.031 EM) while lexical accuracy rises to 0.56–0.65 EM, confirming that the bottleneck in COGS generalization lies in structural composition rather than lexical coverage.

SIGMORPHON difficulty structure. For SIGMORPHON reinflection, we define two complementary difficulty axes:

1. **Edit-distance buckets.** We compute the Levenshtein distance (Levenshtein, 1966) between the lemma and the gold target form and define three analytical buckets for this study: low (0–2 character edits), medium (3–5), and high ($\geq$6). These are study-specific groupings rather than categories provided by the original benchmark. We choose these cutoffs to distinguish minimal surface-form transformations, moderate transformations, and larger transformations while preserving interpretable differences in the empirical distributions of English and Turkish. Low-edit examples typically involve simple suffixation (e.g., `schmear`→`schmeared`, edit = 2). Medium-edit examples involve longer affixation or minor stem changes (e.g., Turkish `kuzu`→`kuzunuzda`, edit = 5, expressing locative + 2nd person plural possessive). High-edit examples involve multiple stacked affixes or stem-internal changes (e.g., Turkish `vade`→`vadelerimizden`, edit = 10, expressing ablative + 1st person plural possessive + plural). We use Levenshtein edit distance as an operational proxy for surface-form transformation difficulty, not as a complete theory of morphological complexity.
2. **Rare-MSD subset.** We count how often each MSD pattern appears in training data and label patterns appearing $\leq$3 times as "rare." This captures reinflection patterns that the model has seen few or no training examples of.

The distribution of these difficulty dimensions differs sharply between languages:

Bucket / Subset	English	Turkish
Low edit (0–2)	747 (75%)	101 (10%)
Medium edit (3–5)	253 (25%)	334 (33%)
High edit (6+)	0 (0%)	565 (57%)
Rare MSD	0 (0%)	293 (29%)
Common MSD	1,000 (100%)	707 (71%)

Table 5: Difficulty structure in the CoNLL–SIGMORPHON 2017 English and Turkish dev sets. English morphology is concentrated in low- and medium-edit transformations with no rare-MSD examples in our conditions. Turkish includes many high-edit cases and a substantial rare-MSD subset.

Because English has no high-edit or rare-MSD

examples, we analyze English primarily through edit-distance groupings (low vs. medium). For Turkish, we use both edit-distance buckets and the rare-MSD subset; the latter serves as the primary hard-subset metric in the main text because it captures a qualitatively distinct axis of difficulty—unfamiliar morphological patterns—rather than merely longer transformations.

B Model Architecture and Training Details

B.1 Architecture

All experiments use an encoder–decoder Transformer (Vaswani et al., 2017) with shared vocabulary embedding and sinusoidal positional encoding. We use the PyTorch `nn.Transformer` with `batch_first=True`. Decoding is performed via greedy search (no beam search) at evaluation time. The same architecture template is used for both COGS and SIGMORPHON; only the model size and vocabulary differ.

Hyperparameter	COGS	SIGMORPHON
Encoder layers	3	2
Decoder layers	3	2
Hidden size (d_{model})	256	128
Attention heads	4	4
FFN size (d_{ff})	512	256
Dropout	0.1	0.1
Tokenization	word-level	character-level
Max source length	64	128
Max target length	128	64

Table 6: Architecture configurations. COGS uses word-level tokenization for English semantic parsing. SIGMORPHON uses character-level tokenization for morphological reinflection.

B.2 Training Configuration

All experiments use AdamW (Loshchilov and Hutter, 2019) with the hyperparameters shown in Table 7.

For the low-resource SIGMORPHON condition (100 training examples), we adjust batch size to 16, increase training to 500 epochs (≈3,000 steps), and evaluate every 50 steps. All other hyperparameters remain the same.

No hyperparameter tuning was performed; the same configuration is used for all seeds. All experiments were conducted on a single NVIDIA RTX 3060 Ti (8 GB VRAM).

Hyperparameter	COGS	SIGMORPHON (medium)
Optimizer	AdamW	AdamW
Learning rate	3×10^{-4}	3×10^{-4}
(β_1, β_2)	(0.9, 0.98)	(0.9, 0.98)
Weight decay	0.01	0.01
Gradient clipping	1.0	1.0
Batch size	32	32
LR schedule	linear warmup + constant	linear warmup + constant
Warmup steps	1,000	200
Training duration	60 epochs (≈45k steps)	100 epochs (≈3.1k steps)
Mixed precision	FP16 (AMP)	FP16 (AMP)
Eval / save interval	every 1,000 steps	every 100 steps
Seeds	1, 2, 3	1, 2, 3

Table 7: Training hyperparameters. All experiments use the same optimizer and schedule family; only duration and warmup differ to account for dataset size.

C Readiness Feature Construction

We compute readiness features at a sampled subset of training checkpoints: every 3rd saved checkpoint plus the final checkpoint (yielding ≈16 checkpoints per seed for COGS out of 45 total, and ≈20 for SIGMORPHON out of ≈60 total). At each checkpoint, we extract three signals.

C.1 R1: OOD Token Confidence

We compute the mean teacher-forced log-probability of gold target tokens on held-out examples. Specifically, for each held-out example (x, y), we feed the gold source x and the gold target prefix $y_{<t}$ into the model at each position t and record $\log p(y_t \mid y_{<t}, x)$. We then average across all non-padding target positions within each example to obtain a per-example score, and average across all held-out examples to obtain the checkpoint-level R1 signal.

For COGS, the held-out set is the full generalization split (21,000 examples). For SIGMORPHON, it is the development set (1,000 examples).

This signal measures how confidently the model assigns probability to the correct output tokens on out-of-distribution inputs. Unlike accuracy, which is binary and typically near zero in the pre-transition regime, log-probability provides a continuous graded signal that can increase before accuracy does.

C.2 R2: Structural Probe Accuracy

We train lightweight logistic regression probes on checkpoint encoder representations to recover coarse structural attributes, following the general probing methodology used to test the recoverability of linguistic or structural information from learned representations (Alain and Bengio, 2016; Hewitt and Manning, 2019; Belinkov, 2022). The procedure is:

1. Extract the encoder output at each position and compute the mean-pooled sentence representation (averaging over non-padding positions).
2. For COGS, we define three binary/multi-class probe tasks:
   - **CP recursion**: whether the source contains $\geq$2 occurrences of "that" (indicating embedded clauses).
   - **PP recursion**: whether the source contains $\geq$2 prepositional phrases.
   - **Length bucket**: coarse length category ($\leq$10 tokens, 11–20, 21+).

   For SIGMORPHON, we define analogous probes:
   - **MSD frequency**: whether the example's MSD is rare ($\leq$3 occurrences in training) or common.
   - **Edit bucket**: the edit-distance category (low / medium / high).
   - **Transform type**: the kind of morphological transformation (identical, suffix-only, prefix-only, stem-change, mixed).
3. Fit a logistic regression classifier (L-BFGS solver, max 1,000 iterations) using `StandardScaler`-normalized representations. Training examples serve as probe training data; held-out examples serve as probe evaluation data.
4. R2 is the average accuracy across all probe tasks at each checkpoint.

This signal measures how well the encoder's learned representations linearly encode task-relevant structural distinctions, independent of whether the decoder can use them for generation.

C.3 R3: Structural Neighbor Alignment

We measure whether held-out examples are nearest in representation space to structurally analogous training examples rather than to merely lexically similar ones.

1. Compute cosine similarity between each held-out example's mean-pooled encoder representation and all training representations.
2. For each held-out example, identify its $k = 10$ nearest training neighbors.
3. For COGS: compute the fraction of top-k neighbors that share the same structural class (matching on both CP-recursion and PP-recursion attributes simultaneously).
4. For SIGMORPHON: compute the fraction of top-k neighbors that share the same `transform_type` (e.g., suffix-only, stem-change).
5. Average across all held-out examples.

This signal captures whether the encoder organizes representations by structural similarity rather than superficial form similarity, in the broader spirit of representation-space analyses that compare the geometry of learned neural representations over training or across models (Raghu et al., 2017; Morcos et al., 2018; Kornblith et al., 2019).

Lexical neighbor alignment (COGS only). As a control, we also compute a lexical neighbor alignment score: the fraction of top-k neighbors that share $\geq$50% of content words (measured by Jaccard similarity on the set of non-stopword tokens, with threshold 0.5). This control distinguishes structural reorganization from mere lexical clustering: if representations are organized by lexical overlap rather than structural composition, lexical neighbor alignment would be high while structural neighbor alignment would be low.

C.4 Composite Readiness Score

The composite readiness score at each checkpoint is formed by:

1. Standardizing each component across all sampled checkpoints within a seed:

$$z_i = \frac{x_i - \bar{x}}{\sigma_x}$$

   where x_i is the raw value at checkpoint i, $\bar{x}$ is the mean, and σ_x is the standard deviation across all sampled checkpoints for that seed.
2. Averaging the three z-scores:

$$\text{readiness}_t = \frac{z(\text{R1}_t) + z(\text{R2}_t) + z(\text{R3}_t)}{3}$$

This construction ensures that each component contributes equally regardless of its raw scale. For ablation analyses, we also compute partial composites: R1-only, R2-only, R3-only, and R1+R2.

D Predictive Analysis Methodology

D.1 Predictive Correlation

We define predictive correlation as the Pearson correlation between the readiness score at checkpoint t and the future change in held-out performance:

$$r = \mathrm{corr}\big(\mathrm{readiness}_t,\ \ \mathrm{perf}_{t+k} - \mathrm{perf}_t\big)$$

where k is the number of evaluation intervals to look ahead. We use $k = 3$ throughout (i.e., 3,000 training steps ahead for COGS, 300 steps for SIGMORPHON).

A positive predictive correlation means that higher readiness at time t is associated with larger subsequent performance improvement—readiness *leads* performance. A negative correlation means that readiness increases only after performance has already risen—readiness *lags* performance.

D.2 Lead Time

Lead time measures the temporal gap between when readiness first crosses a high percentile and when generalization performance first crosses a sustained threshold:

1. Compute the 75th percentile of readiness scores across all sampled checkpoints for a given seed.
2. Identify the first checkpoint at which readiness exceeds this percentile.
3. Identify the first checkpoint at which generalization EM exceeds a fixed threshold (e.g., 30% or 50%) for $\geq$3 consecutive evaluation intervals (the "sustained-transition" criterion).
4. Lead time = transition step $-$ readiness crossing step.

Positive lead time indicates that readiness rises before the generalization transition. We report lead time at both 30% and 50% generalization thresholds where available.

D.3 Matched-Performance Analysis

To control for the possibility that predictive correlation is driven by both readiness and future gain increasing together over training time, we construct matched-performance checkpoint pairs:

1. Identify all pairs of checkpoints (i, j)—potentially from different seeds—such that $|\mathrm{gen_em}_i - \mathrm{gen_em}_j| \leq 0.02$.
2. For each pair, compute $\Delta\mathrm{readiness} = \mathrm{readiness}_i - \mathrm{readiness}_j$ and $\Delta\mathrm{future_gain} = \mathrm{future_gain}_i - \mathrm{future_gain}_j$.
3. If readiness is genuinely predictive beyond current performance level, Δreadiness and Δfuture_gain should be positively correlated.

This analysis asks: *among checkpoints with similar current performance, do internal readiness differences still predict different future trajectories?*

D.4 Low-GEN Regime Analysis

We further focus on the pre-transition regime by restricting to checkpoints with low generalization accuracy (gen_em ≤ 0.05 or ≤ 0.10). Within this restricted set, we perform a median split on readiness and compare mean future gains between the two groups. This analysis asks: *within the period where the model has not yet generalized, are there internal signatures that distinguish checkpoints that are "about to" generalize from those that are not?*

E Transition Definitions and Sensitivity

We define the *transition step* at threshold θ as the first training step at which generalization EM $\geq \theta$ for at least 3 consecutive evaluation intervals. The consecutive-interval requirement prevents transient fluctuations from being counted as sustained transitions.

Threshold	Seed 1	Seed 2	Seed 3
0.30	30,000	26,000	32,000
0.50	38,000	–	–
0.70	–	–	–

Table 8: COGS transition steps by threshold and seed. All seeds reach sustained 30% generalization. Only Seed 1 reaches sustained 50% generalization within 45k training steps. No seed reaches sustained 70% generalization, consistent with the known difficulty ceiling of COGS for small Transformers.

For Turkish SIGMORPHON reinflection (rare-MSD accuracy as the hard-subset metric), the 50% threshold is reached much earlier—between 1,200 and 1,500 steps—and the learning curve rises gradually and monotonically rather than exhibiting a delayed onset. For English SIGMORPHON (using overall dev accuracy), the 50% threshold is reached between 400 and 600 steps.

This contrast in learning dynamics is central to the paper's thesis: readiness is predictive in transition-like regimes (COGS) but not in gradual-learning regimes (SIGMORPHON).

F Full Quantitative Results

F.1 COGS Per-Seed Results

Seed	Max Dev EM	Max GEN EM	Max Lexical	Max Structural	Trans.@30	Trans.@50	Pred. Corr. (R1+R2+R3)
1	0.998	0.561	0.651	0.027	30k	38k	+0.578
2	0.998	0.533	0.617	0.031	26k	–	+0.288
3	0.998	0.480	0.557	0.024	32k	–	+0.086
Mean	0.998	0.525	0.608	0.027	–	–	+0.317

Table 9: Full COGS summary across three seeds. All seeds show strong ID performance (dev EM $\approx$1.0), delayed generalization onset, and a pronounced lexical–structural asymmetry. Predictive correlation is positive in all three seeds.

F.2 Component Ablation

Component	Pred. Corr. (mean $\pm$ std)	Seed 1	Seed 2	Seed 3
R1 only (OOD confidence)	$+0.265 \pm 0.077$	+0.365	+0.250	+0.179
R2 only (probe accuracy)	$+0.150 \pm 0.415$	+0.707	+0.032	-0.288
R3 only (neighbor alignment)	$+0.113 \pm 0.245$	-0.233	+0.282	+0.291
R1 + R2	$+0.246 \pm 0.264$	+0.595	+0.189	-0.044
R1 + R2 + R3 (composite)	$+0.317 \pm 0.202$	+0.578	+0.288	+0.086

Table 10: COGS readiness component ablation. R1 (OOD confidence) is the most consistently positive component across seeds. The full composite (R1+R2+R3) achieves the highest mean predictive correlation and the lowest cross-seed variance among multi-component combinations. Individual components show higher variance across seeds, suggesting that they capture partially complementary information.

F.3 Low-GEN Regime Analysis

Group	Thresh	n	GEN EM	Readiness	FG_1	FG_3	FG_5	OOD	Probe
Low-R, gen$\leq$0.05	0.05	6	0.015	-1.196	-0.007	0.007	0.022	-0.431	0.968
High-R, gen$\leq$0.05	0.05	6	0.011	-0.211	0.003	0.026	0.038	-0.373	0.971
Low-R, gen$\leq$0.10	0.10	9	0.028	-1.018	-0.006	0.014	0.034	-0.403	0.969
High-R, gen$\leq$0.10	0.10	9	0.037	+0.078	0.017	0.046	0.078	-0.339	0.971

Table 11: Low-GEN regime comparison. FG_k = mean future gain at k evaluation intervals ahead. At the gen$\leq$0.10 threshold, high-readiness checkpoints show 3.3$\times$ larger future gain at $k = 3$ (0.046 vs. 0.014) despite comparable current generalization accuracy, indicating that readiness captures internal structure beyond current performance.

F.4 Turkish SIGMORPHON Per-Seed Results

Seed	Max Dev Acc	Max Rare MSD Acc	Trans.@50	Lead Time@50	Pred. Corr.
1	0.722	0.642	1,500	–	-0.953
2	0.710	0.662	1,200	–	-0.959
3	0.750	0.703	1,300	–	-0.927
Mean	0.727	0.669	–	–	-0.946

Table 12: Full Turkish SIGMORPHON summary. Negative lead times are omitted because readiness lags rather than leads performance in this setting. Predictive correlation is strongly negative in all seeds, indicating that readiness signals lag rather than lead performance improvement.

F.5 English SIGMORPHON Per-Seed Results

Seed	Max Dev Acc	Trans.@50	Lead Time@50	Pred. Corr.
1	0.839	500	–	-0.989
2	0.836	500	–	-0.993
3	0.850	600	–	-0.997
Mean	0.842	–	–	-0.993

Table 13: Full English SIGMORPHON summary. English shows even stronger negative predictive correlation than Turkish, consistent with its simpler, more monotonically gradual learning trajectory. English is included to confirm that the gradual-learning pattern is a task-family effect rather than a language-specific artifact.

F.6 Cross-Dataset Comparison

Dataset	Task	Easy–Hard Asymmetry	Transition-Like?	Pred. Corr.	Predictive?
COGS	Compositional gen.	yes (lex $\gg$ struct)	yes	$+0.317 \pm 0.202$	yes
SIGMORPHON Turkish	Morph. reinflection	yes (common $>$ rare)	no	-0.946 ± 0.014	no
SIGMORPHON English	Morph. reinflection	minimal	no	-0.993 ± 0.003	no

Table 14: Cross-dataset regime comparison. Readiness is predictive of future generalization gains only on COGS, where learning exhibits a delayed, transition-like onset. On both SIGMORPHON languages, where learning is gradual, readiness lags performance (negative predictive correlation).

G Resource-Level Comparisons

We additionally evaluate low-resource (100 training examples) conditions for both English and Turkish to test whether the gradual-learning pattern persists under stronger data scarcity.

Language	Resource	n_{train}	Max Dev Acc	Pred. Corr.	Predictive?
Turkish	medium	1,000	0.669	-0.946 ± 0.014	no
Turkish	low	100	0.024	-0.051 ± 0.144	no
English	medium	1,000	0.842	-0.993 ± 0.003	no
English	low	100	0.116	-0.563 ± 0.101	no

Table 15: Resource-level comparison. Low-resource models achieve much lower performance overall. Turkish low-resource models barely learn (2.4% dev accuracy), yielding a near-zero predictive correlation. English low-resource models reach 11.6% accuracy with a moderately negative correlation. In neither case does readiness become positively predictive, confirming that the morphological reinflection task family does not exhibit the transition-like learning dynamics that would support predictive readiness.

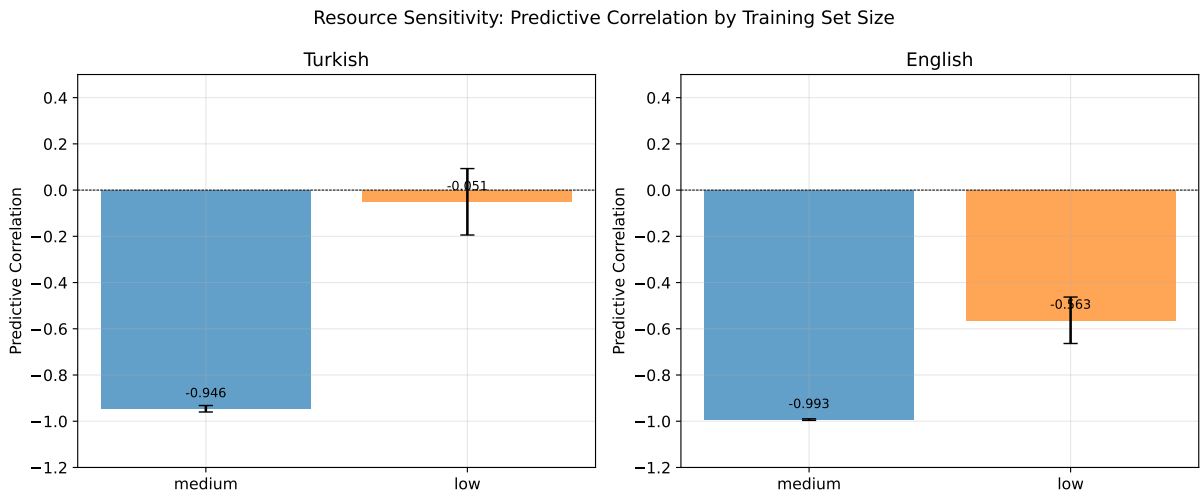

Figure 7: Predictive correlation by resource level and language. Reducing training data weakens the magnitude of the negative correlation but never produces a positive predictive signal.

H Trajectory Plots

H.1 COGS Seed-Level Readiness Trajectories

Figure 8: Composite readiness (dashed, right axis) versus generalization EM (solid, left axis) for COGS across three seeds. In all seeds, readiness begins rising during the early phase of training while generalization EM remains near zero, consistent with internal reorganization preceding the generalization transition. In Seed 1, readiness crosses the 75th percentile approximately 7,000 steps before the 50% generalization transition.

H.2 COGS Readiness Component Decomposition

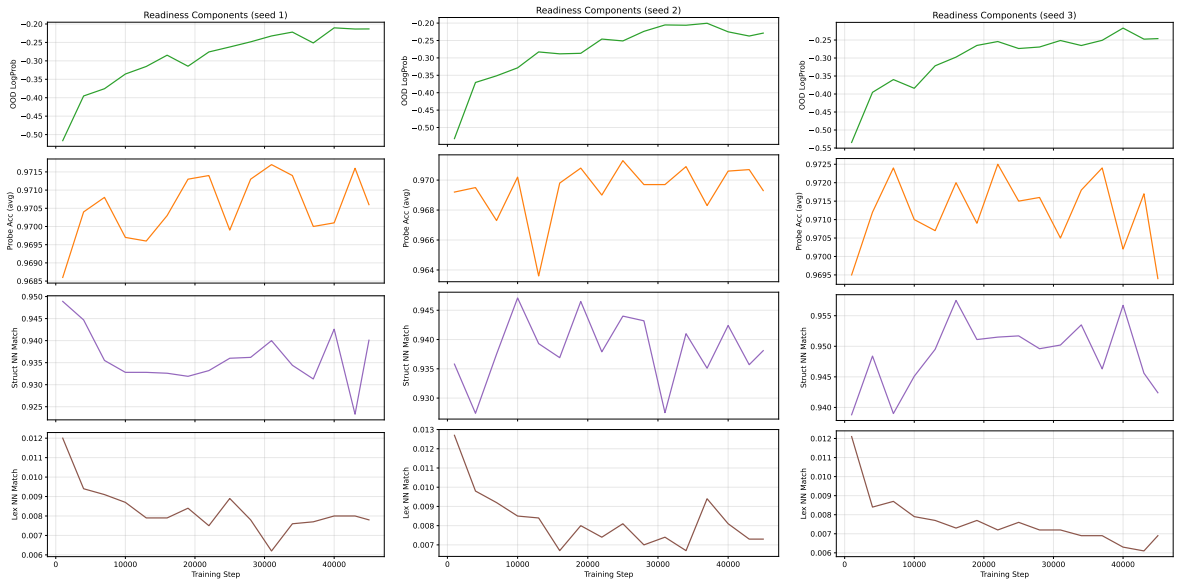

Figure 9: Individual readiness components over training for COGS. Top to bottom: OOD token log-probability (R1), structural probe accuracy (R2), structural neighbor alignment (R3), and lexical neighbor alignment (control). R1 shows the clearest upward trend during the pre-transition period. Lexical neighbor alignment remains near zero throughout, confirming that the encoder organizes representations by structural rather than lexical similarity.

H.3 Turkish SIGMORPHON Readiness Trajectories

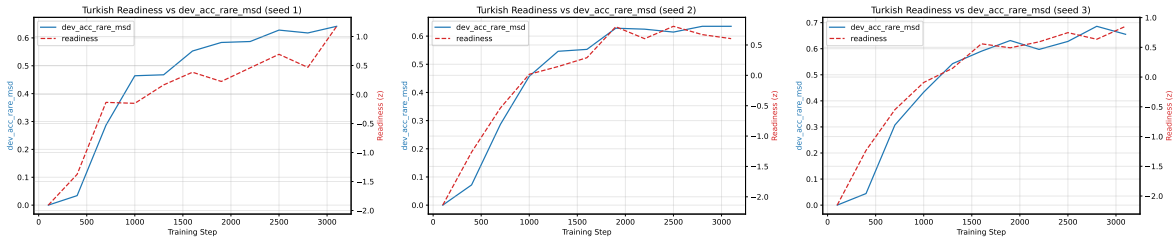

Figure 10: Readiness versus rare-MSD accuracy for Turkish medium-resource SIGMORPHON across three seeds. Unlike COGS, readiness co-moves with or lags performance rather than leading it, consistent with gradual learning.

H.4 English SIGMORPHON Readiness Trajectories

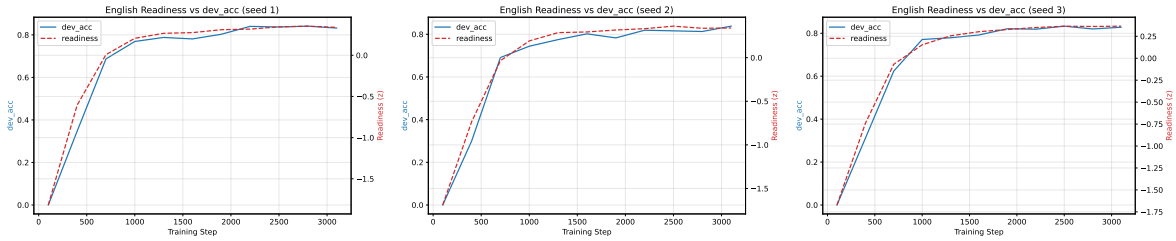

Figure 11: Readiness versus English medium-resource SIGMORPHON development accuracy across three seeds. In addition to overall development accuracy, the corresponding learning-curve plots report the low-edit and medium-edit subsets. The same lagging pattern as Turkish is evident, confirming that the gradual-learning characteristic is a task-family effect.

I Regime Diagnostics

I.1 Matched-Performance Scatter

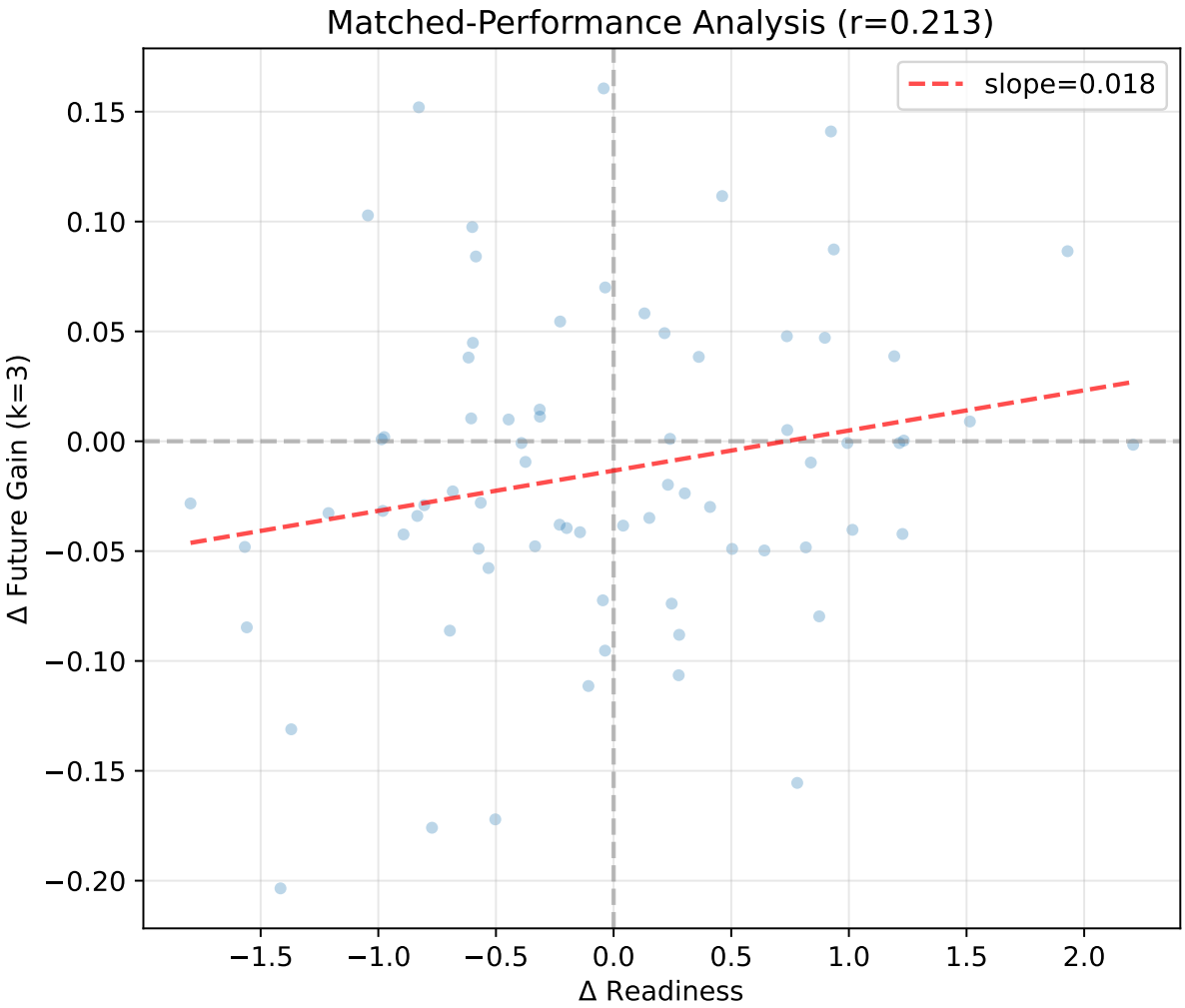

Figure 12: Matched-performance analysis for COGS. Each point represents a pair of checkpoints (potentially from different seeds) with similar current generalization EM ($|\Delta\text{gen_em}| \leq 0.02$). The x-axis shows the difference in readiness between the two checkpoints; the y-axis shows the difference in future performance gain ($k = 3$ intervals ahead). The positive correlation ($r = 0.213$, 75 pairs) indicates that, even at matched current performance, higher readiness is associated with larger future improvement.

I.2 Checkpoint State Space

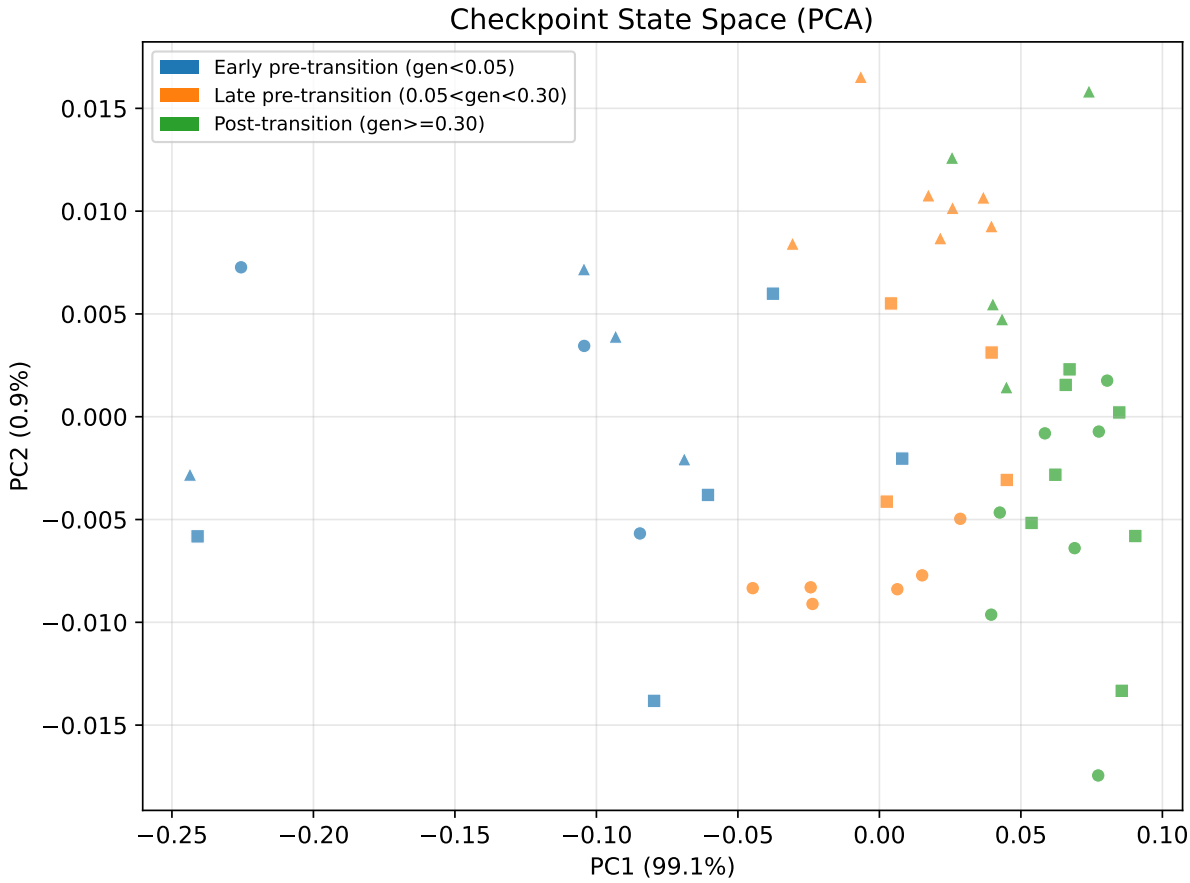

Figure 13: PCA projection of the checkpoint state vector [OOD confidence, probe accuracy, structural neighbor alignment] for all sampled COGS checkpoints across three seeds. Points are colored by generalization phase: early pre-transition (gen$<$0.05, blue), late pre-transition (0.05$\leq$gen$<$0.30, orange), and post-transition (gen$\geq$0.30, green). Marker shapes distinguish seeds ($\circ$ = Seed 1, $\square$ = Seed 2, $\triangle$ = Seed 3). The clear separation between phases indicates that the pre-transition interval is not homogeneous in internal state space; instead, checkpoints trace a structured trajectory from early to late pre-transition states before crossing into the post-transition regime.

I.3 Structural vs. Lexical Neighbor Alignment

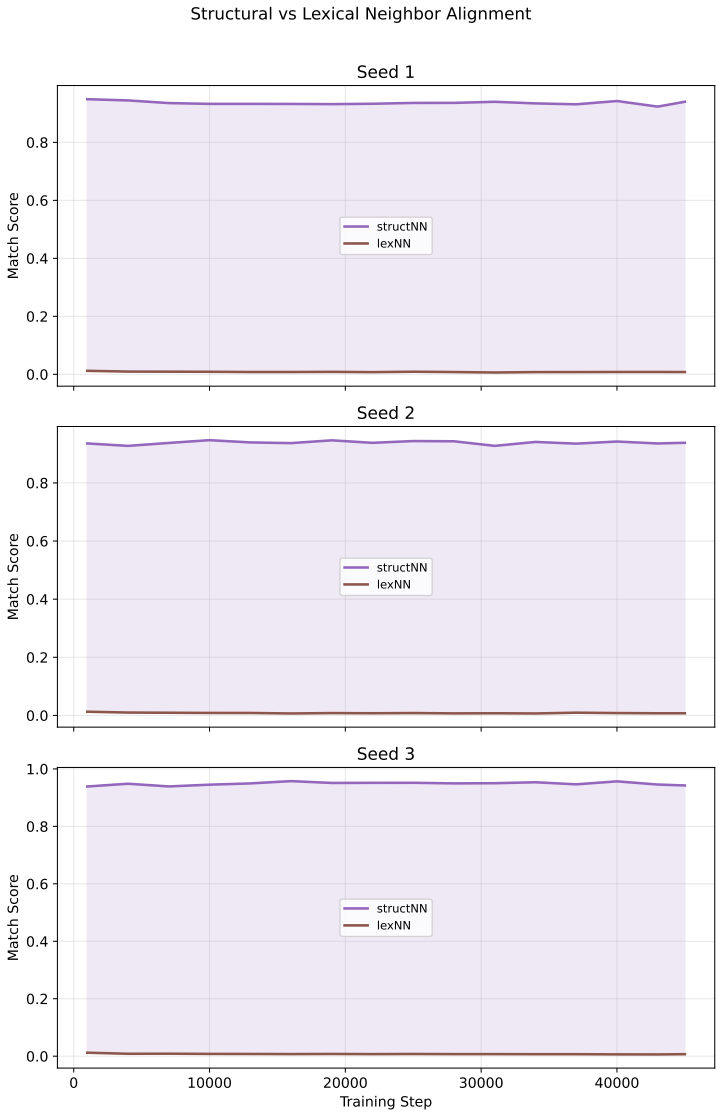

Figure 14: Structural neighbor alignment (structNN, purple) versus lexical neighbor alignment (lexNN, brown) over training for each seed. The shaded region highlights the gap. Structural alignment is consistently high (≈0.93–0.94 throughout training), while lexical alignment remains near zero (≈0.008). This confirms that the encoder organizes generalization-split representations primarily by structural properties (recursion type) rather than by surface lexical overlap—a necessary condition for the structural readiness signals to be meaningful rather than artifacts of lexical memorization.

J English Morphological Reinflection

We additionally trained English medium-resource reinflection models as a supplementary comparison. In this revised version, the English learning-curve plots include overall development accuracy together with the low-edit and medium-edit subsets. High-edit curves are omitted because the English conditions studied here contain no high-edit examples.

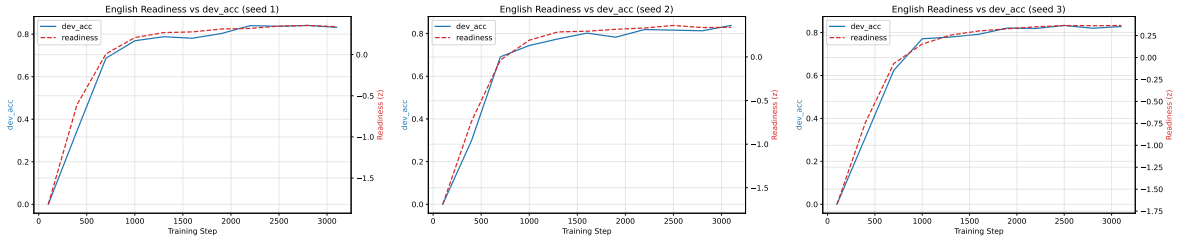

Figure 15: Seed-specific readiness versus accuracy plots for English medium-resource SIGMORPHON reinflection.

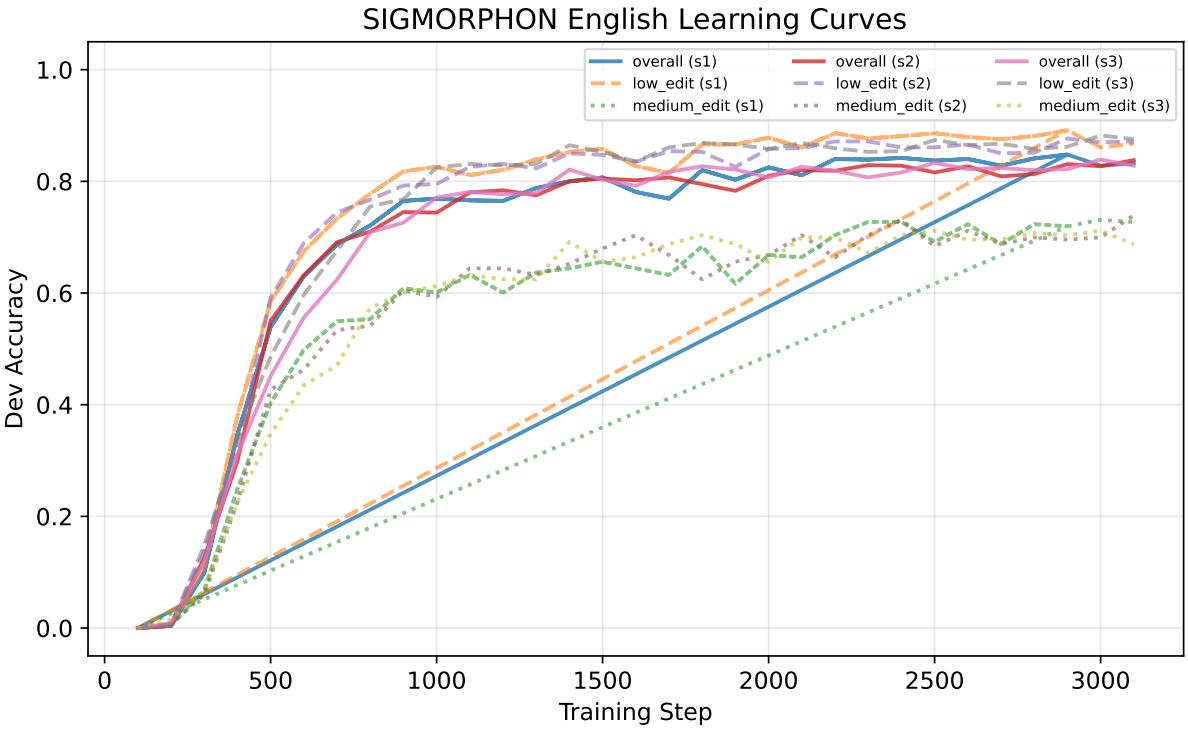

Figure 16: Learning curves for English medium-resource SIGMORPHON reinflection. In addition to overall development accuracy, the plot reports low-edit and medium-edit subsets. English shows gradual improvement across both buckets rather than a delayed transition-like onset.

English is included to test whether the broader gradual-learning picture extends beyond a single morphology language. In the current version, however, English is used as a supplementary comparison based primarily on edit-distance difficulty rather than as a fully matched rare-MSD subset analysis parallel to Turkish.