Data Shapes What, Scale Shapes How: Lexical Learning Dynamics in LLMs

Anonymous ACL submission

Abstract

We investigate how model scale shapes lexical acquisition in LLMs, introducing trajectory comparison as a methodological tool and three new convergence metrics. Tracking surprisal dynamics across four Pythia model sizes and 153 training checkpoints, we report two main findings. First, scale drives a global rise in inter-word surprisal coherence: synonym curves become more correlated at larger scales (dcor: 0.745 at 70M to 0.953 at 1.4B), but baseline correction reveals that the *advantage* of semantic relatedness over random pairs is invariant with scale ($\Delta \approx 0.15$, $R^2 < 0.01$), indicating that what scales is broad inter-word coherence rather than semantic abstraction specifically. Second, learning dynamics are shaped primarily by training data rather than by model capacity: surprisal trajectories are highly similar across model sizes (dcor $\in [0.69, 0.86]$, Table 1), and our new metrics confirm that larger models converge faster but not along qualitatively different paths. Together, these results call for baseline-corrected analyses and longitudinal evaluation when assessing lexical competence in LLMs.

1 Introduction

LLMs demonstrate a remarkable ability to acquire linguistic competence, including word meaning, grammar, and more (Marvin and Linzen, 2018; Linzen and Baroni, 2021). This raises a fundamental and intriguing question: how can a seemingly simple objective (predicting a masked word) lead to the acquisition of the extensive knowledge required to use language correctly? What factors (e.g., corpus size, input diversity, model capacity) influence or enable this learning process?

One productive approach to this question is to study *learning dynamics*: rather than evaluating what a model knows at a fixed endpoint, one tracks how that knowledge builds up across training. Prior work has characterised these dynamics at the level of individual surprisal curves, showing that word-level acquisition is non-monotonic (Chang et al., 2024) and that surprisal does not always decrease over training (Shafiabadi and Wisniewski, 2025).[1]

These studies share a common methodological limitation: they describe how individual trajectories evolve in isolation, without asking how trajectories relate to one another. Yet comparing trajectories is precisely what is needed to address the questions above. Whether two model sizes are driven by the same learning dynamics, or whether semantically related words follow coupled acquisition paths, cannot be read off individual curves: it requires their systematic comparison. Without this, one cannot distinguish a global, scale-driven rise in inter-word coherence from a semantically specific coupling between related words, nor assess whether scale changes *what* is learned or merely *how fast*.

We address this gap by proposing trajectory comparison as a first-class methodological tool for studying lexical acquisition in LLMs. Using the Pythia model suite (Biderman et al., 2023), which provides 153 intermediate checkpoints saved at regular intervals throughout pre-training, we track surprisal trajectories for several hundred words across four model sizes. These checkpoints give access to the full learning process rather than its endpoint, making it possible to compare not just what different models know, but how and when they come to know it. Our contributions are threefold.

Trajectory comparison across model sizes. By directly comparing learning trajectories between Pythia models, we show that surprisal curves are strikingly similar across scales, providing strong evidence that *what* a model learns is shaped primarily by training-data statistics, with model capacity

[1] Surprisal is the negative log-probability of a word given its context; lower values reflect more confident prediction. It serves as a natural proxy for lexical acquisition: as a model learns to represent a word's meaning and distributional behaviour, it assigns higher probability to that word in appropriate contexts, and surprisal decreases accordingly.

governing convergence speed rather than trajectory shape.

Trajectory comparison across semantically related words. By comparing surprisal curves across synonym pairs and inflected forms, and critically by correcting for a global baseline of inter-word coherence, we show that the advantage of synonym pairs over random pairs is invariant with scale. This reveals that what scales is broad inter-word coherence rather than semantic abstraction specifically.

New metrics for trajectory shape. Existing work relies on a threshold-based convergence measure and a variability metric computed over the final 25 % of training (Chang et al., 2024). A richer metric toolkit is needed to disentangle *how* scale affects trajectories: whether it primarily shortens early acquisition, reduces total surprisal burden, or smooths out erratic fluctuations are distinct claims that a single measure cannot separate. We introduce three complementary convergence metrics (linear slope, normalised AUC, and half-life) that capture respectively the overall rate of decline, the total surprisal burden, and the timing of early acquisition, and complement these with the variogram $\gamma(1)$ as a measure of local smoothness throughout the entire training run.

The remainder of the article is organised as follows: Section 2 reviews related work; Section 3 describes the experimental setting; Section 4 presents the results; and Section 5 offers concluding remarks.

2 Related Work

Our work builds directly on Chang et al. (2024), who characterise individual surprisal trajectories across Pythia checkpoints, and on Shafiabadi and Wisniewski (2025), who show that surprisal is not a sufficient proxy for lexical acquisition. Beyond these immediate precursors, our work connects to three broader lines of research: scaling laws, which ask how model capacity affects endpoint performance; probing and representational geometry, which examine what information is encoded in static representations; and the literature on non-monotonic learning and catastrophic forgetting, which analyses instabilities during training.

Scaling laws The relationship between model scale and task performance follows power laws in compute, data, and parameters (Kaplan et al., 2020). Hoffmann et al. (2022) refined these estimates to show that current large models are substantially undertrained relative to a compute-optimal regime, shifting the focus from parameter count to the data-to-parameter ratio. These analyses describe *endpoint* performance as a function of scale. Our work asks a complementary, finer-grained question: how do the *dynamics* of word-level learning change across scales, and can the global rise in inter-word coherence be dissociated from semantically specific coupling through baseline correction?

Probing and representational geometry Probing studies examine what information is encoded in static LLM representations at a given training stage (Hewitt and Liang, 2019; Voita and Titov, 2020). These methods ask whether a target property can be decoded from frozen activations. Our approach is complementary: rather than probing geometric properties of representations, we track *behavioural* evidence of learning over time via surprisal dynamics. The two perspectives are independent (a model can encode syntactic geometry early whilst still exhibiting non-monotonic surprisal dynamics) and both are needed for a full picture of lexical acquisition.

Non-monotonic learning and catastrophic forgetting The non-monotonic trajectories we observe echo, but differ from, catastrophic forgetting (McCloskey and Cohen, 1989; Kemker et al., 2018): in that literature, sequential learning on new tasks degrades performance on old ones. Here, the non-monotonicity occurs *within* a single training run on a fixed dataset, arising from optimisation fluctuations rather than distribution shift. Measuring trajectory shape via smoothness and half-life rather than only endpoint performance or forgetting rate thus requires dedicated metrics, three of which we introduce and compare here.

3 Experimental Setting

Our goal is to understand how LLMs acquire word meaning over the course of training. Following Chang et al. (2024), we operationalise this by tracking how the surprisal of individual words evolves across training checkpoints. Beyond describing individual trajectories, we pursue two further objectives: understanding what *drives* differences in trajectory shape (via corpus-level word features), and comparing trajectories *across* words, specifically across synonyms and inflected forms, to test whether models abstract away from surface form during learning. We describe in turn the models and

evaluation protocol, the metrics used to quantify trajectory shape, and the wordbanks and word-level features on which the analyses are computed.

3.1 Models and Evaluation Protocol

Models We consider `Pythia` (Biderman et al., 2023), a suite of GPT-based transformers trained on a deduplicated subset of `The Pile` (300 billion tokens), with models ranging from 70 million to 12 billion parameters[2] and 153 checkpoints saved at regular intervals. A key strength of `Pythia` is that its training data are fully documented, allowing us to track exactly how often and in which contexts each word has been seen at any given checkpoint, a property we exploit directly in our analysis of frequency effects (Section 4).

Implementation For each word in the wordbank, we collected 30 sentences from Reddit posts published after 2020, so as to minimise overlap with `The Pile`.[3] At every checkpoint, we computed the model's estimated probability of each word given its context and report the average surprisal across all selected sentences, using the method of Pimentel and Meister (2024) for subword-based probability estimation.

3.2 Quantifying Trajectory Shape

Characterising learning trajectories requires capturing two complementary aspects of their shape: how *quickly* surprisal converges, and how *smoothly* it evolves over time.

Convergence speed A single parametric summary of convergence speed is insufficient: as previous works have shown, surprisal curves are non-monotonic, so the global rate of decline, the integrated surprisal burden, and the timing of early convergence can each behave differently. We therefore use three complementary metrics, all new with respect to the threshold-based measure of Chang et al. (2024):

- The slope a of a linear fit $S_w(t) = a \times t + b$ captures the *overall* rate of decline across all checkpoints.[4]
- The normalised AUC, $\text{AUC} = \int_0^{T_{\max}} S_w(t)\,\mathrm{d}t \;/\; (T_{\max} \cdot S_w(0))$, aggregates surprisal over the full training run. Unlike the slope, it is sensitive to the shape of the *entire* trajectory, including transient increases: a word that temporarily spikes in surprisal mid-training accumulates more area even if its final level matches that of a smoother curve. Lower AUC indicates faster overall convergence.
- The half-life $t_{1/2}$ is the first training step at which surprisal drops halfway to its minimum. It focuses specifically on *early* convergence dynamics, is model-free, and is insensitive to late-training fluctuations. Lower $t_{1/2}$ indicates faster early convergence.

Together, these three metrics allow us to disentangle whether model scale primarily changes the speed of initial acquisition, the asymptotic level reached, or the total surprisal accumulated over training.

Smoothness As shown by Chang et al. (2024) and Shafiabadi and Wisniewski (2025), not all surprisal curves that converge at the same speed do so in the same way: some decline steadily, others oscillate or lurch before settling. Smoothness captures this variability in trajectory shape independently of convergence speed, allowing us to ask whether erratic learning is a property of specific words (e.g. highly polysemous or infrequent ones) or of model capacity more broadly.

We measure smoothness via the variogram at lag $h = 1$:

$$\gamma(1) = \frac{1}{2(T-1)} \sum_{t=1}^{T-1} \big(S_w(t+1) - S_w(t)\big)^2 \quad (1)$$

where T is the total number of checkpoints.[5] Intuitively, $\gamma(1)$ is the average squared step between consecutive checkpoints: if surprisal moves in small, regular increments, $\gamma(1)$ is small; if the curve lurches, jumping up at one checkpoint and falling sharply at the next, $\gamma(1)$ is large.

It therefore measures *local* irregularity throughout the entire training run, making it sensitive to sudden jumps wherever they occur. This contrasts with

[2] We analysed all models up to 2.8b parameters; inference costs for larger models were prohibitive, and preliminary results suggested qualitatively similar behaviour to the 2.8b model.

[3] Our code and data will be made available upon publication.

[4] The linear model is the best-fitting model for all words in our dataset. After LOESS smoothing, mean R^2 ranges from 0.32 to 0.60 across model sizes; on raw curves 0.16–0.26, reflecting the non-monotonic nature of the trajectories.

[5] Although variograms are commonly used in geostatistics for spatial data, they are well-defined for one-dimensional time series (Cressie, 1993). We used the implementation of Mälicke et al. (2022).

the *variability (steps)* metric of Chang et al. (2024), which captures fluctuations as the standard deviation of surprisal over the last 25 % of pre-training only. Figure 3 (Appendix B) illustrates the range of curve shapes associated with different values of $\gamma(1)$.

3.3 Measuring Trajectory Similarity

Comparing learning trajectories across words requires a similarity measure robust to non-linear dynamics. We use distance correlation (dcor) introduced by (Székely et al., 2007) as the primary measure: it quantifies dependence between two time series of arbitrary length, detects both linear and non-linear associations, and returns a value in $[0, 1]$, where 0 indicates independence.

To guard against artefacts specific to any single measure, we also report three complementary statistics. Kendall's τ is rank-based and captures monotone co-movement between two curves, making it sensitive to consistent ordering even when the relationship is non-linear. Chatterjee's ξ (Chatterjee, 2021) measures whether one curve is approximately a deterministic function of the other, without assuming a direction of effect; because it is asymmetric and specifically sensitive to functional dependence, it provides a view that is qualitatively different from dcor. Mutual information (Cover and Thomas, 1991) makes no assumptions about the form of the relationship and captures any statistical dependence, including non-linear and non-monotone associations.

Each measure thus has different sensitivities and failure modes; agreement across all four provides strong evidence that observed patterns are not methodological artefacts.

3.4 Wordbanks and Word-Level Features

Wordbanks We use three wordbanks depending on the properties under investigation. The *primary wordbank* comprises 200 words sampled to cover a broad range of training frequencies. We filtered the vocabulary to retain words occurring more than 100,000 times and consisting exclusively of alphabetical characters,[6] computed decile thresholds on word counts, and sampled up to 20 words per decile. This quantile-based strategy ensures balanced coverage across the frequency range, in keeping with the Zipfian distribution of natural language.

[6]This threshold eliminates foreign words, misspellings, URLs, and non-linguistic tokens.

Two supplementary wordbanks target specific linguistic phenomena. The first consists of 40 *synonym pairs* identified by GPT-4o, used to test whether semantically equivalent words follow similar acquisition trajectories. The second comprises 30 *verbs* represented in four morphological forms (infinitive, third-person singular, past tense, present participle), yielding 120 items; base-form frequency was varied systematically to disentangle form identity from frequency effects.

Word-level features To investigate what drives differences in trajectory shape, we associate each word with three corpus-level features.

Training frequency is the total number of occurrences of the word in `The Pile`, as documented in the Pythia training logs. Frequency is a well-established predictor of lexical acquisition speed (Chang et al., 2024): high-frequency words receive more gradient updates per checkpoint and are expected to converge earlier and more smoothly. Frequency values are binned into deciles to allow non-parametric group comparisons.

Semantic ambiguity is operationalised as the number of senses attested in the UFSAC corpus (Vial et al., 2018), which provides context-grounded WordNet sense annotations drawn from a large balanced corpus. We prefer this count over the full WordNet inventory because it reflects senses that are actually attested in naturally occurring text, rather than the theoretical maximum. Polysemous words may exhibit more erratic trajectories if different senses are reinforced by different distributional contexts across training. Sense counts are binned into four groups via the Jenks natural breaks method (Jenks, 1967), which minimises within-group variance.

Subword token count is the number of subword pieces into which the word is segmented by the Pythia tokeniser. This feature is introduced here as a proxy for morphological complexity: words split into many subwords receive weaker direct signal per occurrence, since the model's prediction target is spread across multiple tokens rather than concentrated on a single one. This is expected to increase trajectory irregularity independently of raw frequency.

4 Results

Surprisal Trajectories Across Model Sizes Our first experiment reproduces prior observations by examining the evolution of surprisal across all three

	70M	160M	410M	1.4B
70M	—	0.704 ±0.204	0.730 ±0.200	0.692 ±0.184
160M		—	0.843 ±0.168	0.798 ±0.152
410M			—	0.863 ±0.121

Table 1: Mean (± std) dcor between the surprisal curves of the same words computed by two different Pythia model sizes (frequency corpus, n=155 words). Higher values indicate more similar learning trajectories.

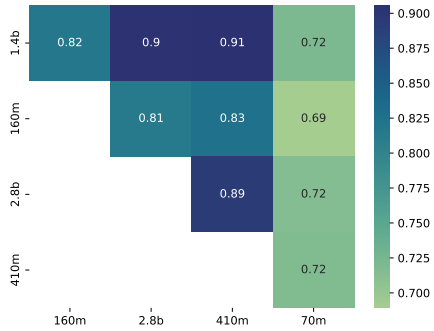

Figure 1: Similarity between surprisal curves across Pythia model sizes: mean dcor between the trajectories of the same 155 words as computed by each pair of models.

	Smoothness	Convergence speed			
	$\bar{\gamma}_1$ [95 % CI]	$\bar{a}$ [95 % CI] $\times 10^{-5}$, $a<0 \Rightarrow \searrow$	AUC [95 % CI]	$\bar{t}_{1/2}$ [95 % CI], k-steps	$\searrow$ surp.
Pythia					
70M	3.46 [2.9, 4.2]	−1.9 [−2.2, −1.6]	0.667 [0.64, 0.70]	5.0 [3.2, 7.5]	76.8 %
160M	3.55 [2.9, 4.5]	−1.0 [−1.2, −0.88]	0.611 [0.58, 0.64]	3.7 [1.8, 6.2]	55.5 %
410M	3.85 [3.1, 4.8]	−1.1 [−1.2, −0.95]	0.576 [0.54, 0.61]	2.0 [1.6, 2.4]	88.4 %
1.4B	3.76 [3.1, 4.7]	−1.0 [−1.1, −0.93]	0.552 [0.52, 0.58]	0.22 [0.03, 0.47]	93.5 %

Table 2: Learning curve dynamics by model size. $\bar{\gamma}_1$: mean variogram (smoothness). $\bar{a}$: mean slope ($\times 10^{-5}$; $\downarrow$ = faster). AUC: mean normalised area under curve ($\downarrow$ = faster). $\bar{t}_{1/2}$: mean half-life in k steps ($\downarrow$ = faster early convergence). $\searrow$: proportion with significant decreasing trend (Mann–Kendall, BH-FDR, $\alpha = 0.05$). CIs: 95 % BCa.

wordbanks, verifying whether acquisition is non-monotonic (Chang et al., 2024) and whether surprisal does not necessarily decrease during training (Shafiabadi and Wisniewski, 2025). Table 2 reports the distribution of the four convergence metrics and the proportion of words with a decreasing trend.

The results are consistent with prior observations and reveal a clear scaling effect. The linear slope a shows that the proportion of words with declining surprisal varies substantially: from 55.5 % (160M) to 93.5 % (1.4B). The AUC decreases monotonically with scale (from 0.667 at 70M to 0.552 at 1.4B), indicating that larger models accumulate substantially less surprisal and converge more consistently across the entire training run. The half-life $t_{1/2}$ reveals the sharpest scaling effect: the 1.4B model reaches its halfway point in 0.22×10^3 steps (0.15 % of total training), compared to 5.0×10^3 steps (3.5 %) for the 70M model; the wide CI for 1.4B ($[0.03 \times 10^3$ steps, 0.47×10^3 steps$]$) further reflects high word-to-word variability in early acquisition.

These three metrics tell a consistent story: scale accelerates convergence, reduces total surprisal burden, and compresses early acquisition into a vanishingly small fraction of training. Smoothness ($\gamma(1)$), however, does not follow this monotone pattern (3.46 at 70M, 3.85 at 410M): trajectory regularity is not a simple function of capacity. This dissociation is informative, as convergence speed and smoothness capture distinct phenomena that a single metric would conflate. The 160M model is a notable outlier: despite its intermediate size, it has the lowest proportion of decreasing words (55.5 %), whilst its AUC (0.611) and $t_{1/2}$ (3.7×10^3 steps) are consistent with the broader size ordering, suggesting its high non-monotonicity is a local trajectory property rather than a global convergence shortfall.

A direct comparison using dcor reinforces this picture: pairwise correlations range from 0.69 (1.4B vs. 70M) to 0.86 (1.4B vs. 410M), with most pairs exceeding 0.80 (Table 1). Taken together, the convergence metrics and trajectory-similarity analysis converge on the same conclusion: learning dynamics are governed more by training-data statistics than by parameter count, with model capacity shaping the speed of acquisition but not its fundamental character.

Influence of Word Characteristics on Learning Dynamics To assess whether word features influence surprisal curve dynamics, we compare all pairwise combinations of feature bins (Section 3) using three complementary statistical tests:

- The Kolmogorov–Smirnov (KS) test detects any difference in distribution shape, including tails and dispersion, without assuming a particular form of shift.
- The Mann–Whitney U (MW) test targets stochastic dominance: whether values in one group are systematically larger than in another, making it sensitive to location shifts that KS

may miss.

- Cohen's d quantifies the standardised mean difference independently of sample size, allowing us to assess whether significant differences are also of meaningful magnitude.

All pairwise p-values within each (feature, model-size) family are corrected using BH-FDR ($\alpha = 0.05$). Together, KS, MW, and d provide a more complete picture than any single test: a feature may produce large effects ($d > 0.8$) even when formal significance rates are moderate.

Table 3 reports, for each model size and feature, the proportion of significant pairwise comparisons (KS and MW, BH-corrected) and mean Cohen's d.[7] The higher the proportion, the more evidence that a feature systematically influences curve dynamics.

The results reveal a clear hierarchy of robustness. Smoothness is the most robust outcome: subword token count produces 50 % to 67 % significant comparisons across all model sizes ($d = 1.5$–1.7), and word frequency 53 % to 64 % KS and 53 % to 67 % MW ($d = 1.0$–1.1). These large effect sizes reflect the direct link between distributional density and trajectory regularity: infrequent or highly sub-tokenised words receive sparser gradient updates, producing higher $\gamma(1)$ regardless of capacity. Semantic ambiguity affects smoothness at approximately 50 % of KS comparisons ($d \approx 0.5$), a modest signal likely reflecting the partial correlation between polysemy and frequency.

AUC shows a consistent frequency signal after correction: 53 % to 58 % significant KS and 40 % to 49 % MW ($d \approx 1.0$), confirming that frequency primarily affects total surprisal burden rather than instantaneous rate of decline.

The linear slope a, by contrast, loses most of its signal after BH correction: only the 70M model shows residual frequency sensitivity (47 %), whilst 160M, 410M, and 1.4B drop to 0 % to 7 %. This collapse indicates that the slope is dominated by noise and ill-suited as a summary statistic. The dissociation between smoothness and AUC on one hand, and the slope on the other, is theoretically informative: smoothness captures step-to-step irregularity driven by local gradient density, AUC captures the global surprisal burden shaped by corpus frequency, and the slope conflates both while being sensitive to neither after correction.

[7] Results for WordNet senses are similar to those for UFSAC senses; see Appendix A.

	smoothness KS / MW (d)	conv. speed (a) KS / MW (d)	conv. speed (AUC) KS / MW (d)
#sub tokens (6 comparisons)			
70M	50 % / 67 % (d=1.5)	50 % / 50 % (d=1.6)	0 % / 0 % (d=0.5)
160M	50 % / 50 % (d=1.5)	33 % / 17 % (d=0.9)	0 % / 0 % (d=0.5)
410M	50 % / 50 % (d=1.5)	0 % / 0 % (d=0.5)	0 % / 0 % (d=0.6)
1.4B	67 % / 67 % (d=1.7)	0 % / 0 % (d=1.0)	0 % / 0 % (d=0.7)
# occ. in training data (45 comparisons)			
70M	53 % / 53 % (d=1.1)	47 % / 47 % (d=1.0)	53 % / 44 % (d=1.0)
160M	64 % / 67 % (d=1.0)	7 % / 7 % (d=0.5)	56 % / 49 % (d=1.0)
410M	58 % / 58 % (d=1.0)	0 % / 0 % (d=0.4)	58 % / 42 % (d=1.0)
1.4B	60 % / 62 % (d=1.1)	7 % / 4 % (d=0.5)	53 % / 40 % (d=1.0)
# UFSAC senses (6 comparisons)			
70M	50 % / 50 % (d=0.5)	50 % / 33 % (d=0.6)	33 % / 33 % (d=0.5)
160M	50 % / 50 % (d=0.5)	33 % / 0 % (d=0.6)	33 % / 33 % (d=0.6)
410M	50 % / 33 % (d=0.5)	0 % / 0 % (d=0.3)	33 % / 33 % (d=0.6)
1.4B	50 % / 33 % (d=0.5)	0 % / 0 % (d=0.4)	33 % / 33 % (d=0.6)

Table 3: Proportion of significant pairwise comparisons (KS = Kolmogorov–Smirnov; MW = Mann–Whitney U; d = Cohen's d). All p-values are BH-FDR corrected within each (feature, model-size) family ($\alpha = 0.05$). Cells show KS / MW % and mean effect size. Large d (>0.8) indicate practical significance even where significance rates are modest.

Taken together, these results suggest that learning dynamics are more tightly linked to statistical factors, such as token exposure frequency, than to linguistic factors such as semantic complexity or ambiguity. Statistical factors thus account for variation in curve *smoothness* and *convergence speed*, but they do not address whether models learn related words (synonyms or inflected forms of the same verb) in a coordinated way. Crucially, all model sizes are exposed to identical training data in the same order, so any systematic difference in semantic coupling across scales cannot be attributed to corpus statistics alone. This motivates a complementary analysis: rather than characterising individual trajectories, we ask whether the surprisal curves of semantically related word pairs co-evolve, and whether the strength of this coupling varies with model capacity.

Investigating the Role of Surface Form in Word Meaning Acquisition We now ask whether learning dynamics depend on the surface form of a word, or whether models abstract across surface forms to track something closer to semantic sense. A key methodological consideration is that any observed correlation between related-word curves must be

evaluated against a random baseline, since all word pairs share a common global learning dynamic (Section 4 above); we address this with an explicit gap analysis after reporting the absolute results.

We conduct two series of experiments: one comparing surprisal curves for synonym pairs, the other for inflected forms of the same verb. If the model fails to abstract from surface form, the surprisal trajectories of related words should not be closely coupled, since surprisal is partly determined by word-specific factors (frequency, subword structure) that differ across forms.

To contextualise the correlations, we compute, for each model, the mean correlation between 40 randomly selected word pairs as a baseline. Table 4 reports dcor as the primary measure alongside three additional measures (τ, ξ, and MI) reported here alongside dcor to ensure robustness.

Table 4 reveals four consistent patterns. Even random pairs show substantial correlations (dcor $\in [0.49, 0.87]$), reflecting the common downward trend shared by all words, a global baseline that must be accounted for before drawing conclusions about semantic coupling. Above this baseline, synonym pairs are more tightly coupled than chance at every scale, with the gap widening with model size (dcor $= 0.745$ at 70M vs. 0.953 at 1.4B). All three additional measures confirm this scaling trend (τ from $[0.29, 0.39]$ to $[0.37, 0.48]$, ξ from $[0.18, 0.26]$ to $[0.26, 0.35]$, and MI from $[0.33, 0.45]$ to $[0.40, 0.53]$) ruling out artefacts of any single method. Notably, ξ is lower than dcor in absolute terms (0.22–0.30 vs. 0.74–0.95), as expected: ξ specifically measures functional (asymmetric) dependence and is known to be conservative, whereas dcor measures symmetric mutual dependence. Its monotone increase with model size is nonetheless unambiguous. Finally, all four measures are statistically significant ($p < 0.05$) for $\geq 90\,\%$ of synonym pairs, confirming that the coupling is not driven by a small number of outlier pairs.

Baseline correction: global coherence vs. semantic specificity The absolute dcor values in Table 4 rise with scale for all three word-pair types, including random pairs (CI midpoint: 0.59 at 70M, 0.78 at 1.4B). This concurrent rise in the random-pair baseline is a confound: any scale-driven increase in synonym dcor could reflect either genuine improvement in semantic abstraction or a more general increase in inter-word surprisal coherence shared

	mean [95 % CI]		
	`synonyms` dcor / τ / ξ / MI	`verb` forms	random pairs
Pythia			
70M	0.745 [0.67,0.81] τ[0.29,0.39] ξ[0.18,0.26] MI[0.33,0.45]	0.81 $\pm$0.05	[0.49,0.69]
160M	0.859 [0.78,0.90] τ[0.27,0.40] ξ[0.20,0.29] MI[0.33,0.45]	0.85 $\pm$0.05	[0.60,0.78]
410M	0.908 [0.82,0.95] τ[0.34,0.45] ξ[0.23,0.33] MI[0.36,0.53]	0.90 $\pm$0.04	[0.70,0.86]
1.4B	0.953 [0.93,0.96] τ[0.37,0.48] ξ[0.26,0.35] MI[0.40,0.53]	0.90 $\pm$0.05	[0.70,0.87]
2.8B	0.90 $\pm$0.06	0.91 $\pm$0.04	0.77 $\pm$0.10

Table 4: Mean correlation between surprisal curves within word groups (95 % BCa bootstrap CIs, Efron and Tibshirani 1993). Verb-form and 2.8b entries: mean $\pm$ std from the full-corpus analysis.

by *all* word pairs. To separate these two sources of variation, we compute the baseline-corrected gap $\Delta = \text{dcor}_{\text{syn}} - \text{dcor}_{\text{rand}}$ for each model size, using the midpoint of the random-pair CI as the baseline (Figure 2).

The gap remains positive at all scales (0.13–0.17), confirming that synonym pairs are consistently more tightly coupled than chance. However, it does not grow with model size: a linear regression of Δ on log(params) yields $R^2 < 0.01$, $p = 0.96$, effectively zero slope. Whilst larger models therefore exhibit higher absolute dcor for synonyms, this increase is largely attributable to the scale-driven rise in global inter-word coherence rather than to an improvement in semantic abstraction per se.

The verb-form gap, by contrast, *decreases* with scale (from 0.22 at 70M to 0.12 at 1.4B), as the growing baseline erodes the apparent advantage of morphological relatedness. This asymmetry between the synonym and verb-form trajectories is theoretically informative. Synonymy represents a distributional overlap largely independent of surface form: synonym pairs occur in similar contexts irrespective of model size, so the surprisal coupling reflecting this overlap remains stable across scales. Morphological relatedness, by contrast, is tied directly to surface form: inflected variants share distributional contexts (especially at smaller scales where token-frequency effects dominate), but as the model grows and the global baseline rises, this surface-form signal is increasingly subsumed by the broader coherence shared by all word pairs.

The divergence between verb-form and synonym gap trajectories thus provides indirect evidence that larger models develop representations that more cleanly separate morphological from semantic organisation. More broadly, these results call for caution when interpreting absolute surprisal correla-

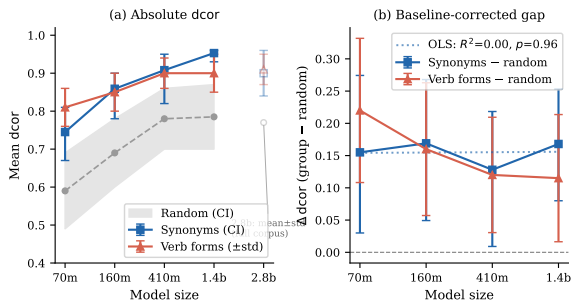

Figure 2: **(a)** Absolute dcor for synonyms, verb forms, and random pairs (shaded band = 95 % CI) across model sizes. **(b)** Baseline-corrected gap $\Delta = \text{dcor}_{\text{group}} - \text{dcor}_{\text{rand}}$ for synonyms and verb forms. The dashed horizontal line marks $\Delta = 0$; the dotted curve is the OLS fit of Δ on log(params) for synonyms ($R^2 < 0.01$, $p = 0.96$). Open markers in (a) denote 2.8b values from the full-corpus analysis (mean $\pm$ std).

tions as evidence of emergent semantic abstraction: the appropriate reference is not zero but the scale-specific random baseline.

5 Discussion and Conclusion

We set out to investigate two questions: whether model scale changes *what* LLMs learn or merely *how fast*, and whether the growing coupling between synonym surprisal curves with scale reflects genuine semantic abstraction or a more general rise in inter-word coherence. We address both by proposing trajectory comparison as a first-class analytical tool, applied to four Pythia model sizes across 153 training checkpoints.

Finding 1: scale drives global inter-word coherence, not semantic abstraction specifically. Absolute dcor between synonym surprisal curves increases monotonically with model size (0.745 at 70M to 0.953 at 1.4B), but baseline correction reveals that the *advantage* of semantic relatedness over random pairs is invariant with scale ($\Delta \approx 0.15$, $R^2 < 0.01$; Figure 2). What scales, therefore, is a general rise in inter-word surprisal coherence shared by all word pairs, not a specifically semantic improvement. Whether this global coherence itself reflects the emergence of structured representations or merely a tightening of the training signal remains an open question that absolute correlation values alone cannot settle.

Finding 2: learning dynamics are data-driven. Surprisal trajectories are highly similar across model sizes (dcor $\in [0.69, 0.86]$; Table 1). The AUC and half-life metrics, which disentangle total surprisal burden from early acquisition speed, confirm that larger models converge faster, but not along qualitatively different paths: the ordering and relative shape of trajectories are governed by training-data statistics, not by model capacity.

These findings have direct implications for evaluation. The data-driven nature of learning trajectories implies that benchmarks applied at a fixed checkpoint may conflate corpus-driven and capacity-driven effects; longitudinal evaluation is needed to disentangle them. The invariance of the synonym gap with scale implies that claims about emergent semantic abstraction must be grounded in baseline-corrected analyses rather than absolute correlation values, and that any claimed scaling effect should be tested against a scale-matched random baseline, a methodological precaution that generalises beyond this study.

Several directions follow naturally from this work. The most pressing is a controlled replication with frequency-matched synonym and random-pair sets, which would allow a cleaner test of whether the residual synonym gap ($\Delta \approx 0.15$) reflects genuine semantic coupling or a shared distributional profile. Extending the trajectory comparison framework to other linguistic levels (syntax, morphology, named entities among others) would test whether the data-driven pattern observed here is specific to lexical semantics or a general property of pre-training dynamics. Finally, applying the same methodology to instruction-tuned or RLHF-trained models would reveal whether post-training alignment disrupts the corpus-driven trajectory structure, or leaves it largely intact.

6 Limitations

First, our reliance on word-level surprisal may overlook contextual nuances and syntactic interactions that influence learning. Second, our experiments are limited to English; future work could explore multilingual generalisation.

Third, and most importantly given our baseline-correction result, disentangling global inter-word coherence from semantically specific coupling requires experiments with frequency-matched synonym and random-pair sets: the current synonym wordbank was not designed to control for differences in base-frequency between synonymous and randomly paired words, which may contribute to the observed gap independently of semantic relatedness. Nevertheless, the stability of the gap across

four model sizes ($\Delta \in [0.13, 0.17]$, $R^2 < 0.01$) is consistent with a genuine semantic signal, since a purely frequency-driven artefact would be expected to scale with the baseline rather than remain constant.

Addressing this limitation would allow a cleaner test of whether models develop *semantic* representations that transcend surface form, or whether the synonym advantage is itself an artefact of shared distributional profiles.

Finally, our experiments required substantial computational resources: surprisal trajectories were computed on a server equipped with two NVIDIA RTX A5000 GPUs (24 GB VRAM each), processing approximately 7,200 sentences (mean length: 28 tokens) at each of the 770 model–checkpoint pairs (5 Pythia sizes $\times$ 154 checkpoints), for a total of roughly 5.5 million model forward passes.

References

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar Van Der Wal. 2023. Pythia: a suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23, pages 2397–2430. JMLR.org.

Tyler A. Chang, Zhuowen Tu, and Benjamin K. Bergen. 2024. Characterizing learning curves during language model pre-training: Learning, forgetting, and stability. *Transactions of the Association for Computational Linguistics*, 12:1346–1362.

Sourav Chatterjee. 2021. A new coefficient of correlation. *Journal of the American Statistical Association*, 116(536):2009–2022.

Thomas M. Cover and Joy A. Thomas. 1991. *Elements of Information Theory*. Wiley, New York.

Noel A. C. Cressie. 1993. *Statistics for Spatial Data*, revised edition. Wiley, New York.

Bradley Efron and Robert Tibshirani. 1993. *An Introduction to the Bootstrap*. Chapman & Hall, New York.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2733–2743, Hong Kong, China. Association for Computational Linguistics.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, and 1 others. 2022. Training compute-optimal large language models. *Preprint*, arXiv:2203.15556.

George F. Jenks. 1967. The data model concept in statistical mapping. In *International Yearbook of Cartography*, volume 7, pages 186–190.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *Preprint*, arXiv:2001.08361.

Ronald Kemker, Marc McClure, Angelina Abitino, Tyler L. Hayes, and Christopher Kanan. 2018. Measuring catastrophic forgetting in neural networks. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pages 3390–3398, New Orleans, Louisiana, USA.

Tal Linzen and Marco Baroni. 2021. Syntactic structure from deep learning. *Annual Review of Linguistics*, 7(Volume 7, 2021):195–212.

Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Brussels, Belgium. Association for Computational Linguistics.

Michael McCloskey and Neal J. Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. In Gordon H. Bower, editor, *Psychology of Learning and Motivation*, volume 24, pages 109–165. Academic Press, New York.

Mirko Mälicke, Romain Hugonnet, Helge David Schneider, Sebastian Müller, Egil Möller, and Johan Van de Wauw. 2022. mmaelicke/scikit-gstat: A scipy flavoured geostatistical variogram analysis toolbox.

Tiago Pimentel and Clara Meister. 2024. How to compute the probability of a word. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18358–18375, Miami, Florida, USA. Association for Computational Linguistics.

Nazanin Shafiabadi and Guillaume Wisniewski. 2025. Beyond surprisal: A dual metric framework for lexical skill acquisition in LLMs. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6636–6641, Abu Dhabi, UAE. Association for Computational Linguistics.

Gábor J. Székely, Maria L. Rizzo, and Nail K. Bakirov. 2007. Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35(6):2769 – 2794.

	smoothness	speed of convergence
#wordnet senses (6 comparisons)		
70M	50.0 %	50.0 %
160M	50.0 %	33.3 %
410M	33.3 %	0.0 %
1.4B	50.0 %	33.3 %

Table 5: Effect of WordNet sense count on the dynamics of surprisal curves. AUC results are omitted as they are consistent with those reported for UFSAC senses in Table 3.

Loïc Vial, Benjamin Lecouteux, and Didier Schwab. 2018. UFSAC: Unification of sense annotated corpora and tools. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Elena Voita and Ivan Titov. 2020. Information-theoretic probing with minimum description length. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 183–196, Online. Association for Computational Linguistics.

A Impact of WordNet senses

Table 5 presents the effect of the number of senses listed in WordNet.

B Examples of surprisal curves

Figure 3 shows representative surprisal trajectories spanning the full range of smoothness values ($\gamma(1)$) observed in the frequency corpus. Each panel displays one word across all 154 checkpoints; the y-axis is standardised across panels to allow direct comparison of curve shapes.

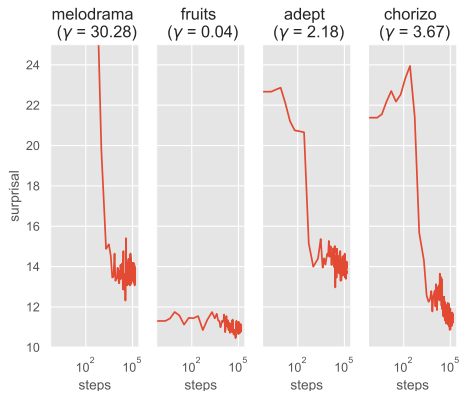

Figure 3: Examples of surprisal curves illustrating different $\gamma(1)$ values. The y-axis has been standardised across plots.