The Frequency Confound in Language-Model Surprisal and Metaphor Novelty

Anonymous ACL submission

Abstract

Language Model surprisal is widely used as a proxy for contextual predictability and has recently been reported to correlate with metaphor novelty. However, surprisal is tightly interwined with lexical frequency. We study this interaction on novelty scores of metaphoric words within their context. We analyse measures from 8 Pythia model sizes, and 154 intermediate checkpoints. Across settings, word frequency has stronger associations with novelty than surprisal. Across training stages, surprisal–novelty association peaks at an early stage and then falls again, mirroring a similarly timed increase in surprisal–frequency association. These results suggest that the often-reported optimal LM surprisal settings may incorrectly associate contextual predictability with novelty and processing difficulty.

1 Introduction

Linguistic creativity is often viewed as a distinctive capacity of human language use (Hockett, 1960; Yule, 2006). The recent fluency of state-of-the-art LLMs raises a natural question: to what extent do LLMs acquire linguistic creativity in ways that resemble human cognition, and how can we measure it reliably? (Dinu et al., 2025)

A well-known instance of creative language is *novel metaphor* (Bowdle and Gentner, 2005; Do Dinh et al., 2018). In conceptual metaphor theory, metaphors arise through mappings between a source and a target domain (Lakoff and Johnson, 1980). Crucially, metaphorical mappings vary in novelty: many metaphors are highly conventionalised, while novel metaphors introduce less familiar mappings and can require greater interpretive effort to understand (Cardillo et al., 2012; Philip, 2016). This makes metaphor novelty a useful testbed for studying creativity-related behaviour in LLMs, as it connects a graded phenomenon (conventional → novel) to processing effort.

A common computational proxy for contextual predictability is *surprisal*. Surprisal has a long history as an operationalisation of predictability and is widely used to model processing difficulty, including naturalistic reading behavior (Goodkind and Bicknell, 2018; Oh and Schuler, 2023a; Shain et al., 2024). However, surprisal is not a pure measure of contextual prediction; it is intertwined with lexical statistics such as word frequency (Shain, 2024; Tjuatja et al., 2025). Recent work has shown that both model scale and the amount of pretraining can substantially affect surprisal-based fits to human reading times, with word frequency playing a key explanatory role in these effects (Oh et al., 2024).

In the specific case of metaphor novelty, Momen et al. (2026) reported significant correlations between novelty annotations and LM surprisal, alongside systematic effects of model size. Building on these observations, we further investigate metaphor novelty through the lens of the surprisal–frequency interaction: when surprisal appears to correlate well with novelty, is it capturing contextual predictability, lexical frequency, or a mixture of both?

In this paper, we analyse associations between surprisal, word frequency, and metaphor novelty across (i) Pythia model sizes and (ii) pretraining stages, using intermediate checkpoints to probe training dynamics. Our results show that word frequency is a substantially stronger predictor of novelty than surprisal across settings, and that the configurations where surprisal performs "best" are also those where it aligns most closely with word frequency. We therefore caution against interpreting strong early-training or small-model surprisal–novelty associations as straightforward support for surprisal-based accounts of metaphor novelty, and we argue that progress requires clearer theoretical and methodological separation between predictability and lexical-frequency effects.

2 Data & Methods

This section describes our experimental set-up to measure the association between word frequency and surprisal on the one hand and metaphor novelty on the other hand.

2.1 Dataset

We base our study on the VU Amsterdam Metaphor Corpus (VUAMC) (Steen et al., 2010) (see Table 1). Every word in VUAMC is annotated as either a metaphoric word or not. Building on VUAMC, Do Dinh et al. (2018) collected crowdsourced metaphor novelty ratings for the 15,155 metaphoric content words in VUAMC and converted them into continuous scores in the range (-1, +1), where -1 denotes the most conventional and +1 the most novel. They additionally binarised these scores using a 0.5 threshold, resulting in 353 metaphors labelled as novel out of the 15,155 content metaphoric words. In our experiment, we use these 15,155 instances, each consisting of a sentence context, a target word (content metaphoric word), and an associated metaphor novelty score.

2.2 Model Suite

To examine the effects of model scale and pretraining progress (data/steps), we use the Pythia model suite (Biderman et al., 2023). Pythia consists of decoder-only causal LLMs at 8 sizes (70M–12B parameters), all trained on the same 300B-token pretraining corpus [1] in the same order. For each model, Pythia provides 154 intermediate checkpoints saved every 1,000 training steps (corresponds to additional ≈2M tokens seen during these steps), and denser early checkpoints at steps $\{1, 2, 4, 8, 16, 32, 64, 128, 256, 512\}$. The exact pretraining sequences seen at each checkpoint can be reconstructed using available scripts.

2.3 Surprisal

For causal LMs, surprisal, computed for a word w_i is $\text{Surprisal}(w_i) = -\log p(w_i \mid w_{<i})$[2]. In our experiment, we measure word-level surprisal for metaphoric target word(s) in their sentence-level context by running an independent, teacher-forced forward pass per sentence and recording the target word surprisal. To map token probabilities to a target word, we locate the word's character offsets in the sentence and sum token-level surprisals over the minimal token span in the sequence's tokenisation that covers these offsets. We additionally apply word-probability corrections for leading-whitespace / BOW-style tokenisation confounds (Pimentel and Meister, 2024; Oh and Schuler, 2024), and we prepend a BOS token so surprisal is defined even when the target is the first word in a sentence.

2.4 Word Frequency

We compute the *negative log frequency* of each target metaphoric word using two complementary estimates.

Negative Log Frequency in General Language Use: We compute the negative log frequency of each target metaphoric word using the Python library *wordfreq*[3], which provides corpus-independent frequency estimates aggregated from multiple large-scale sources, rather than deriving counts from a single corpus. We treat this as an estimate of word frequency for an "average English speaker", and hereafter denote it as **NLF-Human**.

Negative Log Frequency in Pythia's Pretraining Data: For each target metaphoric word, we tokenise its sentence using Pythia's tokeniser and, at each checkpoint, count occurrences of the target word's subtoken sequence in the pretraining tokens seen up to that checkpoint. We treat the negative log of these frequencies as a checkpoint-specific estimate of word frequency, and hereafter denote it as **NLF-LM**.

2.5 Experiment

We compute surprisal using all 8 Pythia model sizes and at each of the 154 checkpoints of the Pythia-70M variant. We likewise compute NLF-LM at each of these checkpoints, and NLF-Human (once) per metaphoric word. We quantify surprisal–novelty and NLF–novelty associations using Pearson's and Spearman's correlation coefficients, and we report the Area Under the ROC Curve (AUC) as an estimate of discriminating novel metaphors. Surprisal–NLF associations are estimated using Spearman's correlation.

3 Results

Detailed numerical results are provided in Appendix B. Figures 1–4 visually illustrate these results across model sizes and checkpoints.

[1] The Pile (Gao et al., 2020)

[2] We use log of base e for all *log* computations in our study.

[3] https://pypi.org/project/wordfreq/

3.1 Associations with Metaphor Novelty

Model Scale: Figure 1 shows surprisal–novelty and NLF–novelty association across Pythia model sizes. Here, NLF-LM is computed at the final checkpoint (300B tokens), and is therefore identical across sizes. Overall, NLF has a clearly stronger association with novelty than surprisal, with NLF-Human yielding slightly higher estimates ($\rho = .66, r = .66, AUC = .90$) than NLF-LM ($\rho = .63, r = .60, AUC = .90$). We also observe a consistent negative effect of model scale on the surprisal–novelty association.

Pretraining Progress: Figure 2 reports associations across the 154 checkpoints of Pythia-70M, with NLF-LM computed separately at each checkpoint. The NLF-LM–novelty association is essentially constant across checkpoints, except for a small deviation at the earliest steps (1–4). In contrast, the surprisal–novelty association is very weak in the first checkpoints, then rises sharply after 64 training steps (134M tokens), and peaks after 128 steps (268M tokens), where it approaches the NLF estimates ($\rho = .62, AUC = .90$). After this peak, the surprisal–novelty association converges to ($\rho = .45, AUC = .83$). Yet, it never reaches the same association strength as NLF.

3.2 Correlations between Surprisal and NLF

Model Scale: Figure 3 reports NLF–surprisal Spearman correlations across Pythia model sizes. Here, both NLF-Human and NLF-LM are fixed across sizes, while surprisal changes. Across sizes, surprisal shows moderate correlations with both NLF estimates ($\rho \in [.40, .61]$). Correlations decrease with model size, mirroring the negative scale effect observed for associations with novelty (Figure 1), suggesting that larger models' surprisal diverges from NLF. Overall, surprisal correlates slightly more with NLF-LM (max $\rho = .61$) than with NLF-Human (max $\rho = .57$).

Pretraining Progress: Figure 4 shows NLF–surprisal correlations across the 154 checkpoints of Pythia-70M. Here, NLF-Human is fixed, while NLF-LM and surprisal vary with checkpoint. The correlation pattern closely matches the trends in Figure 2: correlations are weak at the earliest checkpoints, rise sharply after 64 training steps (134M tokens), and peak after 128 steps (268M tokens), reaching $\rho = .95$ with NLF-LM and $\rho = .89$ with NLF-Human. Correlations then gradually converge over subsequent checkpoints.

4 Discussion and Conclusion

Word Frequency: Our results agree with previous work (Do Dinh et al., 2018; Reimann and Schefler, 2024) showing that word frequency (NLF) is strongly associated with metaphor novelty scores, and these associations are substantially stronger than those obtained from surprisal–novelty for any LM variant. Notably, frequency correlates more strongly with novelty than it does with surprisal itself. We further observe that NLF-Human aligns slightly more with novelty (human-based) than NLF-LM, whereas NLF-LM aligns slightly more with surprisal (LM-based) than NLF-Human. Overall, however, differences between the two frequency estimates are small and do not change overall trends, suggesting that estimating frequencies from relatively small amounts of data is enough (at least for our task) when large-scale estimations are expensive.

Inverse Scale Effect: The surprisal–novelty association decreases with model size, replicating prior results on the same dataset across other model families (GPT-2, Llama 3, Qwen 2.5) (Momen et al., 2026), and on datasets of reading time (Oh and Schuler, 2023b). We additionally show that the same negative scaling trend holds for correlations between surprisal and word frequency: as model size increases, surprisal becomes less aligned with frequency.

Pretraining Amount: The strongest association occurs early—after 128 pretraining updates for Pythia-70M (≈268M tokens seen)—and additional training weakens this association. A qualitatively similar non-monotonic effect has been reported for reading times: surprisal predicts reading time best at an intermediate pretraining amount (about 2B tokens), after which further pretraining reduces predictive power (Oh and Schuler, 2023a). The close similarity between the checkpoint trends for surprisal–novelty (Figure 2) and surprisal–frequency (Figure 4) highlights the extent to which word frequency can confound surprisal-based analyses of linguistic and psycholinguistic targets.

Surprisal: Surprisal achieves its strongest association with metaphor novelty when computed from the smallest model (70M) and at relatively early training (128 steps; ≈268M tokens). However, these are also the settings in which surprisal is most closely aligned with word frequency. We therefore caution against treating these *"best"* as-

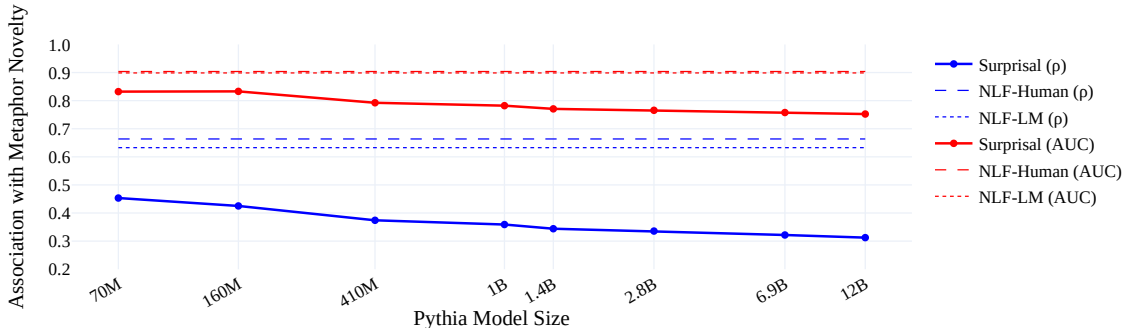

Figure 1: Effect of model size on associations between metaphor novelty scores and Surprisal (**solid**); Negative Log Word Frequency in general language use (NLF-Human) (**dash**); and Negative Log Word Frequency in Pythia's pretraining data (NLF-LM) (**dots**). Blue lines track Spearman correlation, and red lines track AUC to detect novel metaphors ($score \geq 0.5$).

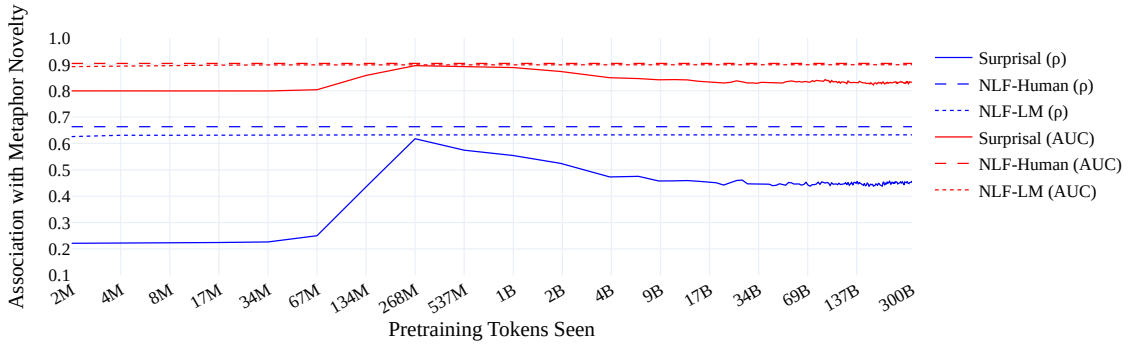

Figure 2: Effect of pretraining data/steps for Pythia-70M on associations between metaphor novelty scores and Surprisal (**solid**); Negative Log Word Frequency in general language use (NLF-Human) (**dash**); and Negative Log Word Frequency in Pythia's pretraining data (NLF-LM) (**dots**). Blue lines track Spearman correlation, and red lines track AUC.

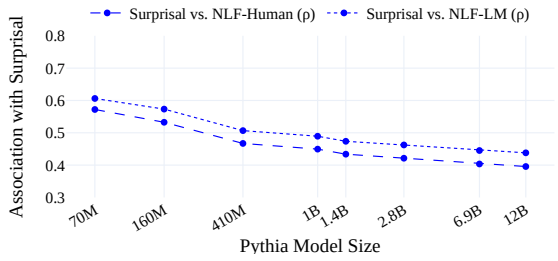

Figure 3: Effect of model sizes on correlation between Surprisal and NLF-Human (**dash**); and NLF-LM (**dots**).

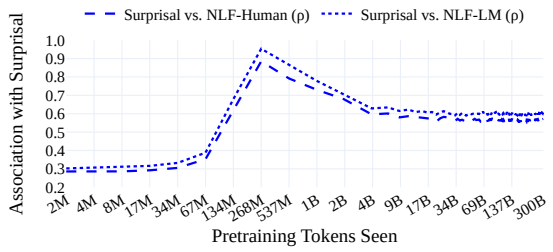

Figure 4: Effect of pretraining data/steps of Pythia-70M on surprisal–NLF correlations.

sociation values as direct evidence for surprisal-as-predictability accounts of metaphor novelty (and processing difficulty in general): in these settings, surprisal may primarily reflect lexical frequency, and possibly more than contextual predictability. At the same time, our findings should not be interpreted as implying that larger models or extensive pretraining produce intrinsically poor estimates of predictability. Rather, they motivate further efforts to develop more accurate novelty annotations that reflect true theories of the novelty dimension in creativity (and potentially processing difficulty more broadly) and to clarify how these constructs relate to surprisal-based accounts.

Conclusion: We analysed the associations between metaphor novelty scores, surprisal and word frequency across different model sizes and pretraining stages. The results mainly caution against interpreting strong surprisal–novelty results from small or early-trained models as straightforward evidence for surprisal-as-predictability accounts of metaphor novelty. Our resources will be publicly available.[4]

[4]Data and code: https://tinyurl.com/57av85ac

Limitations

Due to computational constraints, we do not compute surprisal for intermediate checkpoints of the larger Pythia models, and we restrict the checkpoint-level analysis to the smallest variant (Pythia-70M). Although this choice is consistent with our model-scale findings (smaller models yield stronger associations), evaluating intermediate checkpoints for larger models remains necessary to verify whether the observed training-dynamics trends hold across model sizes.

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. *Preprint*, arXiv:2304.01373.

Brian F. Bowdle and Dedre Gentner. 2005. The career of metaphor. *Psychological Review*, 112(1):193–216.

Eileen R. Cardillo, Christine E. Watson, Gwenda L. Schmidt, Alexander Kranjec, and Anjan Chatterjee. 2012. From novel to familiar: Tuning the brain for metaphors. *NeuroImage*, 59(4):3212–3221.

Anca Dinu, Andra-Maria Florescu, and Alina Resceanu. 2025. A comparative approach to assessing linguistic creativity of large language models and humans. *Procedia Computer Science*, 270:1292–1301.

Erik-Lân Do Dinh, Hannah Wieland, and Iryna Gurevych. 2018. Weeding out conventionalized metaphors: A corpus of novel metaphor annotations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1424, Brussels, Belgium. Association for Computational Linguistics.

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.

Adam Goodkind and Klinton Bicknell. 2018. Predictive power of word surprisal for reading times is a linear function of language model quality. In *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2018)*, pages 10–18, Salt Lake City, Utah. Association for Computational Linguistics.

Charles F. Hockett. 1960. The origin of speech. *Scientific American*, 203(3):88–97.

George Lakoff and Mark Johnson. 1980. *Metaphors We Live By*. University of Chicago Press, Chicago, IL.

Omar Momen, Emilie Sitter, Berenike Herrmann, and Sina Zarrieß. 2026. Surprisal and metaphor novelty: Moderate correlations and divergent scaling effects. *Preprint*, arXiv:2601.02015.

Byung-Doh Oh and William Schuler. 2023a. Transformer-based language model surprisal predicts human reading times best with about two billion training tokens. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1915–1921, Singapore. Association for Computational Linguistics.

Byung-Doh Oh and William Schuler. 2023b. Why does surprisal from larger transformer-based language models provide a poorer fit to human reading times? *Transactions of the Association for Computational Linguistics*, 11:336–350.

Byung-Doh Oh and William Schuler. 2024. Leading whitespaces of language models' subword vocabulary pose a confound for calculating word probabilities. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3464–3472, Miami, Florida, USA. Association for Computational Linguistics.

Byung-Doh Oh, Shisen Yue, and William Schuler. 2024. Frequency explains the inverse correlation of large language models' size, training data amount, and surprisal's fit to reading times. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2644–2663, St. Julian's, Malta. Association for Computational Linguistics.

Gill Philip. 2016. Conventional and novel metaphors in language. In Elena Semino and Zsófia Demjén, editors, *The Routledge Handbook of Metaphor and Language*, page 14. Routledge, London / New York.

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.

Sebastian Reimann and Tatjana Scheffler. 2024. When is a metaphor actually novel? annotating metaphor novelty in the context of automatic metaphor detection. In *Proceedings of the 18th Linguistic Annotation Workshop (LAW-XVIII)*, pages 87–97, St. Julians, Malta. Association for Computational Linguistics.

Cory Shain. 2024. Word frequency and predictability dissociate in naturalistic reading. *Open Mind*, 8:177–201.

Cory Shain, Clara Meister, Tiago Pimentel, Ryan Cotterell, and Roger Levy. 2024. Large-scale evidence for logarithmic effects of word predictability on reading time. *Proceedings of the National Academy of Sciences*, 121(10):e2307876121.

G.J. Steen, A.G. Dorst, J.B. Herrmann, A.A. Kaal, T. Krennmayr, and T. Pasma. 2010. *A method for linguistic metaphor identification. From MIP to MIPVU.* Number 14 in Converging Evidence in Language and Communication Research. John Benjamins.

Lindia Tjuatja, Graham Neubig, Tal Linzen, and Sophie Hao. 2025. What goes into a LM acceptability judgment? rethinking the impact of frequency and length. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2173–2186, Albuquerque, New Mexico. Association for Computational Linguistics.

George Yule. 2006. *The Study of Language*, 3 edition. Cambridge University Press, Cambridge, UK.

A Dataset Statistics

In Table 1, we demonstrate the statistics of the dataset used in our experiment.

B Numerical Results

Detailed numerical results of our study are listed in Tables 2, 3, 4 and 5.

Genre	# Metaphors	L_{sent} mean | std.	Novelty Score mean | std.	# Novel $>= 0.5$
Fiction	3170	26.0 | 16.5	-.005 | .271	94
News	4712	29.9 | 14.2	.000 | .257	132
Academic	5499	34.9 | 16.0	.003 | .239	102
Conversation	1774	17.5 | 15.9	-.000 | .236	25
All	15155	29.4 | 16.5	.000 | .251	353

Table 1: Distributions and statistics of the dataset under study. # Met. is the number of metaphor words, Score_{nov} is the BWS novelty scores, # Nov. is the number of novel metaphors (Score_{nov} >= 0.5). L_{sent} is the length of sentences in words.

Model	Pearson (r)	Spearman (ρ)	AUC
NLF-Human	**.656**	**.664**	**.904**
NLF-LM	**.599**	**.633**	**.899**
Pythia-70M	**.448**	**.453**	.832
Pythia-160M	.426	.425	**.833**
Pythia-410M	.382	.374	.792
Pythia-1B	.371	.359	.782
Pythia-1.4B	.357	.344	.771
Pythia-2.8B	.351	.336	.766
Pythia-6.9B	.338	.322	.758
Pythia-12B	.330	.312	.752

Table 2: Spearman Correlation and AUC estimates between **Metaphor Novelty Scores** and **Surprisal**; **Negative Log Word Frequency in general language use (NLF-Human)**; and **Negative Log Word Frequency in Pythia's pretraining data (NLF-LM)** across different model sizes. All reported estimates are significant at the 0.001 level.

Model	NLF-Human (ρ)	NLF-LM (ρ)
Pythia-70M	**.572**	**.606**
Pythia-160M	.533	.573
Pythia-410M	.467	.507
Pythia-1B	.449	.489
Pythia-1.4B	.434	.474
Pythia-2.8B	.421	.462
Pythia-6.9B	.404	.445
Pythia-12B	.396	.438

Table 3: Spearman Correlation estimates between **Surprisal** and **Negative Log Word Frequency in general language use (NLF-Human)**; and **Negative Log Word Frequency in Pythia's pretraining data (NLF-LM)** across different model sizes. All reported estimates are significant at the 0.001 level.

# Steps	# Pretraining Tokens	Surprisal		NLF-LM	
		ρ	AUC	ρ	AUC
1	2M	.221	.799	.626	.891
2	4M	.221	.799	.631	.894
4	8M	.221	.799	.631	.896
8	17M	.224	.800	.631	.897
16	34M	.226	.799	.631	.898
32	67M	.250	.804	.632	.898
64	134M	.435	.858	.632	**.899**
128	268M	**.618**	**.895**	.632	.899
256	537M	.574	.893	**.633**	.899
512	1B	.554	.888	.633	.899
1,000	2B	.524	.873	.633	.899
2,000	4B	.473	.849	.633	.899
3,000	6B	.475	.846	.633	.899
4,000	8B	.458	.842	.633	.899
5,000	10B	.458	.842	.633	.899
6,000	13B	.459	.841	.633	.899
7,000	15B	.457	.836	.633	.899
8,000	17B	.453	.833	.633	.899
9,000	19B	.451	.833	.633	.899
12,000	25B	.459	.838	.633	.899
17,000	36B	.448	.832	.633	.899
22,000	46B	.447	.831	.633	.899
27,000	57B	.446	.836	.633	.899
32,000	67B	.453	.835	.633	.899
37,000	78B	.446	.839	.633	.899
42,000	88B	.450	.843	.633	.899
47,000	99B	.445	.833	.633	.899
52,000	109B	.444	.829	.633	.899
57,000	120B	.443	.827	.633	.899
62,000	130B	.442	.831	.633	.899
67,000	141B	.446	.831	.633	.899
72,000	151B	.441	.830	.633	.899
77,000	161B	.447	.828	.633	.899
82,000	172B	.445	.824	.633	.899
87,000	182B	.440	.826	.633	.899
92,000	193B	.442	.831	.633	.899
97,000	203B	.443	.826	.633	.899
102,000	214B	.449	.828	.633	.899
107,000	224B	.448	.829	.633	.899
112,000	235B	.453	.833	.633	.899
117,000	245B	.447	.828	.633	.899
122,000	256B	.455	.835	.633	.899
127,000	266B	.451	.830	.633	.899
132,000	277B	.451	.835	.633	.899
137,000	287B	.448	.829	.633	.899
143,000	300B	.453	.832	.633	.899

Table 4: Spearman Correlation and AUC estimates between **Metaphor Novelty Scores** and **Surprisal**; and **Negative Log Word Frequency in Pythia's pretraining data (NLF-LM)** across pretraining steps of Pythia-70M, reporting the amount of pretraining data tokens seen at each step.

# Steps	# Pretraining Tokens	NLF-Human (ρ)	NLF-LM (ρ)
1	2M	.286	.302
2	4M	.286	.306
4	8M	.286	.309
8	17M	.292	.316
16	34M	.305	.332
32	67M	.352	.388
64	134M	.623	.679
128	268M	**.885**	**.953**
256	537M	.792	.866
512	1B	.730	.778
1,000	2B	.679	.705
2,000	4B	.597	.629
3,000	6B	.601	.634
4,000	8B	.580	.615
5,000	10B	.587	.623
6,000	13B	.583	.609
7,000	15B	.579	.614
8,000	17B	.578	.606
9,000	19B	.572	.609
12,000	25B	.583	.614
17,000	36B	.572	.605
22,000	46B	.566	.595
27,000	57B	.566	.599
32,000	67B	.578	.612
37,000	78B	.563	.602
42,000	88B	.570	.604
47,000	99B	.564	.600
52,000	109B	.562	.598
57,000	120B	.559	.595
62,000	130B	.557	.591
67,000	141B	.560	.595
72,000	151B	.559	.597
77,000	161B	.565	.601
82,000	172B	.563	.597
87,000	182B	.557	.595
92,000	193B	.557	.592
97,000	203B	.557	.594
102,000	214B	.565	.597
107,000	224B	.563	.597
112,000	235B	.571	.604
117,000	245B	.562	.595
122,000	256B	.572	.607
127,000	266B	.568	.602
132,000	277B	.567	.602
137,000	287B	.563	.600
143,000	300B	.572	.606

Table 5: Spearman Correlation estimates between **Surprisal** and **Negative Log Word Frequency in general corpora**; and **Negative Log Word Frequency in pretraining corpora** across pretraining steps of Pythia-70M, reporting the amount of pretraining data tokens seen at each step.