

BEYOND SURPRISAL: A DUAL METRIC FRAMEWORK FOR LEXICAL SKILL ACQUISITION IN LLMS

COLING 2025

Tuesday 21st January, 2025

Nazanin Shafiabadi¹ Guillaume Wisniewski²

¹Sorbonne Université, CNRS, ISIR

²Université Paris Cité, CNRS, LLF

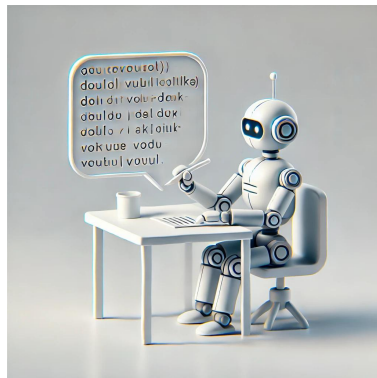
INTRODUCTION

Introduction and Motivation

- LLMs excel at diverse NLP tasks: *generating poetry, mimicking human-like conversations, ...*

But...

- How reliable are they with words?
- How can we better assess lexical skill acquisition?



Surprisal: A Glimpse into Word Understanding

Definition

Surprisal (S) = $-\log_2 p(w|c)$, where w is the target word, and c is the context.

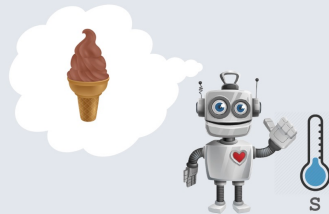
Explanation

Lower surprisal = higher probability of the word in the context.

Example:

“I like [MASK] ice cream.”

[MASK] = “chocolate”



Limitations of Surprisal

Problem

Surprisal alone doesn't guarantee true understanding; it might just reflect word frequency.

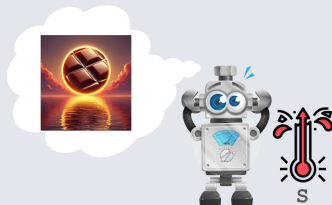
Need

To assess a model's ability to **avoid incorrect word usage**.

Example:

“The [MASK] sets in the horizon.”

[MASK] = “chocolate”



ANTI-SURPRISAL

Anti-surprisal: Measuring Error Avoidance

Definition

Anti-surprisal (AS) = $-\log_2 p(w|c^-)$, where c^- represents an inappropriate context.

Explanation

Higher anti-surprisal = lower probability of the word in the inappropriate context.

Example:

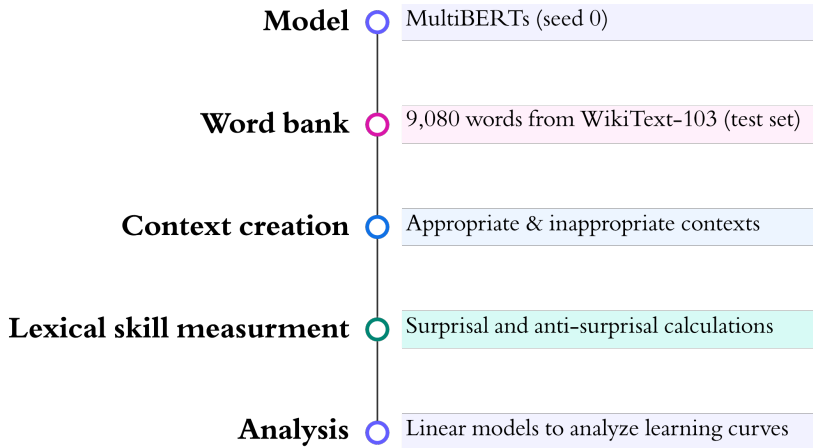
“The [MASK] sets in the horizon.”

[MASK] = “chocolate”

This word
doesn't
fit here.



EXPERIMENTAL SETUP



Context Generation

- **Original sentence pair:**

*"They signed a **contract**. It was legally binding."*

- **Target: "contract"**

- **Appropriate context:**

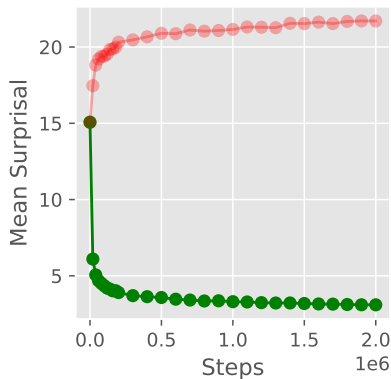
"They signed a [MASK]. It was legally binding."

- **Inappropriate context:**

"They signed a contract. It was [MASK] binding."

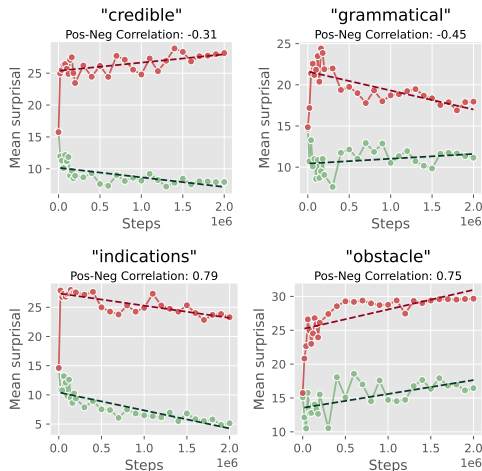
KEY FINDINGS

Corpus-Level Findings



Key observation: Surprisal decreases, anti-surprisal increases

Beyond the Average: Word-Level Learning Dynamics



Key observation: Non-uniform learning patterns

Quantifying Word-Level Trends

Table 1: The prevalence of surprisal and anti-surprisal trend categories across all the words in the word bank.

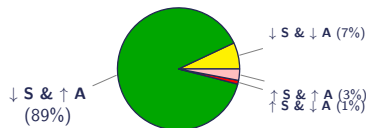
Anti-surprisal	Surprisal	
	Increasing	Decreasing
Increasing	2.7%	91.5%
Decreasing	0.4%	5.4%

- The majority of words (**91.5%**) show decreasing surprisal and increasing anti-surprisal.
- **8.5%** exhibit other trends. $\Rightarrow$ **lexical acquisition is non-uniform.**

The Role of Word Frequency

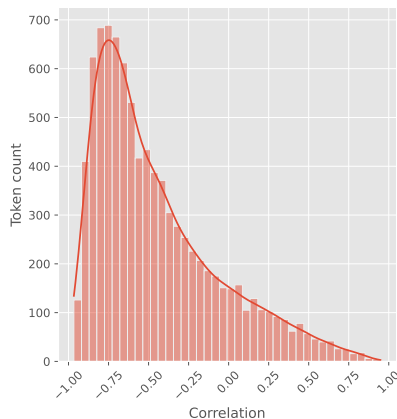
Anti-surprisal	Surprisal	
	Increasing	Decreasing
Increasing	99% 	78% 
Decreasing	100% 	100% 

Table 2: Proportion of infrequent words in each trend.



Key Observation: Most unlearned words are infrequent, but most infrequent words are learned.

Correlation Between Surprisal and Anti-Surprisal



Key observation: Most words show strong negative correlation, but some show weak or positive correlations.

CONCLUSION

Discussion

■ Key Findings:

- Combining surprisal and anti-surprisal provides a more complete picture of lexical acquisition.
- Lexical acquisition is a non-uniform process, with different words following unique learning trajectories.
- Word frequency plays a significant, but not exclusive, role.

■ Implication:

- Combining correct word use and avoidance of misuse reveals deeper linguistic competence, enabling better LLM training and evaluation.

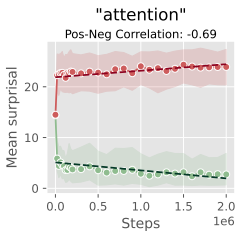
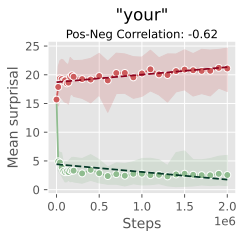
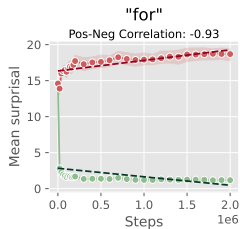
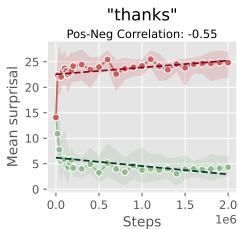
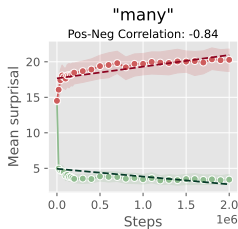
Limitations and Future Directions

■ Limitations:

- Reliance on artificially generated negative contexts.
- Study limited to English.
- Evaluation conducted on a single BERT-type model (MultiBERTs).

■ Future Directions:

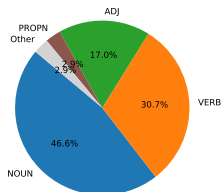
- Develop more sophisticated negative context generation methods.
- Explore the generalizability of these findings across languages.
- Validate findings on other models and model architectures.
- Investigate other factors influencing lexical acquisition.



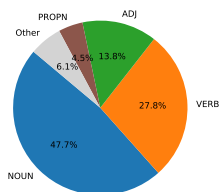
APPENDIX

Appendix A: The Role of POS Tags in Lexical Learning

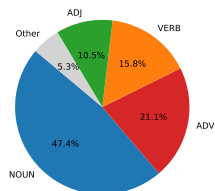
↗ Surprisal - ↗ Anti-surprisal



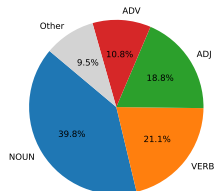
↘ Surprisal - ↗ Anti-surprisal



↗ Surprisal - ↘ Anti-surprisal



↘ Surprisal - ↘ Anti-surprisal



Key observation: Minimal Impact of POS on Lexical Learning