# Word Surprisal Correlates with Sentential Contradiction in LLMs

Ning Shi, Bradley Hauer, David Basil, John Zhang and Grzegorz Kondrak
{ning.shi,gkondrak}@ualberta.ca

## Introduction

**Contradiction**: A premise *P* contradicts a hypothesis *H* when *P* entails the negation of *H*. (MacCartney and Manning, 2009)

$$CON(P, H) \Leftrightarrow P \models \neg H$$

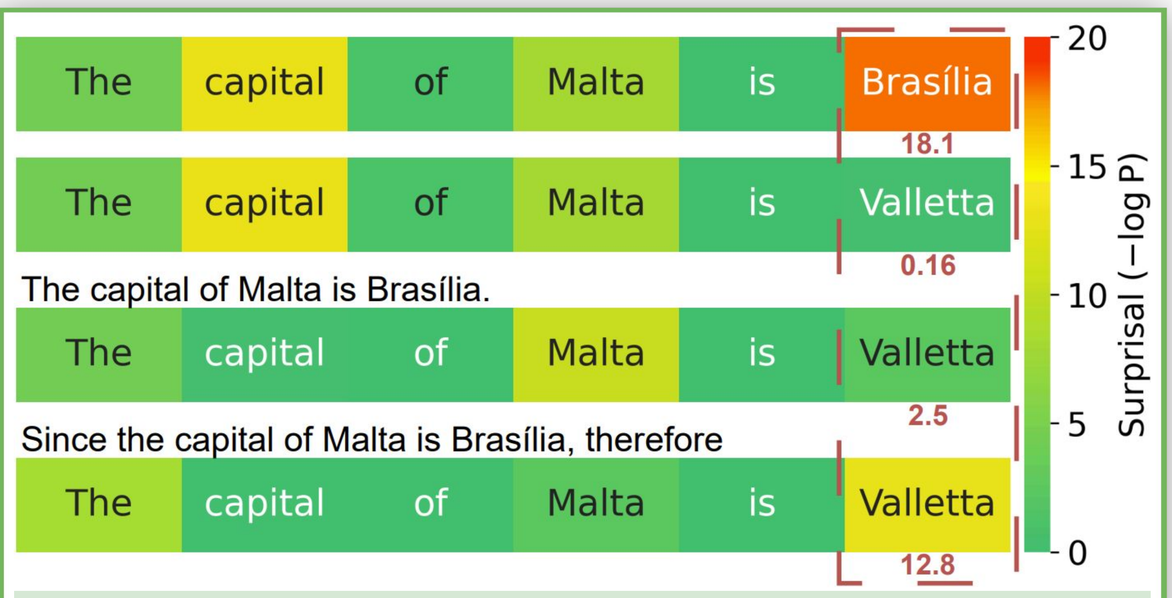
**Surprisal** measures the unpredictability of a word given a preceding context. (Hale, 2001)

$$Surprisal(w \mid C) = -\log P(w \mid C)$$

**Open question**: How does a language model respond to semantic inconsistency between two given sentences?

**We posit that**:

***When H is conditioned on P, the surprisal magnitude of content words in H is positively correlated with the presence of a contradiction, CON(P, H).***




Surprisal reflects context: **False** premises, especially under causal framing, can make **True** hypotheses **surprising**.

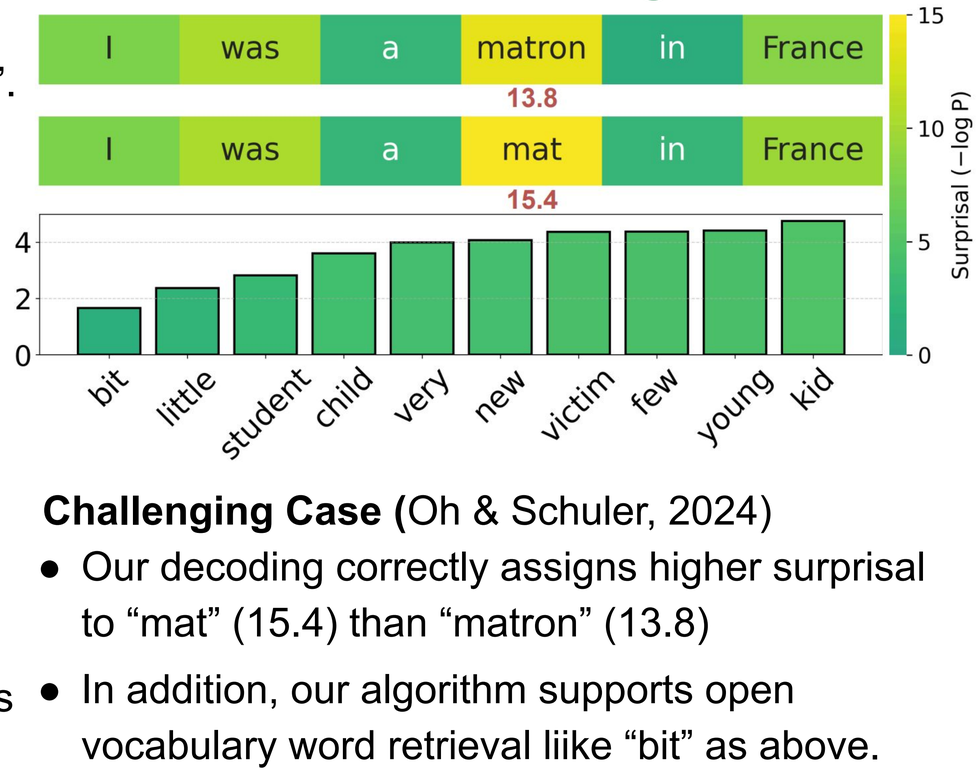
## Token-to-Word Decoding

**Token ≠ Word**: "*My name is ___*"
P("*John*") includes probability mass for "*Johnathan*".

**Constrained Expansion:**
Only expand token sequences that form valid word prefixes.

**Dynamic Normalization**: At each decoding step, renormalize probability mass over the valid token subset.

**InjectEOW**: Insert an explicit End-of-Word (EOW) option so completed words receive probability mass independently of longer continuations.




**Challenging Case (**Oh & Schuler, 2024)

- Our decoding correctly assigns higher surprisal to "mat" (15.4) than "matron" (13.8)
- In addition, our algorithm supports open vocabulary word retrieval liike "bit" as above.

**Algorithm 1** Token-to-Word Decoding

**Input:** tokenized context $\mathbf{S}^C$, beam width W, beam depth D, language model $\mathrm{F}(\cdot)$

$B_0 \leftarrow \{\langle 0, \mathbf{S}^C \rangle\}$
**for** $d \in \{1, \ldots, D\}$ :
  $B \leftarrow \varnothing$
  **for** $\langle p, \mathbf{S} \rangle \in B_{d-1}$ :
    **if** $\mathbf{S}.\mathrm{last}() = EOW$ :
      $B.\mathrm{add}(\langle p, \mathbf{S} \rangle)$; **continue**
    $\mathcal{P} \leftarrow \mathrm{F}(\mathbf{S})$
    **if** $d = 1$ :
      $\mathcal{P} \leftarrow \mathrm{Normalize}(P, \mathcal{S}_{\mathrm{bow}})$
    **else**:
      $\mathcal{P} \leftarrow \mathrm{InjectEOW}(\mathcal{P}, \mathcal{S}_{\mathrm{bow}})$
      $\mathcal{P} \leftarrow \mathrm{Normalize}(\mathcal{P}, \mathcal{S}_{\mathrm{mid}})$
    **for** $(p', s) \in \mathrm{Top}(\mathcal{P}, \mathrm{W})$ :
      $B.\mathrm{add}(\langle p \cdot p', \mathbf{S} \circ s \rangle)$
  $B_d \leftarrow B.\mathrm{top}(\mathrm{W})$
**return** $B_D.\mathrm{sort}()$

## Experiments

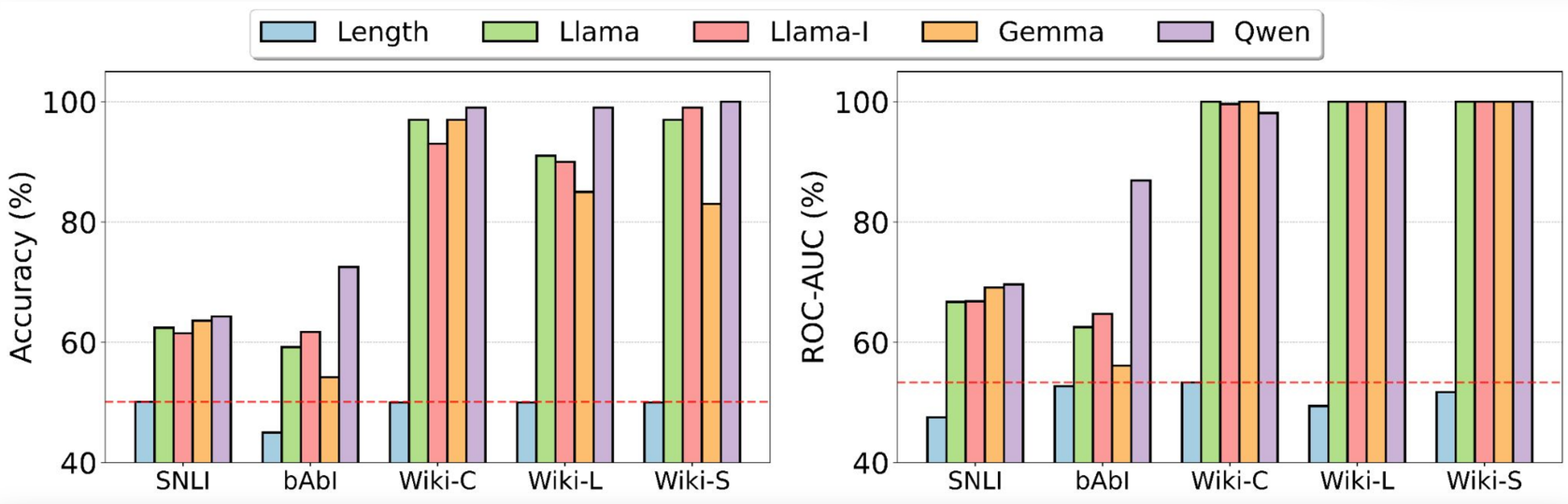
**TEMP Context:** "*Since {P}, therefore {H}*"
**Direct Surprisal**: $-\log P(w \mid C)$
**Mean Aggregation**: Average surprisal across hypothesis content words
**Classifier**: Tune a threshold on the validation set and predict contradiction when the mean surprisal exceeds.

***Word-level surprisal achieves statistically significant performance in both Accuracy and ROC-AUC across datasets (e.g., SNLI) and model families (e.g., Qwen).***

## Analysis

**Correlation over Causation**:
The model (e.g., Llama) favors distributional correlation over causal reasoning ("kitchen" over "bedroom").

**Negation Failure**:
The model also struggles with negation. It fails to raise surprise for words that should be logically inconsistent ("not in the").





## Conclusion

- Introduced **token-to-word decoding** for reliable word surprisal.
- Found a clear **positive correlation** between surprisal and contradiction.
- Demonstrated **robustness** across methods, models, and datasets.
- Revealed key model **limitations**:
  - dependency on word correlation
  - weak negation understanding

https://github.com/ShiningLab/CON2LM

UNIVERSITY OF ALBERTA