

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

Nazanin Shafiabadi[✉] and Guillaume Wisniewski[✉]

[✉] Sorbonne Université, ISIR, 75 005 Paris, France

[✉] Université Paris Cité, LLF, CNRS, 75 013 Paris, France

nazanin.shafiabadi@isir.upmc.fr, guillaume.wisniewski@u-pariscite.fr

Abstract

We investigate how model scale shapes lexical acquisition in LLMs, introducing trajectory comparison as a methodological tool for studying how individual words are learned over the course of training. Tracking surprisal dynamics across Pythia model sizes and checkpoints, we report two findings. First, synonym curves become more correlated at larger scales, but once corrected against a frequency-matched random baseline this advantage shows no detectable scale trend: what grows with scale is broad inter-word coherence, not synonym-specific coupling. Second, learning dynamics are shaped primarily by training data rather than by model capacity, larger models converging faster but along qualitatively similar 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 the next 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

¹Surprisal is the negative log-probability of a word given

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 154 intermediate checkpoints spanning the whole of pre-training, we track surprisal trajectories for several hundred words across five 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: *what* a model learns is shaped primarily by training-data statistics, with capacity governing convergence speed rather than trajectory shape.

Trajectory comparison across semantically related words. By comparing surprisal curves across synonym pairs and inflected forms, and correcting

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.

for a global baseline of inter-word coherence, we show that the advantage of synonyms over random pairs has no detectable scale trend once frequency is controlled: what scales is broad inter-word coherence, not synonym-specific coupling.

Metrics newly applied to 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 apply 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.

2 Related Work

Our work builds directly on Chang et al. (2024), who characterise individual surprisal trajectories across Pythia checkpoints, and on Shafabadi 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, taken up in turn below: scaling laws, probing and representational geometry, and non-monotonic learning and catastrophic forgetting.

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 rather than endpoint performance or forgetting rate calls for a different metric toolkit; the four measures set out in Section 3 are, to our knowledge, new to this problem.

3 Experimental Setting

Our goal is to understand lexical learning, as measured by surprisal, 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 to compare trajectories across words, and the wordbanks and word-level features underpinning the analyses.

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 parameters² and 154 public checkpoints,

²Inference costs beyond 2.8B were prohibitive. All three wordbanks were evaluated at all five sizes.

unevenly spaced: after step 0 they follow powers of two up to step 512, then a grid of 1,000 steps up to step 143,000, so early training is sampled far more densely than late training. 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 up to 30 sentences from Reddit posts published after 2020. The Pile was assembled and released in 2020 (Gao et al., 2020), so posts postdating it cannot appear in the training corpus verbatim.³ We capped collection at 30 sentences per word as a practical balance between per-word estimation stability and the cost of evaluating every sentence at all 154 checkpoints across up to five model sizes; actual coverage varies across words, since not every word yields 30 matching sentences meeting our minimum-score threshold, i.e. drawn from a post with a score of at least 30 (the Limitations section reports the resulting coverage). 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. Rather than scoring a single canonical BPE segmentation, this method marginalises over the subword tokenisations consistent with a word’s surface form.⁴ The same tokeniser is shared across all Pythia sizes, so a given word is segmented identically regardless of model size.

Confidence intervals Surprisal is computed per sentence at every model size, and all confidence intervals reported in this paper (including the trajectory-similarity and frequency-matched-baseline analyses of Section 4) are 95 % BCa bootstraps (Efron and Tibshirani, 1993) over this per-

³Code, including the frequency-matched baseline generation scripts, is available at https://gitlab.huma-num.fr/gwisniewski/lexical_learning_dynamics; the wordbanks (with Pile frequencies and subword-token counts), the evaluation corpora and the per-checkpoint surprisal data are archived under DOI <https://doi.org/10.34847/nk1.243621pi> (CC-BY-4.0). Together they enable exact replication of every table and figure in this paper.

⁴We use our own version of the `wordsp probability` library implementing this method, which fixes a checkpoint-cache leak and a tokeniser end-of-sequence bug affecting some Pythia checkpoints (Wisniewski, 2026); it is available at <https://gitlab.huma-num.fr/gwisniewski/probability-of-a-word>.

sentence distribution, with a fixed resampling seed so that the reported bounds are exactly reproducible.

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, none of which has previously been applied to surprisal trajectories, where prior work relies on a single threshold-based measure (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.⁵
- The normalised AUC, $\text{AUC} = \int_0^{T_{\max}} S_w(t) dt / (T_{\max} \cdot S_w(0))$, aggregates surprisal over the full training run. Unlike the slope, it is sensitive to the *entire* trajectory, including transient increases: a mid-training spike accumulates extra area even where the 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 overall rate of decline, or the total surprisal accumulated.

Smoothness As shown by Chang et al. (2024) and Shafabadi 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

⁵We fit a linear model for its interpretability, not because the trajectories are linear: after LOESS smoothing, mean R^2 ranges from 0.32 to 0.60 across model sizes, and on raw curves only 0.16–0.26, reflecting the non-monotonic nature of the trajectories. Section 4 shows that little of its signal survives p -value correction.

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 $h = 1$, a standard descriptor of local variability in time-series analysis:

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

where T is the total number of checkpoints.⁶ Intuitively, $\gamma(1)$ is the average squared step between consecutive checkpoints: small when surprisal moves in regular increments, large when the curve lurches.

It therefore measures *local* irregularity throughout the entire training run, making it sensitive to sudden jumps wherever they occur. This contrasts with 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 5 (Appendix I) 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

⁶Although 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 Mälikic et al. (2022).

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. Our *primary wordbank* comprises 200 words sampled to cover a broad range of training frequencies, of which the 155 that yielded usable evaluation sentences enter every analysis reported below (Appendix A). We filtered the vocabulary to retain words occurring more than 100,000 times and consisting exclusively of alphabetical characters,⁷ 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.

Two supplementary wordbanks target specific linguistic phenomena. The first consists of 20 *synonym pairs* identified by GPT-4o and manually validated against four criteria: substitutability in context, matching part of speech, absence of a dominant additional sense that would break the synonymy, and presence of both words in our surprisal data; the wordbank is 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; the six unordered pairs derived from these four forms for a given verb are referred to below as its *within-lemma pairs*; base-form frequency was varied systematically to disentangle form identity from frequency effects.

Frequency-matched baseline Because synonym pairs are not matched in frequency to random pairs by construction, we complement the random baseline used in Section 4 with a frequency-matched control. For each synonym pair (w_1, w_2) with Pile frequencies (f_1, f_2) , we sample a control pair (r_1, r_2) from the same $\log_{10}$ -frequency decile as w_1 and w_2 respectively, drawing candidates from

⁷This constraint eliminates foreign words, misspellings, URLs, and non-linguistic tokens.

words already present in our surprisal data to guarantee computable trajectories. We repeat this procedure over 200 independent draws and report the mean and 95 % BCa bootstrap confidence interval of $d_{\text{cor,rand,matched}}$ across draws, together with the empirical p -value (the proportion of draws in which the matched baseline reaches the synonym value).

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 prediction target is spread across several tokens rather than concentrated on one. We expect this to increase trajectory irregularity independently of frequency.

4 Results

Surprisal Trajectories Across Model Sizes Our first experiment reproduces prior observations by examining the evolution of surprisal across all three wordbanks, verifying whether acquisition is non-monotonic (Chang et al., 2024) and whether surprisal does not necessarily decrease during training (Shafabadi and Wisniewski, 2025). Table 1

	Smoothness	Convergence speed		
	$\bar{\gamma}_1$ [95 % CI]	AUC [95 % CI]	$t_{1/2}$ [95 % CI], k-steps	$\searrow$ surp.
<i>Pythia</i>				
70M	3.46 [2.9, 4.2]	0.667 [0.64, 0.70]	5.0 [3.3, 7.9]	76.1 %
160M	3.55 [2.9, 4.5]	0.611 [0.58, 0.64]	3.7 [2.1, 7.1]	54.2 %
410M	3.85 [3.2, 4.8]	0.576 [0.55, 0.61]	2.0 [1.6, 2.5]	88.4 %
1.4B	3.76 [3.1, 4.7]	0.552 [0.52, 0.58]	0.22 [0.05, 0.64]	93.5 %
2.8B	4.13 [3.4, 5.2]	0.543 [0.51, 0.57]	2.7 [1.9, 4.9]	89.7 %

Table 1: Learning curve dynamics by model size (five Pythia sizes). $\searrow$: proportion of words with a significant decreasing trend (Mann–Kendall, BH-FDR, $\alpha = 0.05$). The mean linear slope $\bar{a}$, redundant here with the more robust non-parametric $\searrow$ trend test, is reported in Table 10 (Appendix E).

reports the smoothness and convergence-speed metrics together with the proportion of words with a decreasing trend (the linear slope is deferred to Appendix E).

The results are consistent with prior observations and reveal a clear scaling effect. AUC falls monotonically with scale: larger models accumulate less surprisal over the run. The half-life shows the sharpest effect, compressing early acquisition from thousands of steps at 70M to a few hundred at 1.4B, under 1 % of training. Its confidence interval at that size is correspondingly wide, so the compression is far from uniform across words. Unlike AUC, however, the half-life does not keep falling: at 2.8B it rises back to 2.7×10^3 steps. This reversal coincides with the least regular trajectories of the suite ($\bar{\gamma}_1 = 4.13$, the highest value in Table 1), which suggests that the half-life is poorly identified at that scale rather than that early acquisition genuinely slows down.

These metrics tell a largely consistent story: scale reduces the total surprisal burden monotonically (AUC) and compresses early acquisition into a vanishingly small fraction of training, with the exception of the 2.8B half-life discussed above. Smoothness ($\bar{\gamma}_1$) runs the other way, rising with scale (3.46 at 70M to 4.13 at 2.8B) and not even monotonically (3.85 at 410M against 3.76 at 1.4B): larger models have *less* regular trajectories, though the intervals overlap throughout and this is a trend rather than a significant difference. 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 (54.2 %), whilst its AUC (0.611) and $t_{1/2}$ (3.7×10^3 steps) are consis-

	70M	160M	410M	1.4B	2.8B
70M	—	0.704 $\pm$ 0.204	0.730 $\pm$ 0.200	0.692 $\pm$ 0.184	0.727 $\pm$ 0.208
160M		—	0.843 $\pm$ 0.168	0.798 $\pm$ 0.152	0.839 $\pm$ 0.169
410M			—	0.863 $\pm$ 0.121	0.901 $\pm$ 0.141
1.4B				—	0.861 $\pm$ 0.117

Table 2: Mean ($\pm$ std) dcor between the surprisal curves of the same words computed by two different Pythia model sizes (primary wordbank). Higher values indicate more similar learning trajectories.

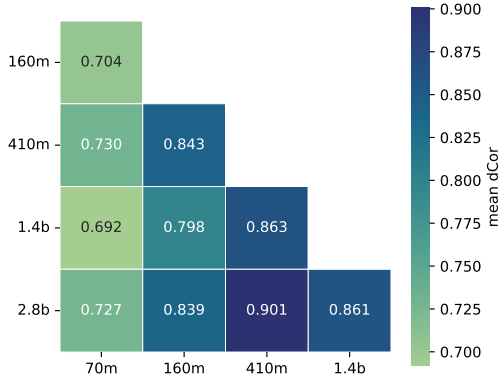


Figure 1: Similarity between surprisal curves across Pythia model sizes (five sizes, 70M–2.8B): mean dcor between the trajectories of the same 155 words as computed by each pair of models.

tent 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 0.90 (2.8B vs. 410M); every pair that excludes the 70M model reaches ≈ 0.80 or above (0.798–0.901), whereas every pair involving it stays near 0.70 (0.692–0.730); Table 2 gives the full matrix and Figure 1 the same values as a heatmap; Appendix G repeats the comparison under τ , ξ and MI. Taken together, these analyses converge on the same conclusion: capacity shapes how fast a word is acquired, not the shape of the trajectory. Since every Pythia size sees the same data in the same order, what governs trajectory shape is something the sizes share rather than what distinguishes them; the training data is the candidate we can interrogate directly, though the Limitations section sets out why it cannot be isolated here from the shared architecture and optimiser. Which of its properties are responsible is taken up next.

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 .⁸ 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 %–67 % significant comparisons across all model sizes (d : 1.5–1.7), and word frequency 53 %–64 % KS and 53 %–67 % MW (d : 1.0–1.1). These large effect sizes reflect the direct link between distributional density and trajectory regularity: infrequent or highly subtokenised words receive sparser gradient updates, producing higher $\gamma(1)$ regardless of capacity. For instance, a common, single-token word such as *family* (56.8M occurrences, top frequency decile) receives many concentrated gradient updates per checkpoint, whereas a rarer, more heavily subtokenised word such as *neurologist* (4 subwords, 188K occurrences, below-median frequency) receives fewer, more fragmented ones. 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: 51 %–58 % significant KS and

⁸Results for WordNet senses are similar to those for UFSAC senses; see Appendix H.

	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 % ($d=0.7$)
2.8B	50 % / 50 % ($d=1.6$)	0 % / 0 % ($d=0.5$)	0 % / 0 % ($d=0.6$)
#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 % ($d=1.0$)
2.8B	58 % / 53 % ($d=1.1$)	0 % / 0 % ($d=0.4$)	51 % / 42 % ($d=0.9$)
#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 % ($d=0.6$)
2.8B	50 % / 33 % ($d=0.5$)	0 % / 0 % ($d=0.4$)	0 % / 0 % ($d=0.5$)

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.

40 %–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, 1.4B and 2.8B drop to 0 %–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.

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 synonym-specific 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 Surprisal-Based Lexical Learning 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); 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 over 20 pairs formed from 40 randomly drawn words. Table 4 reports dcor as the primary measure, alongside τ , ξ and MI as robustness checks.

Figure 2 illustrates the contrast on one representative pair of each kind. Table 4 reveals three consistent patterns. Even random pairs are substantially correlated, reflecting the downward trend shared by all words: a global baseline that must be accounted for before any claim about synonym-specific coupling. Above that baseline, synonym pairs are more tightly coupled than chance at every scale, and all three additional measures move in the same direction with size, which rules out an artefact of any single statistic. The decisive point, however, is that the random-pair baseline itself rises over most of the range (0.551 at 70M to 0.760 at 1.4B, before falling back at 2.8B): absolute coupling values cannot on

	dcor, mean [95 % CI]		
	synonyms	verb forms	random pairs
<i>Pythia</i>			
70M	0.745 [0.67, 0.81]	0.648 [0.59, 0.70]	0.551 [0.46, 0.65]
160M	0.846 [0.75, 0.89]	0.749 [0.69, 0.80]	0.642 [0.55, 0.73]
410M	0.908 [0.75, 0.95]	0.784 [0.71, 0.84]	0.717 [0.61, 0.81]
1.4B	0.941 [0.91, 0.96]	0.782 [0.72, 0.83]	0.760 [0.65, 0.84]
2.8B	0.931 [0.87, 0.96]	0.829 [0.77, 0.87]	0.659 [0.53, 0.77]

Table 4: Distance correlation between surprisal curves within word groups, all five sizes from the same per-sentence pipeline. Synonyms: 20 pairs (Appendix B). Verb forms: the six within-lemma pairs of each of the 30 lemmas (Appendix C), i.e. 180 pairs. The τ , ξ and MI breakdown is reported in Table 11 (Appendix F); all four measures are significant ($p < 0.05$) for at least 90 % of synonym pairs at every size.

their own establish that larger models represent synonymy any better, which is what the baseline correction below sets out to test.

Baseline correction: global coherence vs. semantic specificity Absolute dcor broadly rises with scale for all three pair types, random pairs included — not monotonically, since both the synonym and the random values fall back at 2.8B, but over the bulk of the range. That concurrent rise is a confound: a larger model’s higher synonym dcor may reflect genuine semantic coupling, or merely a general tightening of inter-word coherence. We therefore work with the baseline-corrected gap $\Delta = \text{dcor}_{\text{syn}} - \text{dcor}_{\text{rand}}$, taking the mean of the random-pair distribution reported in Table 4 as the baseline (Figure 4).

The gap stays positive at every scale — synonyms are consistently more coupled than chance — but shows no reliable growth with size ($R^2 = 0.286$, $p = 0.353$). The rise in absolute synonym dcor is thus largely inherited from the global baseline rather than from any gain in synonym-specific coupling. This unmatched comparison is superseded by the frequency-matched analyses presented later in this section and in Appendix D, which provide the primary evidence for the synonym and verb-form gaps respectively.

The verb-form gap behaves differently: against a frequency-matched baseline it falls from 70M to 410M and then levels off, though the downward trend is not significant over five sizes ($R^2 = 0.614$, $p = 0.117$), a limitation of statistical power we return to in the Limitations section. This matched baseline covers only the infinitive/third-person-singular pair rather than all six within-lemma pairs (Appendix D), so it probes a narrower contrast than

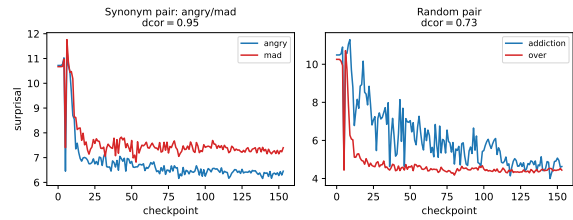


Figure 2: Example surprisal trajectories (1.4B) for a representative synonym pair (*angry/mad*, $\text{dcor} = 0.95$) and a representative random pair (*addiction/over*, $\text{dcor} = 0.73$), illustrating the qualitative difference between synonym-specific coupling and chance overlap. Pairs are the closest to the median dcor within their group, not cherry-picked. The random example is drawn from a wider pool (100 pairs) to locate the median more reliably; that pool never contributes to a reported value.

the dcor values of Table 4. Without frequency matching the raw gap is smaller and erratic, so here too the matched baseline is what makes the signal legible at all.

This asymmetry is theoretically informative. Synonymy is a distributional overlap largely independent of surface form, so its coupling holds steady across scales. Morphological relatedness is tied to surface form: inflected variants share contexts most at small scales, where token-frequency effects dominate, and as the global baseline rises this surface signal is progressively absorbed into the coherence shared by all pairs. The divergence therefore suggests that larger models separate morphological from semantic organisation more cleanly. More broadly, it warns against reading absolute surprisal correlations as evidence of emergent semantic structure: the reference point is not zero but the scale-specific random baseline.

Frequency-matched baseline: ruling out a frequency artefact The unmatched random baseline above is not controlled for frequency, and frequency is a well-established predictor of surprisal dynamics (Section 3); the synonym gap could in principle reflect a frequency confound rather than a semantic one. Table 5 reports dcor_{syn} against the frequency-matched baseline for all five model sizes. At every scale, dcor_{syn} exceeds the upper bound of the baseline’s 95 % BCa CI, and none of the 200 matched draws reaches the synonym value at any size (empirical $p < 0.005$ throughout), ruling out the frequency-artefact hypothesis. The frequency-matched gap (Δ_{freq} : 0.181–0.218) is comparable to the unmatched gap (Δ : 0.18–0.27) and markedly more stable across scales. As with the unmatched

Model	dcor_{syn}	$\overline{\text{dcor}}_{\text{rand}}$	95 % CI	Δ_{freq}
Pythia-70M	0.745	0.564	[0.558, 0.569]	0.181
Pythia-160M	0.846	0.638	[0.632, 0.644]	0.208
Pythia-410M	0.908	0.723	[0.717, 0.730]	0.185
Pythia-1.4B	0.941	0.722	[0.716, 0.729]	0.218
Pythia-2.8B	0.931	0.749	[0.742, 0.755]	0.183

Table 5: Distance correlation for synonym pairs vs. a frequency-matched random baseline (five Pythia sizes, 70M–2.8B; mean over 200 independent draws, 95 % BCa CI). $\Delta_{\text{freq}} = \text{dcor}_{\text{syn}} - \overline{\text{dcor}}_{\text{rand,matched}}$; regression of Δ_{freq} on $\log_{10}(\text{params})$: $R^2 = 0.032$, $p = 0.773$.

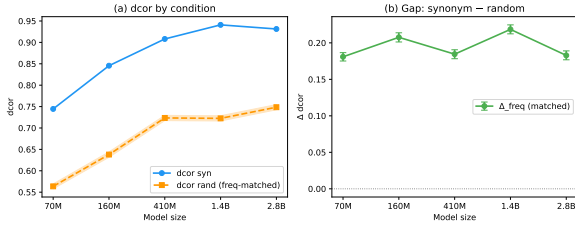


Figure 3: **(a)** dcor_{syn} against the frequency-matched random baseline (95 % BCa CI as a shaded band, too narrow to be visible at this scale) across model sizes. **(b)** Frequency-matched gap $\Delta_{\text{freq}} = \text{dcor}_{\text{syn}} - \text{dcor}_{\text{rand,matched}}$.

gap, Δ_{freq} shows no detectable scale trend (OLS regression on $\log_{10}(\text{params})$: $R^2 = 0.032$, $p = 0.773$).

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 correlation between synonym surprisal curves with scale reflects synonym-specific coupling or a more general rise in inter-word coherence. We address both by proposing trajectory comparison as a first-class analytical tool, applied to five Pythia model sizes across 154 training checkpoints.

Finding 1: scale drives global inter-word coherence, not synonym-specific coupling. Absolute dcor between synonym surprisal curves rises with model size (0.745 at 70M to 0.941 at 1.4B, 0.931 at 2.8B), but baseline correction against a frequency-matched random baseline reveals that the *advantage* of semantic relatedness over random pairs shows no detectable scale trend ($\Delta_{\text{freq}} \approx 0.20$, $R^2 = 0.032$, $p = 0.773$; Table 5, Figure 3). What scales, therefore, is a general rise in inter-word surprisal coherence shared by all word pairs, not a specifically semantic improvement. Whether that coherence reflects structured representations or

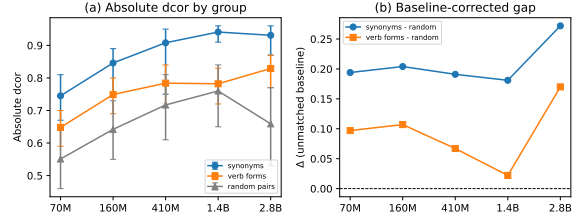


Figure 4: **(a)** Absolute dcor for synonyms, verb forms, and random pairs across model sizes. **(b)** Baseline-corrected (unmatched) gap $\Delta = \text{dcor}_{\text{group}} - \text{dcor}_{\text{rand}}$ for synonyms and verb forms; the dashed line marks $\Delta = 0$, i.e. no advantage over the scale-matched random baseline.

merely a tightening of the training signal is something absolute correlation values cannot settle.

Finding 2: learning trajectories are conserved across scale within the Pythia training setup. Surprisal trajectories are highly similar across model sizes ($\text{dcor} \in [0.69, 0.90]$; Table 2). 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. This behavioural conservation does not establish that larger models lack richer internal representations of word meaning; the representational interpretation should be read as a hypothesis, not a settled conclusion.

These findings have direct implications for evaluation, and yield three practical methodological lessons. First, any claimed scaling effect on semantically-driven behaviour should be tested against a scale-matched random baseline: the absence of a scale trend in the synonym gap only became visible once we corrected for the rise shared by *all* word pairs. Second, the conservation of trajectories across scale implies that benchmarks applied at a fixed checkpoint may conflate corpus- and capacity-driven effects; longitudinal evaluation is needed to disentangle them. Third, if learning dynamics are governed primarily by training-data statistics rather than capacity (Finding 2), then interventions on the data itself (composition, ordering) should shift surprisal trajectories more than interventions on model size — a prediction amenable to controlled pre-training experiments.

Two directions follow: probing representations across checkpoints, to test whether richer structure emerges at scale where behaviour is conserved; and extending it to syntax and morphology.

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, the conservation of surprisal trajectories across Pythia sizes does not, by itself, isolate training data as the causal driver of this conservation: all model sizes share not only the same training corpus but also the same architecture family, optimiser, and training-data order, so architecture- or optimiser-driven convergence could in principle also contribute to the observed similarity. Disentangling these factors would require comparing model families that vary one of data, architecture, or optimiser while holding the others fixed, a cross-suite comparison outside the scope of this study.

Fourth, the scale trends we report rest on five model sizes only. The regressions of Δ and Δ_{freq} on $\log_{10}(\text{params})$, and the apparent decline of the verb-form gap, are correspondingly underpowered: the null results ($p = 0.353$, $p = 0.773$ and $p = 0.117$ respectively) are consistent with the absence of a scale effect but cannot exclude a moderate one. Our claim is therefore that no trend is detectable at this resolution, not that none exists; settling the question would require either the larger Pythia sizes we could not afford to evaluate or a second suite with comparable checkpoint coverage.

Fifth, our surprisal estimates are computed on Reddit sentences published after 2020, chosen to minimise verbatim overlap with the training corpus (Section 3); the linguistic register and topical distribution of post-2020 Reddit discourse may nonetheless differ from the contexts in which a word was originally encountered during training, which could introduce a systematic, if hard to characterise, bias in the resulting surprisal estimates.

Sixth, per-word coverage is uneven. We capped collection at 30 sentences per word and applied the minimum-score threshold of Section 3, but Reddit yielded no usable sentence for 45 of the 200 words in the primary wordbank, which are absent from every analysis; among the 155 that remain, the median word is attested in 13 sentences and only 22 reach the cap. Surprisal estimates are correspondingly noisier for sparsely attested words. Appendix A reports what this attrition does and does not change: the subword-token distribution is unaffected, frequency is mildly skewed towards the retained set,

and the sense-count bins are the ones it thins.

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 5,690 distinct sentences (mean length: 32 tokens) at each of the 154 checkpoints and at each of the five model sizes, for a total of roughly 4.4 million model forward passes.

Ethics Statement

Our evaluation sentences are drawn from public Reddit posts. We retain no author-identifying meta-data (usernames, post identifiers, or subreddits) beyond the post score and coarse publication date used for filtering, and released data are limited to sentence text and per-checkpoint surprisal values. No annotation involving human subjects was carried out.

We use each artefact in accordance with its licence: the Pythia models and code are released under Apache 2.0, the Pile code under MIT and the Pile data under the heterogeneous licences of its constituent components, and UFSAC under MIT. Reddit content carries no open licence; we collected it under Reddit’s API terms and redistribute only the derived material described above.

Use of AI Assistants

We used Claude (Anthropic) to assist with language editing and proofreading of the manuscript. All scientific content and interpretation are the authors’ own.

Acknowledgements

We thank the anonymous reviewers for their detailed and constructive comments, which substantially improved this paper. This work was supported by the DEEPTYPE project, funded by the Agence Nationale de la Recherche (ANR-23-CE38-0003-01).

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.
- Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020. [The Pile: An 800GB dataset of diverse text for language modeling](#). *Preprint*, arXiv:2101.00027.
- 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, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, and 3 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.
- 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.
- Guillaume Wisniewski. 2026. [An experimental-campaign tool for surprisal computation](#). Technical Report hal-05727632, Université Paris Cité; Laboratoire de Linguistique Formelle. HAL Id: hal-05727632.

A Primary wordbank statistics

The 200-word primary wordbank is sampled evenly across frequency deciles by design (Section 3): each of the ten deciles contributes exactly 20 words, spanning 102,707 to 75,112,198 Pile occurrences. Table 6 summarises the distribution of subword-token count and sense count (WordNet and UFSAC) across the 200 words.

From 200 sampled words to 155 analysed All analyses in this paper use the 155 words of the primary wordbank for which we could collect Reddit sentences meeting the minimum-score threshold (Section 3); the remaining 45 yielded none. This attrition is not uniform across the features of Table 6, so it is worth stating what it does and does not affect. The subword-token distribution is unchanged (KS $p = 0.91$, MW $p = 0.21$; mean 2.2 subwords among retained words against 2.3 among discarded ones), so the subword-count results of Table 3 are not affected by it. Frequency is mildly skewed: 72 of the 100 words in the bottom five deciles survive against 83 of the top five (median Pile frequency 363,875 vs. 268,161; MW $p = 0.04$), so the retained set leans slightly frequent whilst still spanning the full range; the frequency deciles used in Table 3 are correspondingly uneven in size (about 14 words per decile in the bottom half against 17 in the top), which the non-parametric tests tolerate but which makes the per-decile proportions noisier at the low end. The clearest effect is on sense count: 36 of the 45 discarded words have no attested UFSAC sense at all, against 52 of the 155 retained, and retention rises from 70 % in the lowest sense bin to 95 %–100 % in the three higher ones. The sense-count comparisons of Table 3 and Appendix H therefore contrast bins whose lowest group is thinned relative to the sampling design. A second, independent caveat compounds this: those comparisons rest on only six pairwise bin comparisons per (feature, model-size) combination, a small cell count at which proportions of significant comparisons are an unstable summary. We report effect sizes (d) alongside these proportions throughout for this reason, and they should be read as the primary evidence of a feature’s influence, not the significance rates on their own. Together the two make the semantic-ambiguity results the weakest of the three feature analyses.

	median	range
# subwords	2	1–4
# WordNet senses	1	0–21
# UFSAC senses	1	0–22

Table 6: Primary wordbank (200 words): distribution of subword-token count and sense count. Subword-token breakdown: 23 words at 1 subword, 117 at 2, 55 at 3, 5 at 4. Both sense counts are right-skewed (median 1, mean 2.6 for WordNet and 1.8 for UFSAC).

B Synonym wordbank

Table 7 lists the full 20-pair synonym wordbank, with Pile occurrence and subword-token counts for each word.

Pair	freq. 1	tok. 1	freq. 2	tok. 2
angry / mad	4,685,683	2	3,722,403	1
ask / inquire	23,767,109	1	656,402	2
begin / start	49,850,461	1	48,199,112	1
big / large	57,674,640	1	55,198,625	1
buy / purchase	17,260,862	1	10,158,842	2
choose / select	16,807,018	1	20,689,642	1
clean / tidy	12,802,716	1	557,055	2
end / finish	123,485,933	1	8,023,989	1
fast / quick	17,486,793	1	11,395,947	1
fix / repair	9,974,274	1	6,388,009	2
happy / joyful	17,394,063	1	430,114	2
assist / help	4,671,686	2	72,035,249	1
concept / idea	10,923,520	1	31,074,834	1
job / occupation	29,979,247	1	2,497,034	2
adore / love	455,342	2	50,399,559	1
rich / wealthy	11,545,890	1	2,040,483	2
observe / see	5,359,071	2	186,199,956	1
intelligent / smart	2,860,663	2	7,832,731	1
inform / tell	2,950,407	1	36,669,742	1
believe / think	33,134,636	1	98,017,007	1

Table 7: Full synonym wordbank (20 pairs): Pile occurrence count (freq.) and Pythia subword-token count (tok.) for each word. Occurrence counts are whole-corpus token counts over the deduplicated Pile, taken from the merged per-step count file released with the code; subword-token counts are those recorded in the wordbank metadata.

C Verb-form wordbank statistics

The verb-form wordbank comprises 30 base verbs sampled across Pile frequency deciles (three per decile), each represented in four morphological forms (infinitive, third-person singular, past tense, present participle), for 120 items in total. Pile occurrence counts range from 55,308 to 64,729,842 (median 4,428,498). Table 8 summarises subword-token count and sense count across the 120 items.

	median	range
# subwords	2	1–3
# WordNet senses	5	1–38
# UFSAC senses	5	1–55

Table 8: Verb-form wordbank (120 items): distribution of subword-token count and sense count. Subword-token breakdown: 25 items at 1 subword, 79 at 2, 16 at 3.

D Verb-form gap: raw vs. frequency-matched baseline

Table 9 contrasts the verb-form gap computed against an unmatched random baseline (as plotted

in Figure 4b) with the same gap computed against a frequency-matched baseline (same procedure as Section 4, applied here to infinitive vs. third-person-singular pairs only, the only pair type this baseline currently supports). The unmatched gap is smaller at every size except 2.8B and erratic across scales, whereas the frequency-matched gap remains clearly positive at every size, including 2.8B. As in the main text, it falls from 70M to 410M (0.200 to 0.131) and then levels off rather than continuing to decline (0.126–0.141 over the top three sizes). As with the synonym result in the main text, this shows that an unmatched random baseline can obscure a genuine signal that a frequency-matched one recovers, reinforcing the case for frequency-matched baselines as standard practice rather than an optional robustness check.

Model	Δ_{raw}	Δ_{freq}
Pythia-70M	0.097	0.200
Pythia-160M	0.107	0.173
Pythia-410M	0.067	0.131
Pythia-1.4B	0.022	0.126
Pythia-2.8B	0.170	0.141

Table 9: Verb-form gap against an unmatched random baseline (Δ_{raw} , all six within-lemma pairs, mean of the Table 4 random-pair column) vs. a frequency-matched baseline (Δ_{freq} , infinitive/third-person-singular pairs only). Both wordbanks use the same 200-draw protocol as Section 4; Δ_{raw} does not depend on the number of draws.

E Mean linear slope by model size

Table 10 reports the mean linear slope $\bar{a}$ omitted from Table 1 in the main text. The slope is negative at every size, in line with the predominance of decreasing trends reported there, but its magnitude does not track the $\searrow$ proportions: the 70M model has the steepest mean slope and one of the lowest proportions of decreasing words. We therefore rely on the Mann–Kendall-based $\searrow$ statistic as the primary measure of directional trend, since it is a non-parametric per-word significance test rather than a single aggregate coefficient.

F Additional similarity measures

Table 11 reports the three complementary measures omitted from Table 4: Kendall’s τ , Chatterjee’s ξ and mutual information, for the synonym wordbank at each model size. All three broadly increase with scale, and their decline at 2.8B is shared with dcor

Model	$\bar{a}$ [95 % CI] ($\times 10^{-5}$)
Pythia-70M	−1.9 [−2.2, −1.6]
Pythia-160M	−1.0 [−1.2, −0.88]
Pythia-410M	−1.1 [−1.2, −0.95]
Pythia-1.4B	−1.0 [−1.1, −0.93]
Pythia-2.8B	−1.2 [−1.3, −1.1]

Table 10: Mean linear slope of surprisal over training by model size (five Pythia sizes, 70M–2.8B; $a < 0 \Rightarrow$ decreasing surprisal). 95 % BCa CIs.

(0.941 to 0.931), which is what rules out an artefact specific to any single statistic. They are not simply redundant with it, however: all three dip at 160M — markedly for τ and MI, marginally for ξ — where dcor instead rises (0.745 to 0.846), so the three measures are more sensitive than dcor to the local irregularity of the 160M trajectories noted in Section 4. ξ is markedly lower in absolute terms, as expected of a measure of functional (asymmetric) dependence.

Model	τ	ξ	MI
Pythia-70M	0.349 [0.29, 0.39]	0.223 [0.18, 0.26]	0.396 [0.33, 0.45]
Pythia-160M	0.319 [0.26, 0.37]	0.221 [0.19, 0.26]	0.364 [0.31, 0.42]
Pythia-410M	0.382 [0.32, 0.43]	0.251 [0.21, 0.29]	0.404 [0.34, 0.48]
Pythia-1.4B	0.420 [0.36, 0.48]	0.312 [0.26, 0.37]	0.556 [0.49, 0.64]
Pythia-2.8B	0.405 [0.34, 0.46]	0.280 [0.23, 0.32]	0.452 [0.38, 0.53]

Table 11: Kendall’s τ , Chatterjee’s ξ and mutual information between the surprisal curves of synonym pairs: mean over the 20 pairs, with its 95 % BCa CI.

G Cross-size similarity under the three complementary measures

Table 2 reports dcor only. Table 12 repeats the same computation, on the same 155 words and the same trajectories, under Kendall’s τ , Chatterjee’s ξ and mutual information.

All three are much lower than dcor in absolute terms, as expected of measures of monotone, functional and informational dependence respectively, so only their ordering is comparable. That ordering is largely preserved: across the ten size pairs, ξ and MI both rank the pairs as dcor does (Spearman $\rho = 0.95$), τ less closely ($\rho = 0.72$). The specific pattern highlighted in Section 4 — every pair involving 70M sits below every pair that excludes it — survives under MI (0.246–0.274 against 0.303–0.437) but, strictly, under neither ξ nor τ : under ξ the 160M–1.4B pair (0.140) falls just below two of the four pairs involving 70M (0.143 and 0.144), and under τ the 160M–2.8B pair scores lowest of all (0.165). The conclusion drawn in Section 4 should

therefore be read as resting on dcor and MI, with ξ and τ agreeing on the broad ordering of the ten pairs but not on the isolation of the smallest model.

Size pair	dcor	τ	ξ	MI
70M–160M	0.704	0.215	0.126	0.274
70M–410M	0.730	0.208	0.144	0.271
70M–1.4B	0.692	0.205	0.121	0.246
70M–2.8B	0.727	0.194	0.143	0.266
160M–410M	0.843	0.269	0.182	0.351
160M–1.4B	0.798	0.231	0.140	0.303
160M–2.8B	0.839	0.165	0.156	0.313
410M–1.4B	0.863	0.389	0.256	0.437
410M–2.8B	0.901	0.274	0.230	0.393
1.4B–2.8B	0.861	0.287	0.221	0.364

Table 12: Mean similarity between the surprisal curves of the same 155 words at two model sizes, under each of the four measures of Section 3. The dcor column reproduces Table 2.

H Impact of WordNet senses

Table 13 presents the effect of the number of senses listed in WordNet.

I Examples of surprisal curves

Figure 5 shows representative surprisal trajectories spanning the full range of smoothness values ($\gamma(1)$) observed in the primary wordbank. Each panel displays one word across all 154 checkpoints; the y-axis is standardised across panels to allow direct comparison of curve shapes.



Figure 5: Examples of surprisal curves illustrating different $\gamma(1)$ values. The y-axis has been standardised across plots.

	smoothness	conv. speed (a)
<i>#wordnet senses</i> (6 comparisons)		
70M	33.3 %	33.3 %
160M	33.3 %	0.0 %
410M	33.3 %	0.0 %
1.4B	50.0 %	0.0 %
2.8B	33.3 %	0.0 %

Table 13: Effect of WordNet sense count on the dynamics of surprisal curves (five Pythia sizes). Proportion of significant pairwise KS comparisons, BH-FDR corrected as in Table 3. The convergence-speed column reports the linear slope a ; AUC results are omitted as they are consistent with those reported for UFSAC senses in Table 3.