Word Surprisal Correlates with Sentential Contradiction in LLMs

Ning Shi, Bradley Hauer, David Basil, John Zhang, Grzegorz Kondrak

Alberta Machine Intelligence Institute (Amii)
Department of Computing Science
University of Alberta, Edmonton, Canada

Sentence-Level Contradiction

A **contradiction** (CON) is a relation between a **premise** (P) and a **hypothesis** (H) such that P entails the negation of H.

$$\text{CON}(P, H) \Leftrightarrow P \models \neg H$$

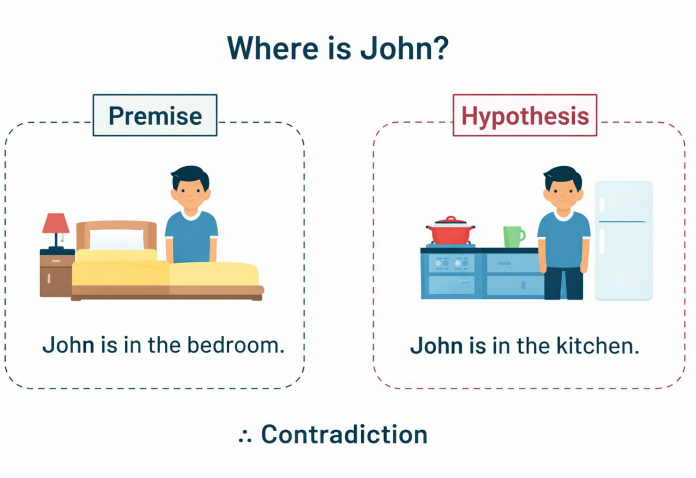
Example:

P = "John is in the bedroom."

H = "John is in the kitchen."

¬ H = "John is not in the kitchen."

P ⊨ ¬ H



An extended model of natural logic (MacCartney & Manning, IWCS 2009)

Word-Level Surprisal

Surprisal measures the unpredictability of a word in context. The surprisal of a **word** (w) given a preceding **context** (C) is defined as:

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

- **Low** probability → **High** surprisal
- **High** surprisal → Model did not expect this word

Context: "The capital of Canada is ___"

- "Ottawa" → Low surprisal
- "Edmonton" → High surprisal

A Probabilistic Earley Parser as a Psycholinguistic Model (Hale, NAACL 2001)

Research Question

In this work, we investigate the relationship between **lexical surprisal** and **sentential contradiction**.

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).

Our Contributions

We bridge lexical uncertainty and sentential semantics through both theory and empirical analysis.

- **First systematic study** of the correlation between lexical surprisal and sentential contradiction in LLMs.
- **Token-to-word decoding algorithm** for accurate word-level surprisal estimation in open-vocabulary settings.
- **Extensive empirical validation** across datasets and model families, demonstrating a **consistent positive** correlation.

Method

Context Construction (how P and H are presented)

- H-only: internal expectations
- CAT: premise conditioning
- TEMP: explicit causal framing

Surprisal Computation (how lexical uncertainty is measured)

- Direct: $-\log P(w \mid C)$
- Relative: normalized against most likely word

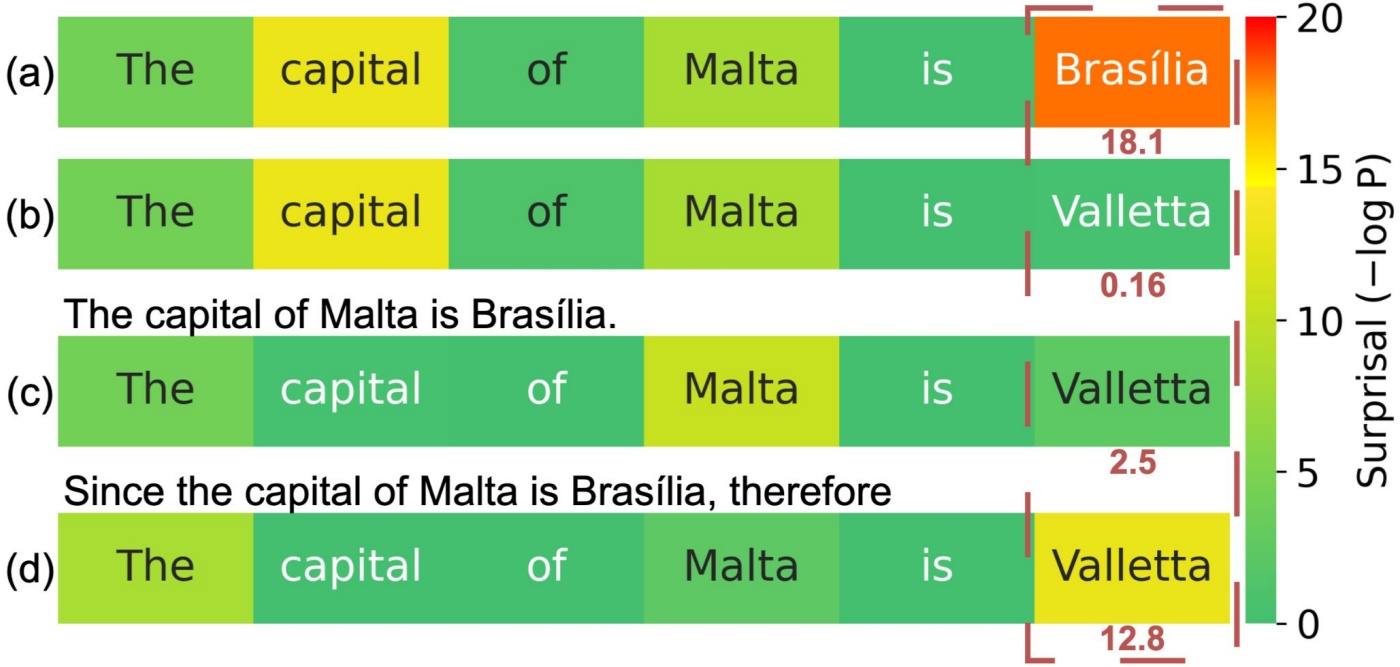
Surprisal Aggregation (How to aggregate to a sentence-level score)

- Last: the final word.
- Max: the highest surprisal
- Mean: the average surprisal

Aggregation plus thresholding **reduces** sentence-level contradiction detection into a word-level surprisal estimation problem.

Method – Context Construction

H-only in (a) and (b); **CAT** in (c); and **TEMP** in (d).



We used TEMP context, Direct surprisal, and Mean aggregation in the main experiments.

Token ≠ Word

Overestimation

Context: "My name is ___"

P("ĠJohn") includes probability mass for "**John**athan"

Token Prefix Ambiguity

What does "ĠK" mean?

Start of: **K**elsey, **K**ait, **K**ari...

Probability Mass Violation

Naively summing token paths can cause total word probability to **exceed 1**.

To name a few...

Leading Whitespaces of Language Models' Subword Vocabulary Pose a Confound for Calculating Word Probabilities (Oh & Schuler, EMNLP 2024)

Token-to-Word Decoding

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.

Algorithm 1 Token-to-Word Decoding

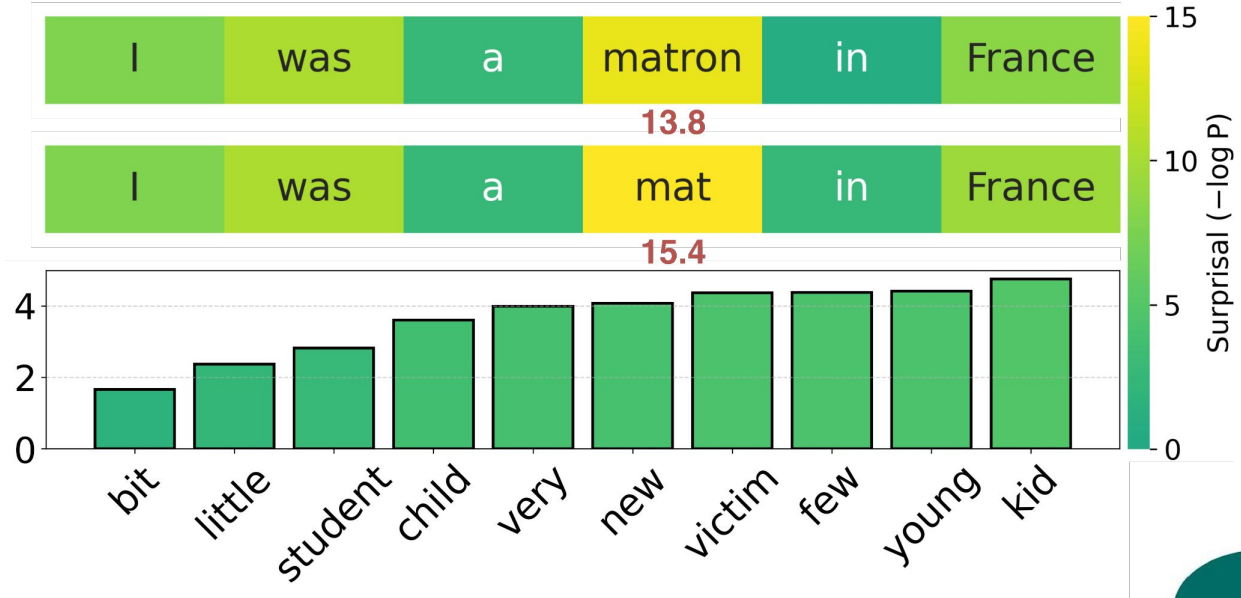
Input: tokenized context $\mathbf{S}^C$, beam width W, beam depth D, language model $\mathrm{F}(\cdot)$

$B_0 \leftarrow \{\langle 0, \mathbf{S}^C \rangle\}$
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 $d = 1$:
      $\mathcal{P} \leftarrow \mathrm{Normalize}(P, \mathcal{S}_{\text{bow}})$
    else:
      $\mathcal{P} \leftarrow \mathrm{InjectEOW}(\mathcal{P}, \mathcal{S}_{\text{bow}})$
      $\mathcal{P} \leftarrow \mathrm{Normalize}(\mathcal{P}, \mathcal{S}_{\text{mid}})$
    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 $B_D.\mathrm{sort}()$

Token-to-Word Decoding

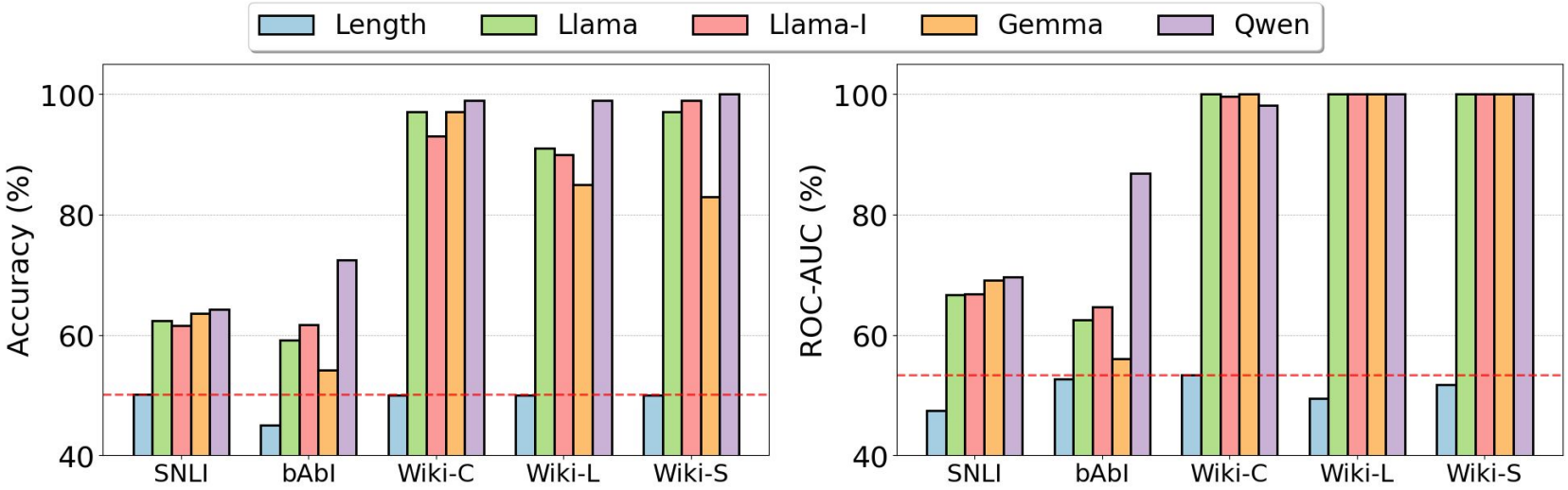
Challenging Case from Oh & Schuler (2024)

Our decoding assigns higher surprisal to “mat” (15.4) than “matron” (13.8), while supporting open-vocabulary word retrieval.



Results

Word-level surprisal achieves statistically **significant** performance (Accuracy and ROC-AUC), and consistently **outperforms** the Length baseline across datasets and model families.



We followed the evaluation and baseline in Can You Learn Semantics Through Next-Word Prediction? The Case of Entailment (Merrill et al., Findings 2024)

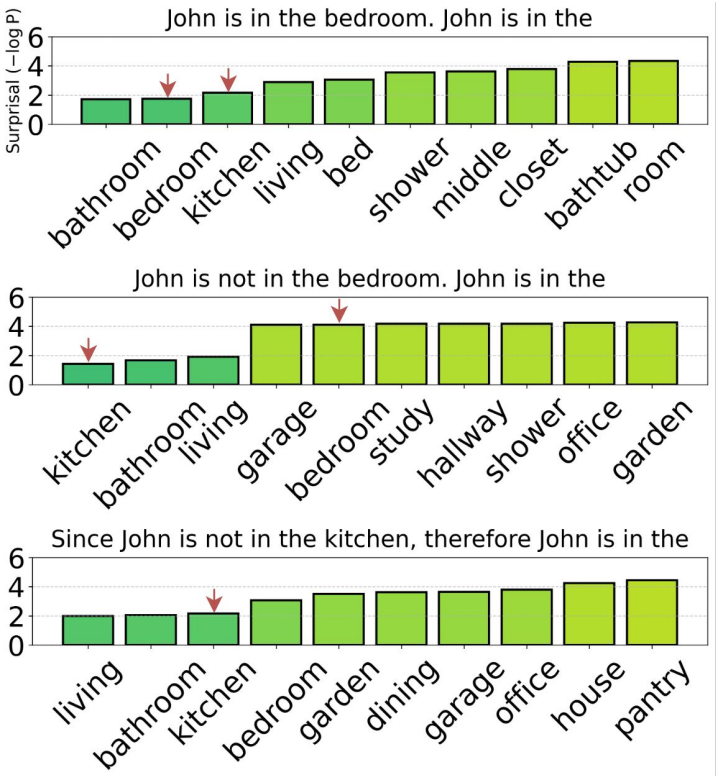
Analysis

Correlation over Causation

The model (Llama) favors distributional correlation over causal reasoning ("kitchen" over "bedroom").

Negation Failure

The model also struggles with negation, failing to raise surprise for words that should be logically inconsistent ("not in the").



Conclusion

We have investigated whether word-level surprisal correlates with sentence-level contradiction in LLMs.

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

github.com/ShiningLab/CON2LM