



# Beyond Surprisal: A Dual Metric Framework for Lexical Skill Acquisition in LLMs

Nazanin Shafiabadi  and Guillaume Wisniewski 

 Sorbonne Université, ISIR, 75 013 Paris, France

 Université Paris Cité, LLF, CNRS, 75 013 Paris, France

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

## Abstract

Many studies have explored when and how LLMs learn to use specific words, primarily by examining their learning curves. While these curves capture a model’s capacity to use words correctly in context, they often neglect the equally important skill of avoiding incorrect usage. In this paper, we introduce a new metric, *anti-surprisal*, which measures a model’s capacity to refrain from using words in inappropriate or unexpected contexts. By examining both correct usage and error avoidance, we offer a more comprehensive perspective on the learning dynamics of LLMs.

## 1 Introduction

There has been considerable interest in investigating when and how LLMs learn to use specific words, as well as the factors influencing this learning (Liu et al., 2021; Chang and Bergen, 2022; Xia et al., 2023; Evanson et al., 2023). Traditionally, lexical skill acquisition has been assessed using learning curves, which track metrics such as accuracy, perplexity, or surprisal to measure a model’s ability to predict words in context over time or across training iterations. These curves offer insights into the model’s learning trajectory, shedding light on how quickly and effectively it grasps word usage and revealing subtleties in how it generalizes or struggles with certain linguistic patterns (Chang et al., 2024).

However, we believe that this traditional approach has limitations, as it provides only a partial view of a model’s ability to “learn”<sup>1</sup> a word: while existing learning curves effectively capture whether a model can use a word in an appropriate context, they do not account for whether the model has also learned when *not* to use the word — a dimension of learning that is equally important yet

<sup>1</sup>Despite the risk of anthropomorphism, we use the term “learn” as is common in the literature. Though imprecise, we believe this term effectively conveys the intuition behind our analyses and the issues we address.

often overlooked. To address this gap, we propose a new metric, “anti-surprisal”, which reflects a model’s ability to refrain from using a word in unexpected or inappropriate contexts.

The measurement of anti-surprisal is of particular importance given that, during training, a language model is solely exposed to “positive” data — that is, correct sentences chosen for their quality (Longpre et al., 2024). However, there is no guarantee that this data is sufficient for the model to detect incorrect sentences: Gold (1967) even demonstrated (mathematically) that it is impossible to infer a formal language from positive examples alone, though within a theoretical framework that does not directly align with language model learning.

In this paper, we demonstrate that examining the dual evolution of a model’s ability to use words correctly in context while avoiding misapplication offers deeper insights than those gained by focusing solely on the probability of correct word usage, thereby capturing a fuller picture of the model’s linguistic capabilities.

## 2 Observing lexical skill acquisition in LLMs

**Conventional Metrics** Most studies on lexical skill acquisition in LLMs focus on the evolution of surprisal<sup>2</sup> for carefully selected target words in a *word bank* across different training stages (i.e., checkpoints). Surprisal is defined as  $-\log_2 p(w|c)$ , where  $w$  is the target word and  $c$  is the context, typically an extract from a corpus in which the target word is masked. A lower surprisal value indicates a higher probability of the word appearing

<sup>2</sup>Although surprisal is a common metric, it is essentially a monotonic transformation of probability and can be viewed as an unnecessary abstraction. Its value lies in its historical connection to psycholinguistics (Levy, 2008), which links surprisal to cognitive processing effort. This convention persists, though whether surprisal adds value beyond simple probability remains debatable.

in that context, suggesting the model has effectively “learned” the word’s usage.

Several works have shown that by identifying the words for which surprisal decreases over time, comparing the convergence values, and analyzing the speed and dynamics of this convergence, we can distinguish the words that are correctly “learned” from those that are not. This approach also allows the identification of factors, such as word frequency, that influence the learning process. A key takeaway from prior research is that studying surprisal curves for individual words offers deeper insights than relying on aggregate surprisal across a test or validation set, an approach typically used to monitor overall training progress.

**Learning not to use a word** We strongly believe that this traditional focus on studying surprisal curves has its limitations and provides only a partial view of a model’s ability to “learn” a word. While it effectively captures whether a model knows how to use a word in the right context, it does not address the complementary question of whether the model has also learned when not to use it — a dimension of word usage that is equally important but often overlooked.

Indeed, by studying surprisal, we assess solely whether a language model recognizes, for instance, the high probability that the masked word is “chocolate” in the sentence “I like [MASK] ice cream.” However, nothing guarantees that this probability is due to the model’s true “understanding” of the context rather than potential confounding factors, the most obvious being the frequency of the target word in the training data. This is why we believe it is also important to study a model’s ability to refrain from using a target word in an inappropriate context: continuing with the previous example, we need to ensure that the model’s estimated probability of “chocolate” being the hidden word in the sentence “The sun sets [MASK] the horizon.” decreases as training progresses.

To assess a model’s ability to avoid using a target word in inappropriate contexts, we introduce a complementary metric to surprisal: anti-surprisal, defined simply as  $-\log_2 p(w|c^-)$ , where  $c^-$  represents a “negative” or inappropriate context. While surprisal is expected to decrease with training, anti-surprisal should exhibit the opposite trend, increasing as the model improves.

As we will show in Section 4, analyzing the joint evolution of surprisal and anti-surprisal during

training provides a more accurate understanding of lexical skill acquisition in LLMs than focusing on surprisal alone.

### 3 Experimental Setup

**Model** For all our experiments, we employed intermediate checkpoints from the MultiBERTs model by Sellam et al. (2022), which offers 25 independent reproductions of BERT training, each using the original BERT hyperparameters. For our analysis, we selected the first of five MultiBERTs models with saved intermediate checkpoints (labeled “seed 0”). This model includes 28 checkpoints, covering every 20,000 steps up to 200,000, and every 100,000 steps up to 2 million.

**Word Bank Construction** Following the methodology outlined by Chang and Bergen (2022), we constructed a word bank of 9,080 unique words drawn from the test portion of the English WikiText-103 corpus.<sup>3</sup> This word bank contains the words we use to monitor the model’s lexical acquisition. Appendix A.1 details the preprocessing steps applied during the construction of this word bank.

**Appropriate Context Selection** To ensure longer context windows, we first split the testing subset of the WikiText corpus into sentence pairs. For each word in our word bank, we then selected the first 512 pairs containing that word, creating a set of examples to evaluate the model’s ability to predict target words in appropriate contexts. In each example, we randomly masked one occurrence of the target word and used a model checkpoint to estimate the probability that the masked word was the target.

**Inappropriate Context Generation** In our experiments, we also evaluate the model’s ability to avoid using target words in inappropriate contexts. To construct a dataset of inappropriate contexts, we used the same set of examples described earlier but masked a different word in each sentence. This approach ensures an equal number of appropriate and inappropriate contexts for each target word, allowing for a direct comparison of surprisal and anti-surprisal values.

**Measuring Lexical Skill Acquisition** For each example in our dataset, we estimate the probability  $p(w|c)$  using a language model checkpoint and

<sup>3</sup>The test portion of the English WikiText-103 corpus has not been used to train the MultiBERT models.

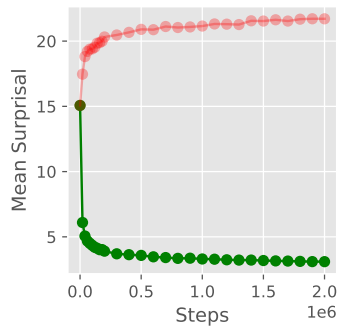


Figure 1: Corpus-level surprisal (green) and anti-surprisal (red) curves.

calculate the corresponding surprisal (or anti-surprisal). We average these values across all instances of  $w$  at each of the 29 distinct training steps, resulting in two learning curves per word: one for surprisal and one for anti-surprisal. These curves enable us to visualize how the model’s “understanding” of each word evolves over time.

To enable quantitative analysis, we model the learning curve for each target word by fitting a linear model that maps time steps to average surprisal (or anti-surprisal) values using `LinearRegression` from the `scikit-learn` library (Pedregosa et al., 2011). The linear model is chosen over more complex functions (e.g., exponential) to capture the “general trend” of the curve, even if the fit to the data is imperfect: the slope of the fitted model indeed provides a simple way to determine whether the modeled value is increasing or decreasing over time.<sup>4</sup>

## 4 Results

**Corpus-Level Analysis** We begin with a control experiment: Figure 1 shows the evolution of surprisal and anti-surprisal aggregated at the corpus level. As expected, surprisal decreases smoothly as learning progresses, while anti-surprisal increases. This indicates that, on average, the probability of using words in appropriate contexts rises over the course of training, while the probability of inappropriate word use declines. Thus, it can be inferred that even without exposure to negative (nonsensical) sentences, the language model can effectively identify and avoid them.

**Word-Level Analysis** On the other hand, analyzing learning curves for individual words, as advo-

<sup>4</sup>Code and data are available at <https://github.com/NazaninShafiabadi/antisurprisal>.

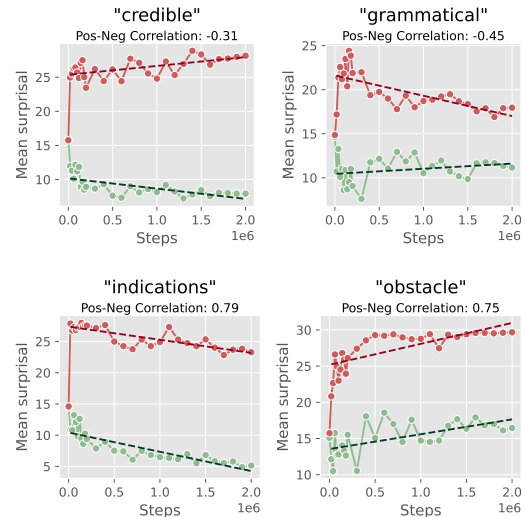


Figure 2: Word-level surprisal (green) and anti-surprisal (red) curves showing the 4 different trends, with dashed lines representing the fitted linear regression models.

cated by Xia et al. (2023), reveals that the model has highly variable learning dynamics from one word to the next. Figure 2 illustrates this by showing four carefully selected words, each with distinct learning trajectories representing all possible combinations of increasing or decreasing surprisal and anti-surprisal curves. The plots highlight the complexity of lexical learning, showing cases where the model’s confidence in word usage either improves, declines or fluctuates in unison or opposition. They also confirm Chang et al. (2024)’s observation that, at certain stages of learning, a language model may “forget” previously acquired word knowledge, as word-level performance is often non-monotonic contrary to what is observed at the corpus-level.

To quantify the prevalence of each trend across the word bank, we categorized all words based on the observed behavior of their surprisal and anti-surprisal curves. Specifically, linear models were fitted to these curves, and the slope of each model was used to determine whether the curve was increasing or decreasing. A cross-tab summarizing these trends is presented in Table 1, showing how often both curves move in the same direction, as well as cases where one curve increases while the other decreases.

The data reveals that the vast majority of words fall into the category of decreasing surprisal and increasing anti-surprisal, which is the expected behavior during training: this indicates that the model is becoming more confident in using these words correctly and avoiding incorrect contexts.

| anti-surprisal | surprisal  |              |
|----------------|------------|--------------|
|                | Increasing | Decreasing   |
| Increasing     | 2.7%       | <b>91.5%</b> |
| Decreasing     | 0.4%       | 5.4%         |

Table 1: Proportion of words in each surprisal and anti-surprisal trend category.





| anti-surprisal | surprisal  |  |
|----------------|--|--|
|                | Increasing   | Decreasing   |
| Increasing     | 99%   | 78%   |
| Decreasing     | 100%  | 100%  |

Table 2: Proportion of infrequent words in each surprisal and anti-surprisal trend category.

However, a notable minority (8.5% of the words in our word bank) exhibit different trends, underscoring the complexity of lexical acquisition beyond what corpus-level or surprisal-focused analyses might suggest. Most importantly, it demonstrates that word acquisition is not uniform: even in models with strong performance on downstream tasks, certain items remain poorly “captured” by the model. This behavior likely reflects lexical items that are inherently ambiguous or underrepresented in the training data, leading to slower or more erratic learning patterns.

Word frequency in the training corpus is likely a significant factor influencing these trends. To test this hypothesis, we examined the proportion of infrequent words (frequency  $\leq 10$  in the test set<sup>5</sup>) within each category, as well as their distribution across categories. The results show that nearly all (99–100%) words not correctly learned by the model (i.e., with increasing surprisal or decreasing anti-surprisal) are infrequent (Table 2). However, a significant majority of these infrequent words (89%) are still learned correctly (Figure 3). This suggests that while the model struggles primarily with infrequent words, frequency is not the sole determinant of lexical acquisition. The fact that most infrequent words are learned correctly highlights the likely influence of other factors, such as contextual diversity, on the model’s generalization.

<sup>5</sup>Despite our best efforts, we were unable to access Multi-BERT’s training data: the Book Corpus is no longer available. We hypothesized that a word’s frequency in the test set could serve as an approximation of its frequency in the training set.

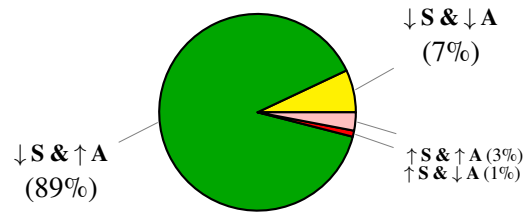


Figure 3: Distribution of infrequent words across surprisal (S) and anti-surprisal (A) trend categories. ↑ and ↓ indicate increasing and decreasing trends respectively.

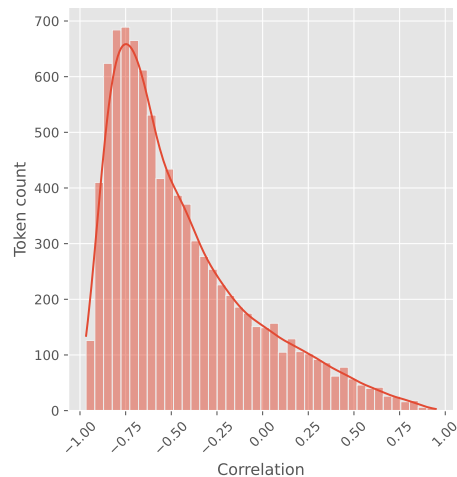


Figure 4: Distribution of token counts across different surprisal and anti-surprisal correlation values.

The remaining 11% may reflect challenges related to ambiguity, limited context, or difficulty in learning specific lexical items.

**Surprisal and Anti-Surprisal Correlation** In our final experiment, we aimed to determine the extent of the relationship between variations in surprisal and anti-surprisal curves. Specifically, we explored whether the model’s improved ability to predict the correct usage of a word also strengthens its ability to identify when not to use it, or if these two skills develop independently. To investigate this, we calculated the Pearson correlation between surprisal and anti-surprisal values for each word during training.

Figure 4 shows the distribution of correlation values for all words in our word bank. It appears that most words exhibit a “moderate” to “strong negative” correlation between surprisal and anti-surprisal, reinforcing the observation that in most cases, surprisal and anti-surprisal are strongly linked and, as already observed, their curves move in opposite directions. However, a minority



of words present weak or positive correlations, suggesting that for certain items, the ability to avoid inappropriate contexts may not align directly with improvements in predicting correct usage. These findings highlight the non-uniform nature of lexical acquisition, echoing earlier observations about the variability in word-level learning dynamics.

## 5 Discussion and Conclusion

We show that combining surprisal and anti-surprisal metrics offers a richer understanding of lexical skill acquisition in LLMs. This dual-metric approach reveals the non-uniform nature of word learning, where certain lexical items exhibit distinct learning trajectories, highlighting the need for comprehensive evaluation methods beyond simple probability metrics. By leveraging both positive and negative examples, we gain a more holistic view of the model’s linguistic capabilities, paving the way for more refined training and evaluation strategies in future research.

## 6 Limitations

While anti-surprisal offers valuable insights into a model’s ability to avoid inappropriate contexts, its interpretation is heavily dependent on the construction of negative examples. Relying on artificially generated inappropriate contexts may fail to capture the complexity of real-world linguistic scenarios, where contextual appropriateness often depends on subtler factors like pragmatics, domain knowledge, or the polysemous nature of words. Moreover, our study is limited to English, which restricts the generalizability of our findings across languages with different syntactic and morphological features. Future work should focus on developing more sophisticated methods for generating negative contexts and expanding the evaluation of anti-surprisal across a wider range of languages and domains.

## Acknowledgments

We are extremely grateful to BENOIT CRABBÉ and TIMOTHÉE BERNARD for their invaluable feedback on an earlier version of this work. Their insightful questions and particularly relevant suggestions were absolutely instrumental in initiating the reflections that ultimately led to this work.

This research was partially funded by the DIAGNOSTIC project supported by the Agence d’Innovation de Défense (grant n° 2022 65 007) and the

DEEPTYPO project supported by the Agence Nationale de la Recherche (ANR-23-CE38-0003-01).

## References

- Tyler A. Chang and Benjamin K. Bergen. 2022. [Word acquisition in neural language models](#). *Transactions of the Association for Computational Linguistics*, 10:1–16.
- Tyler A. Chang, Zhuowen Tu, and Benjamin K. Bergen. 2024. [Characterizing learning curves during language model pre-training: Learning, forgetting, and stability](#). *Preprint*, arXiv:2308.15419.
- Linnea Evanson, Yair Lakretz, and Jean Rémi King. 2023. [Language acquisition: do children and language models follow similar learning stages?](#) In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12205–12218, Toronto, Canada. Association for Computational Linguistics.
- E Mark Gold. 1967. [Language identification in the limit](#). *Information and Control*, 10(5):447–474.
- Matthew Honnibal and Ines Montani. 2018. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. *To appear*.
- Roger Levy. 2008. [Expectation-based syntactic comprehension](#). *Cognition*, 106(3):1126–1177.
- Zeyu Liu, Yizhong Wang, Jungo Kasai, Hannaneh Hajishirzi, and Noah A. Smith. 2021. [Probing across time: What does RoBERTa know and when?](#) In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 820–842, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Shayne Longpre, Gregory Yauney, Emily Reif, Katherine Lee, Adam Roberts, Barret Zoph, Denny Zhou, Jason Wei, Kevin Robinson, David Mimno, and Daphne Ippolito. 2024. [A pretrainer’s guide to training data: Measuring the effects of data age, domain coverage, quality, & toxicity](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3245–3276, Mexico City, Mexico. Association for Computational Linguistics.
- F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.
- Thibault Sellam, Steve Yadlowsky, Jason Wei, Naomi Saphra, Alexander D’Amour, Tal Linzen, Jasmijn Bastings, Iulia Turc, Jacob Eisenstein, Dipanjan Das,

Ian Tenney, and Ellie Pavlick. 2022. [The multiberts: Bert reproductions for robustness analysis](#). *Preprint*, arXiv:2106.16163.

Mengzhou Xia, Mikel Artetxe, Chunting Zhou, Xi Victoria Lin, Ramakanth Pasunuru, Danqi Chen, Luke Zettlemoyer, and Veselin Stoyanov. 2023. [Training trajectories of language models across scales](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13711–13738, Toronto, Canada. Association for Computational Linguistics.

## **A Appendix**

### **A.1 Word Bank Preprocessing**

The word bank was built by first removing tokens from the WikiText corpus containing non-ASCII or non-alphabetic characters. Next, we excluded tokens identified as proper nouns by SpaCy ([Hon-nibal and Montani, 2018](#)), as they provide limited insight into general lexical abilities. Additionally, we filtered out words treated as multiple tokens by the model, ensuring a final list of 9,080 words. This final step avoided the need for aggregating sub-token representations and simplified our analysis.