Do Language Models learn the specificity of parasitic gaps?

Adèle Hénot-Mortier (MIT)

August 31, 2023

Architectures and Mechanisms of Language Processing 29

Introduction

What is the syntactic characterization of parasitic gaps?

- *Wh*-questions introduce a dependency between the *wh*-word and the position the questioned element would have occupied in the corresponding answer. This position is called a gap (__).

(1) a. What did you eat __?
    b. I ate **[an apple]**.

- English allows for parasitic gaps (**PG**, $\_\__{pg}$), i.e. dependencies licensed by another gap (__) in the sentence (Engdahl, 1983).
- Interestingly, PGs do not seem to behave like regular gaps:
  1. not all languages allowing regular gaps allow for parasitic gaps;
  2. PGs are not reconstruction sites as evidenced by anaphor-binding diagnostics;
  3. **PGs typically occur in islands for extraction, such as adjunct clauses.**
- We want to focus on Property 3, sketched below.

(2) What did you discard __ [$_{\text{adjunct}}$ after using $\_\__{pg}$]?

What is the syntactic characterization of parasitic gaps?

- *Wh*-questions introduce a dependency between the *wh*-word and the position the questioned element would have occupied in the corresponding answer. This position is called a gap (__).

(1) a. What did you eat __?
    b. I ate **[an apple]**.

- English allows for parasitic gaps (**PG**, $\_\__{pg}$), i.e. dependencies licensed by another gap (__) in the sentence (Engdahl, 1983).
- Interestingly, PGs do not seem to behave like regular gaps:
  1. not all languages allowing regular gaps allow for parasitic gaps;
  2. PGs are not reconstruction sites as evidenced by anaphor-binding diagnostics;
  3. **PGs typically occur in islands for extraction, such as adjunct clauses.**
- We want to focus on Property 3, sketched below.

(2) What did you discard __ $[_{\text{adjunct}}$ after using $\_\__{pg}]$?

What is the syntactic characterization of parasitic gaps?

- *Wh*-questions introduce a dependency between the *wh*-word and the position the questioned element would have occupied in the corresponding answer. This position is called a gap (__).

(1) a. What did you eat __?
    b. I ate **[an apple]**.

- English allows for parasitic gaps (**PG**, $\_\__{pg}$), i.e. dependencies licensed by another gap (__) in the sentence (Engdahl, 1983).
- Interestingly, PGs do not seem to behave like regular gaps:
  1. not all languages allowing regular gaps allow for parasitic gaps;
  2. PGs are not reconstruction sites as evidenced by anaphor-binding diagnostics;
  3. **PGs typically occur in islands for extraction, such as adjunct clauses.**
- We want to focus on Property 3, sketched below.

(2) What did you discard __ [$_{\text{adjunct}}$ after using $\_\__{pg}$]?

What is the syntactic characterization of parasitic gaps?

- *Wh*-questions introduce a dependency between the *wh*-word and the position the questioned element would have occupied in the corresponding answer. This position is called a gap (__).

(1) a. What did you eat __?
    b. I ate **[an apple]**.

- English allows for parasitic gaps (**PG**, $\_\__{pg}$), i.e. dependencies licensed by another gap (__) in the sentence (Engdahl, 1983).
- Interestingly, PGs do not seem to behave like regular gaps:
  1. not all languages allowing regular gaps allow for parasitic gaps;
  2. PGs are not reconstruction sites as evidenced by anaphor-binding diagnostics;
  3. **PGs typically occur in islands for extraction, such as adjunct clauses.**
- We want to focus on Property 3, sketched below.

(2) What did you discard __ $[_{\text{adjunct}}$ after using $\_\__{pg}]$?

What is the syntactic characterization of parasitic gaps?

- *Wh*-questions introduce a dependency between the *wh*-word and the position the questioned element would have occupied in the corresponding answer. This position is called a gap (__).

(1) a. What did you eat __?
    b. I ate **[an apple]**.

- English allows for parasitic gaps (**PG**, $\_\_{}_{pg}$), i.e. dependencies licensed by another gap (__) in the sentence (Engdahl, 1983).
- Interestingly, PGs do not seem to behave like regular gaps:
  1. not all languages allowing regular gaps allow for parasitic gaps;
  2. PGs are not reconstruction sites as evidenced by anaphor-binding diagnostics;
  3. **PGs typically occur in islands for extraction, such as adjunct clauses.**
- We want to focus on Property 3, sketched below.

(2) What did you discard __ $[_{\text{adjunct}}$ after using $\_\_{}_{pg}]$?

What is the syntactic characterization of parasitic gaps?

- *Wh*-questions introduce a dependency between the *wh*-word and the position the questioned element would have occupied in the corresponding answer. This position is called a gap (__).

(1) a. What did you eat __?
b. I ate **[an apple]**.

- English allows for parasitic gaps (**PG**, $\_\__{pg}$), i.e. dependencies licensed by another gap (__) in the sentence (Engdahl, 1983).
- Interestingly, PGs do not seem to behave like regular gaps:
  1. not all languages allowing regular gaps allow for parasitic gaps;
  2. PGs are not reconstruction sites as evidenced by anaphor-binding diagnostics;
  3. **PGs typically occur in islands for extraction, such as adjunct clauses.**
- We want to focus on Property 3, sketched below.

(2) What did you discard __ $[_{\text{adjunct}}$ after using $\_\__{pg}]$?

What is the syntactic characterization of parasitic gaps?

- *Wh*-questions introduce a dependency between the *wh*-word and the position the questioned element would have occupied in the corresponding answer. This position is called a gap (__).

(1) a. What did you eat __?
    b. I ate **[an apple]**.

- English allows for parasitic gaps (**PG**, $\_\__{pg}$), i.e. dependencies licensed by another gap (__) in the sentence (Engdahl, 1983).
- Interestingly, PGs do not seem to behave like regular gaps:
  1. not all languages allowing regular gaps allow for parasitic gaps;
  2. PGs are not reconstruction sites as evidenced by anaphor-binding diagnostics;
  3. **PGs typically occur in islands for extraction, such as adjunct clauses.**
- We want to focus on Property 3, sketched below.

(2) What did you discard __ $[_{\text{adjunct}}$ after using $\_\__{pg}]$?

Zooming on the "island" property

- Strongly transitive verbs require an object (or gap).

(3) a. Mary { used / discarded } *(the book).
    b. What did you { use / discard } __?

- Islands are constituent from within which a filler-gap dependency cannot be established. Adjuncts are generally strong islands.
- Crucially, **gaps are disallowed within adjuncts, but PGs are OK!**
- This is made clear in (4) due to *using* requiring a gap (strongly transitive), and *discard* being saturated by an overt object in (4a) but not in (4b).

(4) a. * What did you discard it [before using __] ?
    b. What did you discard __ [before using $\_\__{pg}$] ?

Question

- Large Language Models (**LLM**) are exposed to sentences involving regular and parasitic gaps.
- But they are never explicitly taught about the syntactic differences between them.
- **Do LLMs "understand" the specificity of parasitic gaps?**

Modeling of the problem

Models tested

- We tested 4 LLMs built on the **Transformer architecture** (Vaswani et al., 2017): GPT-2 (Radford et al., 2019), XLNet (Yang et al., 2019), BERT (Devlin et al., 2018), and RoBERTa (Liu et al., 2019).[1]
  - BERT and RoBERTa are "**bidirectional**" Transformers, which means that the probability of an individual token can depend on both its left- and right-context.
  - GPT-2 on the other hand, is purely **left-to-right**.
  - XLNet finally, is "structurally" left-to-right, but trained on an objective which allows to incorporate bidirectional information.
- These architectural differences can significantly affect the models' behavior when it comes to evaluating and processing sentences.

[1] Models from Hugging Face, all in their LARGE version.

Models tested

- We tested 4 LLMs built on the **Transformer architecture** (Vaswani et al., 2017): GPT-2 (Radford et al., 2019), XLNet (Yang et al., 2019), BERT (Devlin et al., 2018), and RoBERTa (Liu et al., 2019).[1]
  - BERT and RoBERTa are "**bidirectional**" Transformers, which means that the probability of an individual token can depend on both its left- and right-context.
  - GPT-2 on the other hand, is purely **left-to-right**.
  - XLNet finally, is "structurally" left-to-right, but trained on an objective which allows to incorporate bidirectional information.
- These architectural differences can significantly affect the models' behavior when it comes to evaluating and processing sentences.

[1] Models from Hugging Face, all in their LARGE version.

Models tested

- We tested 4 LLMs built on the **Transformer architecture** (Vaswani et al., 2017): GPT-2 (Radford et al., 2019), XLNet (Yang et al., 2019), BERT (Devlin et al., 2018), and RoBERTa (Liu et al., 2019).[1]
  - BERT and RoBERTa are "**bidirectional**" Transformers, which means that the probability of an individual token can depend on both its left- and right-context.
  - GPT-2 on the other hand, is purely **left-to-right**.
  - XLNet finally, is "structurally" left-to-right, but trained on an objective which allows to incorporate bidirectional information.
- These architectural differences can significantly affect the models' behavior when it comes to evaluating and processing sentences.

[1] Models from Hugging Face, all in their LARGE version.

Models tested

- We tested 4 LLMs built on the **Transformer architecture** (Vaswani et al., 2017): GPT-2 (Radford et al., 2019), XLNet (Yang et al., 2019), BERT (Devlin et al., 2018), and RoBERTa (Liu et al., 2019).[1]
  - BERT and RoBERTa are "**bidirectional**" Transformers, which means that the probability of an individual token can depend on both its left- and right-context.
  - GPT-2 on the other hand, is purely **left-to-right**.
  - XLNet finally, is "structurally" left-to-right, but trained on an objective which allows to incorporate bidirectional information.
- These architectural differences can significantly affect the models' behavior when it comes to evaluating and processing sentences.

[1] Models from Hugging Face, all in their LARGE version.

Models tested

- We tested 4 LLMs built on the **Transformer architecture** (Vaswani et al., 2017): GPT-2 (Radford et al., 2019), XLNet (Yang et al., 2019), BERT (Devlin et al., 2018), and RoBERTa (Liu et al., 2019).[1]
    - BERT and RoBERTa are "**bidirectional**" Transformers, which means that the probability of an individual token can depend on both its left- and right-context.
    - GPT-2 on the other hand, is purely **left-to-right**.
    - XLNet finally, is "structurally" left-to-right, but trained on an objective which allows to incorporate bidirectional information.
- These architectural differences can significantly affect the models' behavior when it comes to evaluating and processing sentences.

[1] Models from Hugging Face, all in their Large version.

Modeling grammaticality judgments

- The LLMs were evaluated like human subjects would be, using sentences which varied minimally along critical parameters.
- We used **surprisal**, which has been shown to correlate with language processing effort (Hale, 2001; Levy, 2008), as a proxy for grammaticality.

$$\begin{aligned}\textsc{Grammaticality}(w_t) &\simeq -\textsc{Surprisal}(w_t)\\ &= \log P(w_t|w_1 \ldots w_{t-1})^2\\ \textsc{Grammaticality}(w_1 \ldots w_t) &\simeq -\sum_{i=1}^{t} \textsc{Surprisal}(w_i)\end{aligned}$$

- Measures of surprisal were computed using the Python `minicons` library (Misra, 2022).

[2] In the case of BERT-like bidirectional models, this formula is adapted to a masked language modeling objective: the probability of a word is computed given its left *and* right context.

Modeling grammaticality judgments

- The LLMs were evaluated like human subjects would be, using sentences which varied minimally along critical parameters.
- We used **surprisal**, which has been shown to correlate with language processing effort (Hale, 2001; Levy, 2008), as a proxy for grammaticality.

$$\begin{aligned} \textsc{Grammaticality}(w_t) &\simeq -\textsc{Surprisal}(w_t) \\ &= \log P(w_t | w_1 \dots w_{t-1})^2 \\ \textsc{Grammaticality}(w_1 \dots w_t) &\simeq -\sum_{i=1}^{t} \textsc{Surprisal}(w_i) \end{aligned}$$

- Measures of surprisal were computed using the Python `minicons` library (Misra, 2022).

[2] In the case of BERT-like bidirectional models, this formula is adapted to a masked language modeling objective: the probability of a word is computed given its left *and* right context.

Modeling grammaticality judgments

- The LLMs were evaluated like human subjects would be, using sentences which varied minimally along critical parameters.
- We used **surprisal**, which has been shown to correlate with language processing effort (Hale, 2001; Levy, 2008), as a proxy for grammaticality.

$$\begin{aligned} \textsc{Grammaticality}(w_t) &\simeq -\textsc{Surprisal}(w_t) \\ &= \log P(w_t | w_1 \dots w_{t-1})^2 \\ \textsc{Grammaticality}(w_1 \dots w_t) &\simeq -\sum_{i=1}^{t} \textsc{Surprisal}(w_i) \end{aligned}$$

- Measures of surprisal were computed using the Python `minicons` library (Misra, 2022).

[2] In the case of BERT-like bidirectional models, this formula is adapted to a masked language modeling objective: the probability of a word is computed given its left *and* right context.

Previous work based on a similar methodology

- This general methodology is not new and has been previously used to investigate related phenomena such as various **island** (E. G. Wilcox et al., 2023) and **garden-path effects** (Futrell et al., 2019), **more standard filler-gap dependencies** (E. Wilcox et al., 2018; E. G. Wilcox et al., 2023), **relativization** (Kobzeva et al., 2022).
- More broadly, it is based on seminal work on Language Model explainability pertaining to agreement effects (Linzen et al., 2016; Gulordava et al., 2018, a.o.).
- To our knowledge however, PGs have never been systematically investigated through that lens in the past, and, additionally, our investigation focuses on recent LLMs instead of RNNs.

Previous work based on a similar methodology

- This general methodology is not new and has been previously used to investigate related phenomena such as various **island** (E. G. Wilcox et al., 2023) and **garden-path effects** (Futrell et al., 2019), **more standard filler-gap dependencies** (E. Wilcox et al., 2018; E. G. Wilcox et al., 2023), **relativization** (Kobzeva et al., 2022).
- More broadly, it is based on seminal work on Language Model explainability pertaining to agreement effects (Linzen et al., 2016; Gulordava et al., 2018, a.o.).
- To our knowledge however, PGs have never been systematically investigated through that lens in the past, and, additionally, our investigation focuses on recent LLMs instead of RNNs.

Previous work based on a similar methodology

- This general methodology is not new and has been previously used to investigate related phenomena such as various **island** (E. G. Wilcox et al., 2023) and **garden-path effects** (Futrell et al., 2019), **more standard filler-gap dependencies** (E. Wilcox et al., 2018; E. G. Wilcox et al., 2023), **relativization** (Kobzeva et al., 2022).
- More broadly, it is based on seminal work on Language Model explainability pertaining to agreement effects (Linzen et al., 2016; Gulordava et al., 2018, a.o.).
- To our knowledge however, PGs have never been systematically investigated through that lens in the past, and, additionally, our investigation focuses on recent LLMs instead of RNNs.

Task 0: testing island-sensitivity

Motivation

- The sentences we are interested in here follow the template below:

(5) $$\text{Wh did Subj } V_1 \left\{ \begin{array}{c} \text{pro} \\ \text{\_\_} \end{array} \right\} \left\{ \begin{array}{c} \text{before} \\ \text{after} \\ \text{without} \end{array} \right\} V_2\text{-ing} \left\{ \begin{array}{c} \text{pro} \\ \text{\_\_}(pg) \end{array} \right\} ?$$

- In particular, a sentence such as (6a), containing an object pronoun in the matrix clause but a gap in the embedded clause, is bad due to:
  1. **the gap being located in an adjunct island;**
  2. **the gap not being parasitic on anything.**
- A sentence such as (6b) may be semantically weird out of the blue, but is syntactically OK.

(6) a. * What did you discard **it** after using __?

b. What did you discard __ after using **it**?

- We first test if LLMs are attuned to this contrast in the gap's position. This will allow to ensure that (6a) is a good ungrammatical baseline.

Motivation

- The sentences we are interested in here follow the template below:

(5) Wh did Subj $V_1 \left\{ \begin{array}{c} \text{pro} \\ \text{—} \end{array} \right\} \left\{ \begin{array}{c} \text{before} \\ \text{after} \\ \text{without} \end{array} \right\} V_2\text{-ing} \left\{ \begin{array}{c} \text{pro} \\ \text{—}(pg) \end{array} \right\}$?

- In particular, a sentence such as (6a), containing an object pronoun in the matrix clause but a gap in the embedded clause, is bad due to:
  1. **the gap being located in an adjunct island;**
  2. **the gap not being parasitic on anything.**
- A sentence such as (6b) may be semantically weird out of the blue, but is syntactically OK.

(6) a. * What did you discard **it** after using __?

b. What did you discard __ after using **it**?

- We first test if LLMs are attuned to this contrast in the gap's position. This will allow to ensure that (6a) is a good ungrammatical baseline.

Motivation

- The sentences we are interested in here follow the template below:

$$\text{(5)}\quad \text{Wh did Subj } V_1 \left\{ \begin{array}{c} \text{pro} \\ \text{—} \end{array} \right\} \left\{ \begin{array}{c} \text{before} \\ \text{after} \\ \text{without} \end{array} \right\} V_2\text{-ing} \left\{ \begin{array}{c} \text{pro} \\ \text{—}(pg) \end{array} \right\} ?$$

- In particular, a sentence such as (6a), containing an object pronoun in the matrix clause but a gap in the embedded clause, is bad due to:
  1. **the gap being located in an adjunct island;**
  2. **the gap not being parasitic on anything.**
- A sentence such as (6b) may be semantically weird out of the blue, but is syntactically OK.

(6) a. * What did you discard **it** after using __?

b. What did you discard __ after using **it**?

- We first test if LLMs are attuned to this contrast in the gap's position. This will allow to ensure that (6a) is a good ungrammatical baseline.

Motivation

- The sentences we are interested in here follow the template below:

(5) $$\text{Wh did Subj } V_1 \left\{ \begin{array}{c} \text{pro} \\ \text{—} \end{array} \right\} \left\{ \begin{array}{c} \text{before} \\ \text{after} \\ \text{without} \end{array} \right\} V_2\text{-ing} \left\{ \begin{array}{c} \text{pro} \\ \text{—}(pg) \end{array} \right\} ?$$

- In particular, a sentence such as (6a), containing an object pronoun in the matrix clause but a gap in the embedded clause, is bad due to:
  1. **the gap being located in an adjunct island;**
  2. **the gap not being parasitic on anything.**
- A sentence such as (6b) may be semantically weird out of the blue, but is syntactically OK.

(6) a. * What did you discard **it** after using __?

b. What did you discard __ after using **it**?

- We first test if LLMs are attuned to this contrast in the gap's position. This will allow to ensure that (6a) is a good ungrammatical baseline.

Motivation

- The sentences we are interested in here follow the template below:

(5) $$\text{Wh did Subj } V_1 \left\{ \begin{array}{c} \text{pro} \\ \text{—} \end{array} \right\} \left\{ \begin{array}{c} \text{before} \\ \text{after} \\ \text{without} \end{array} \right\} V_2\text{-ing} \left\{ \begin{array}{c} \text{pro} \\ \text{—}(pg) \end{array} \right\} ?$$

- In particular, a sentence such as (6a), containing an object pronoun in the matrix clause but a gap in the embedded clause, is bad due to:
  1. **the gap being located in an adjunct island;**
  2. **the gap not being parasitic on anything.**
- A sentence such as (6b) may be semantically weird out of the blue, but is syntactically OK.

(6) a. * What did you discard **it** after using __?

b. What did you discard __ after using **it**?

- We first test if LLMs are attuned to this contrast in the gap's position. This will allow to ensure that (6a) is a good ungrammatical baseline.

Motivation

- The sentences we are interested in here follow the template below:

(5) $$\text{Wh did Subj } \text{V}_1 \left\{ \begin{array}{c} \text{pro} \\ \text{—} \end{array} \right\} \left\{ \begin{array}{c} \text{before} \\ \text{after} \\ \text{without} \end{array} \right\} \text{V}_2\text{-ing} \left\{ \begin{array}{c} \text{pro} \\ \text{—}(pg) \end{array} \right\} ?$$

- In particular, a sentence such as (6a), containing an object pronoun in the matrix clause but a gap in the embedded clause, is bad due to:
  1. **the gap being located in an adjunct island;**
  2. **the gap not being parasitic on anything.**
- A sentence such as (6b) may be semantically weird out of the blue, but is syntactically OK.

(6) a. * What did you discard **it** after using __?

b. What did you discard __ after using **it**?

- We first test if LLMs are attuned to this contrast in the gap's position. This will allow to ensure that (6a) is a good ungrammatical baseline.

Motivation

- The sentences we are interested in here follow the template below:

(5) $$\text{Wh did Subj } V_1 \left\{ \begin{array}{c} \text{pro} \\ \text{—} \end{array} \right\} \left\{ \begin{array}{c} \text{before} \\ \text{after} \\ \text{without} \end{array} \right\} V_2\text{-ing} \left\{ \begin{array}{c} \text{pro} \\ \text{—}(pg) \end{array} \right\} ?$$

- In particular, a sentence such as (6a), containing an object pronoun in the matrix clause but a gap in the embedded clause, is bad due to:
  1. **the gap being located in an adjunct island;**
  2. **the gap not being parasitic on anything.**
- A sentence such as (6b) may be semantically weird out of the blue, but is syntactically OK.

(6) a. * What did you discard **it** after using __?

b. What did you discard __ after using **it**?

- We first test if LLMs are attuned to this contrast in the gap's position. This will allow to ensure that (6a) is a good ungrammatical baseline.

Design

(7) a. * What did you V_1 {it, this, that} {before, after, without} V_2-ing __ ?

b. What did you V_1 __ {before, after, without} V_2-ing {it, this, that} ?

- 2 gap/pro configurations (=independent variable), see (7).
- To build the various "frames":
  - **367 pairs of matrix and adjunct verbs** curated to ensure minimal semantic consistence, all strongly transitive and compatible with an inanimate object.[1]
  - **3 possible adjunct-introducing prepositions**: *before*, *after*, *without*;
  - **3 possible pronouns** in place of gaps: *it*, *this*, *that*.
- Totalling to $367 \times 3 \times 3 \times 2 = 6606$ paired sentences.
- Sentence surprisals (normalized by the number of tokens) were computed for each sentence.

[1]Chosen among: 'tell', 'get', 'send', 'love', 'taste', 'kiss', 'notice', 'state', 'make', 'obtain', 'hug', 'hate', 'like', 'assert', 'learn', 'do', 'repair', 'sell', 'discard', 'destroy', 'buy', 'borrow', 'use', 'suspect', 'burn', 'dislike', 'recognize', 'discover', 'say', 'devour'.

Design

(7) a. * What did you V_1 {it, this, that} {before, after, without} V_2-ing __ ?

b. What did you V_1 __ {before, after, without} V_2-ing {it, this, that} ?

- 2 gap/pro configurations (=independent variable), see (7).
- To build the various "frames":
  - **367 pairs of matrix and adjunct verbs** curated to ensure minimal semantic consistence, all strongly transitive and compatible with an inanimate object.[1]
  - **3 possible adjunct-introducing prepositions**: *before*, *after*, *without*;
  - **3 possible pronouns** in place of gaps: *it*, *this*, *that*.
- Totalling to $367 \times 3 \times 3 \times 2 = 6606$ paired sentences.
- Sentence surprisals (normalized by the number of tokens) were computed for each sentence.

[1] Chosen among: 'tell', 'get', 'send', 'love', 'taste', 'kiss', 'notice', 'state', 'make', 'obtain', 'hug', 'hate', 'like', 'assert', 'learn', 'do', 'repair', 'sell', 'discard', 'destroy', 'buy', 'borrow', 'use', 'suspect', 'burn', 'dislike', 'recognize', 'discover', 'say', 'devour'.

Design

(7) a. * What did you V_1 {it, this, that} {before, after, without} V_2-ing __ ?

b. What did you V_1 __ {before, after, without} V_2-ing {it, this, that} ?

- 2 gap/pro configurations (=independent variable), see (7).
- To build the various "frames":
  - **367 pairs of matrix and adjunct verbs** curated to ensure minimal semantic consistence, all strongly transitive and compatible with an inanimate object.[1]
  - **3 possible adjunct-introducing prepositions**: *before*, *after*, *without*;
  - **3 possible pronouns** in place of gaps: *it*, *this*, *that*.
- Totalling to $367 \times 3 \times 3 \times 2 = 6606$ paired sentences.
- Sentence surprisals (normalized by the number of tokens) were computed for each sentence.

[1] Chosen among: 'tell', 'get', 'send', 'love', 'taste', 'kiss', 'notice', 'state', 'make', 'obtain', 'hug', 'hate', 'like', 'assert', 'learn', 'do', 'repair', 'sell', 'discard', 'destroy', 'buy', 'borrow', 'use', 'suspect', 'burn', 'dislike', 'recognize', 'discover', 'say', 'devour'.

Design

(7) a. * What did you V_1 {it, this, that} {before, after, without} V_2-ing __ ?

b. What did you V_1 __ {before, after, without} V_2-ing {it, this, that} ?

- 2 gap/pro configurations (=independent variable), see (7).
- To build the various "frames":
    - **367 pairs of matrix and adjunct verbs** curated to ensure minimal semantic consistence, all strongly transitive and compatible with an inanimate object.[1]
    - **3 possible adjunct-introducing prepositions**: *before*, *after*, *without*;
    - **3 possible pronouns** in place of gaps: *it*, *this*, *that*.
- Totalling to $367 \times 3 \times 3 \times 2 = 6606$ paired sentences.
- Sentence surprisals (normalized by the number of tokens) were computed for each sentence.

[1] Chosen among: 'tell', 'get', 'send', 'love', 'taste', 'kiss', 'notice', 'state', 'make', 'obtain', 'hug', 'hate', 'like', 'assert', 'learn', 'do', 'repair', 'sell', 'discard', 'destroy', 'buy', 'borrow', 'use', 'suspect', 'burn', 'dislike', 'recognize', 'discover', 'say', 'devour'.

Design

(7) a. * What did you V_1 {it, this, that} {before, after, without} V_2-ing __ ?

b. What did you V_1 __ {before, after, without} V_2-ing {it, this, that} ?

- 2 gap/pro configurations (=independent variable), see (7).
- To build the various "frames":
  - **367 pairs of matrix and adjunct verbs** curated to ensure minimal semantic consistence, all strongly transitive and compatible with an inanimate object.[1]
  - **3 possible adjunct-introducing prepositions**: *before*, *after*, *without*;
  - **3 possible pronouns** in place of gaps: *it*, *this*, *that*.
- Totalling to $367 \times 3 \times 3 \times 2 = 6606$ paired sentences.
- Sentence surprisals (normalized by the number of tokens) were computed for each sentence.

[1] Chosen among: 'tell', 'get', 'send', 'love', 'taste', 'kiss', 'notice', 'state', 'make', 'obtain', 'hug', 'hate', 'like', 'assert', 'learn', 'do', 'repair', 'sell', 'discard', 'destroy', 'buy', 'borrow', 'use', 'suspect', 'burn', 'dislike', 'recognize', 'discover', 'say', 'devour'.

Design

(7) a. *What did you V_1 {it, this, that} {before, after, without} V_2-ing __ ?

b. What did you V_1 __ {before, after, without} V_2-ing {it, this, that} ?

- 2 gap/pro configurations (=independent variable), see (7).
- To build the various "frames":
  - **367 pairs of matrix and adjunct verbs** curated to ensure minimal semantic consistence, all strongly transitive and compatible with an inanimate object.[1]
  - **3 possible adjunct-introducing prepositions**: *before*, *after*, *without*;
  - **3 possible pronouns** in place of gaps: *it*, *this*, *that*.
- Totalling to $367 \times 3 \times 3 \times 2 = 6606$ paired sentences.
- Sentence surprisals (normalized by the number of tokens) were computed for each sentence.

[1] Chosen among: 'tell', 'get', 'send', 'love', 'taste', 'kiss', 'notice', 'state', 'make', 'obtain', 'hug', 'hate', 'like', 'assert', 'learn', 'do', 'repair', 'sell', 'discard', 'destroy', 'buy', 'borrow', 'use', 'suspect', 'burn', 'dislike', 'recognize', 'discover', 'say', 'devour'.

Design

(7) a. * What did you V_1 {it, this, that} {before, after, without} V_2-ing __ ?

b. What did you V_1 __ {before, after, without} V_2-ing {it, this, that} ?

- 2 gap/pro configurations (=independent variable), see (7).
- To build the various "frames":
  - **367 pairs of matrix and adjunct verbs** curated to ensure minimal semantic consistence, all strongly transitive and compatible with an inanimate object.[1]
  - **3 possible adjunct-introducing prepositions**: *before*, *after*, *without*;
  - **3 possible pronouns** in place of gaps: *it*, *this*, *that*.
- Totalling to $367 \times 3 \times 3 \times 2 = 6606$ paired sentences.
- Sentence surprisals (normalized by the number of tokens) were computed for each sentence.

[1] Chosen among: 'tell', 'get', 'send', 'love', 'taste', 'kiss', 'notice', 'state', 'make', 'obtain', 'hug', 'hate', 'like', 'assert', 'learn', 'do', 'repair', 'sell', 'discard', 'destroy', 'buy', 'borrow', 'use', 'suspect', 'burn', 'dislike', 'recognize', 'discover', 'say', 'devour'.

Design

(7) a. * What did you V_1 {it, this, that} {before, after, without} V_2-ing __ ?

b. What did you V_1 __ {before, after, without} V_2-ing {it, this, that} ?

- 2 gap/pro configurations (=independent variable), see (7).
- To build the various "frames":
  - **367 pairs of matrix and adjunct verbs** curated to ensure minimal semantic consistence, all strongly transitive and compatible with an inanimate object.[1]
  - **3 possible adjunct-introducing prepositions**: *before*, *after*, *without*;
  - **3 possible pronouns** in place of gaps: *it*, *this*, *that*.
- Totalling to $367 \times 3 \times 3 \times 2 = 6606$ paired sentences.
- Sentence surprisals (normalized by the number of tokens) were computed for each sentence.

[1] Chosen among: 'tell', 'get', 'send', 'love', 'taste', 'kiss', 'notice', 'state', 'make', 'obtain', 'hug', 'hate', 'like', 'assert', 'learn', 'do', 'repair', 'sell', 'discard', 'destroy', 'buy', 'borrow', 'use', 'suspect', 'burn', 'dislike', 'recognize', 'discover', 'say', 'devour'.

Testing & Results

- One-tailed Wilcoxon test for matched pairs: we expect (7a) to be systematically more surprising than (7b).
- **Contrast found in 3/4 models with large effect sizes** (Cliff's Δ).
- This suggests most models prefer gaps outside adjunct islands, *when there is only one gap in the sentence*.

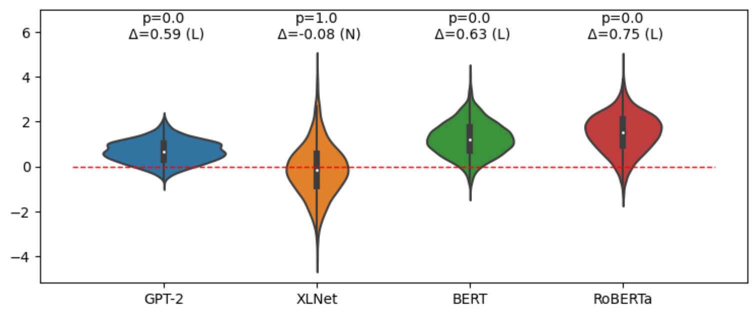


Figure 1: Surprisal contrasts between (7a) and (7b) for all 4 models.

Testing & Results

- One-tailed Wilcoxon test for matched pairs: we expect (7a) to be systematically more surprising than (7b).
- **Contrast found in 3/4 models with large effect sizes** (Cliff's Δ).
- This suggests most models prefer gaps outside adjunct islands, *when there is only one gap in the sentence*.

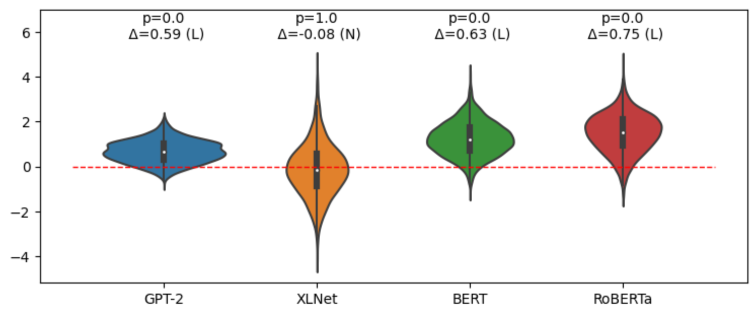


Figure 1: Surprisal contrasts between (7a) and (7b) for all 4 models.

- One-tailed Wilcoxon test for matched pairs: we expect (7a) to be systematically more surprising than (7b).
- **Contrast found in 3/4 models with large effect sizes** (Cliff's Δ).
- This suggests most models prefer gaps outside adjunct islands, *when there is only one gap in the sentence*.

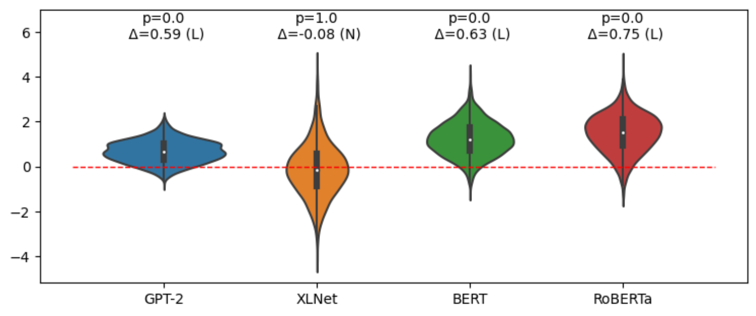


Figure 1: Surprisal contrasts between (7a) and (7b) for all 4 models.

Task 1: PG-licensing at the sentence-level

Design

- Recall the template (5):

$$\text{(5)}\quad \text{Wh did Subj } V_1 \left\{ \begin{array}{c} \text{pro} \\ \_\_ \end{array} \right\} \left\{ \begin{array}{c} \text{before} \\ \text{after} \\ \text{without} \end{array} \right\} V_2\text{-ing} \left\{ \begin{array}{c} \text{pro} \\ \_\_(pg) \end{array} \right\} ?$$

- In Task 0 we focused on single-gap configurations, to confirm a dispreference for gaps within adjuncts.
- Now, we want to verify if LLMs capture the "parasitic" nature of PGs, by comparing:
  1. a **multiple gap configuration** whereby the PG (located in the adjunct) is licensed by a matrix gap, cf. (8b)...
  2. ...to a **single-gap, island-violating configuration** only involving an adjunct gap, cf. (8a)=(7a).

(8) a. * What did you V_1 {it, this, that} {before, after, without} V_2-ing __ ?

b. What did you V_1 __ {before, after, without} V_2-ing __ ?

Design

- Recall the template (5):

(5) Wh did Subj V_1 $\left\{ \begin{array}{c} \text{pro} \\ \_\_ \end{array} \right\} \left\{ \begin{array}{c} \text{before} \\ \text{after} \\ \text{without} \end{array} \right\}$ V_2-ing $\left\{ \begin{array}{c} \text{pro} \\ \_\__{(pg)} \end{array} \right\}$?

- In Task 0 we focused on single-gap configurations, to confirm a dispreference for gaps within adjuncts.
- Now, we want to verify if LLMs capture the "parasitic" nature of PGs, by comparing:
  1. a **multiple gap configuration** whereby the PG (located in the adjunct) is licensed by a matrix gap, cf. (8b)...
  2. ...to a **single-gap, island-violating configuration** only involving an adjunct gap, cf. (8a)=(7a).

(8) a. * What did you V_1 {it, this, that} {before, after, without} V_2-ing __ ?

b. What did you V_1 __ {before, after, without} V_2-ing __ ?

Design

- Recall the template (5):

$$\text{(5)}\quad \text{Wh did Subj } V_1 \left\{ \begin{array}{c} \text{pro} \\ \_\_ \end{array} \right\} \left\{ \begin{array}{c} \text{before} \\ \text{after} \\ \text{without} \end{array} \right\} V_2\text{-ing} \left\{ \begin{array}{c} \text{pro} \\ \_\__{(pg)} \end{array} \right\} ?$$

- In Task 0 we focused on single-gap configurations, to confirm a dispreference for gaps within adjuncts.
- Now, we want to verify if LLMs capture the "parasitic" nature of PGs, by comparing:
  1. a **multiple gap configuration** whereby the PG (located in the adjunct) is licensed by a matrix gap, cf. (8b)...
  2. ...to a **single-gap, island-violating configuration** only involving an adjunct gap, cf. (8a)=(7a).

(8) a. * What did you V_1 {it, this, that} {before, after, without} V_2-ing __ ?

b. What did you V_1 __ {before, after, without} V_2-ing __ ?

Design

- Recall the template (5):

(5) Wh did Subj V_1 $\left\{ \begin{array}{c} \text{pro} \\ \text{—} \end{array} \right\} \left\{ \begin{array}{c} \text{before} \\ \text{after} \\ \text{without} \end{array} \right\}$ V_2-ing $\left\{ \begin{array}{c} \text{pro} \\ \text{—}(\textit{pg}) \end{array} \right\}$?

- In Task 0 we focused on single-gap configurations, to confirm a dispreference for gaps within adjuncts.
- Now, we want to verify if LLMs capture the "parasitic" nature of PGs, by comparing:
  1. a **multiple gap configuration** whereby the PG (located in the adjunct) is licensed by a matrix gap, cf. (8b)...
  2. ...to a **single-gap, island-violating configuration** only involving an adjunct gap, cf. (8a)=(7a).

(8) a. * What did you V_1 {it, this, that} {before, after, without} V_2-ing __ ?

b. What did you V_1 __ {before, after, without} V_2-ing __ ?

Design

- Recall the template (5):

$$\text{(5)}\quad \text{Wh did Subj } V_1 \begin{Bmatrix} \text{pro} \\ \text{\_\_} \end{Bmatrix} \begin{Bmatrix} \text{before} \\ \text{after} \\ \text{without} \end{Bmatrix} V_2\text{-ing} \begin{Bmatrix} \text{pro} \\ \text{\_\_}_{(pg)} \end{Bmatrix} ?$$

- In Task 0 we focused on single-gap configurations, to confirm a dispreference for gaps within adjuncts.
- Now, we want to verify if LLMs capture the "parasitic" nature of PGs, by comparing:
    1. a **multiple gap configuration** whereby the PG (located in the adjunct) is licensed by a matrix gap, cf. (8b)...
    2. ...to a **single-gap, island-violating configuration** only involving an adjunct gap, cf. (8a)=(7a).

(8) a. * What did you V_1 {it, this, that} {before, after, without} V_2-ing __ ?

b. What did you V_1 __ {before, after, without} V_2-ing __ ?

Testing & Results

- Same frames, test and scoring method as in Task 0. Here we expect (8a) to be more surprising than (8b).
- **Contrast found in 2/3 models (the bidirectional ones) which succeeded in Task 0, with medium to large effect sizes.**
- This suggests some models prefer a gap in an adjunct island *when it is parasitic on a matrix gap*, as opposed to when it is not.

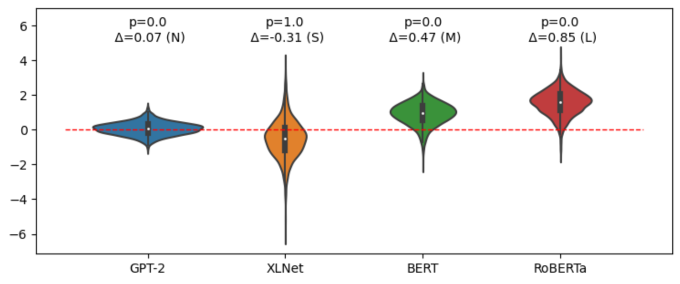


Figure 2: Surprisal contrasts between (8a) and (8b). Note XLNet is not extremely relevant as it failed on Task 0.

Testing & Results

- Same frames, test and scoring method as in Task 0. Here we expect (8a) to be more surprising than (8b).
- **Contrast found in 2/3 models (the bidirectional ones) which succeeded in Task 0, with medium to large effect sizes**.
- This suggests some models prefer a gap in an adjunct island *when it is parasitic on a matrix gap*, as opposed to when it is not.

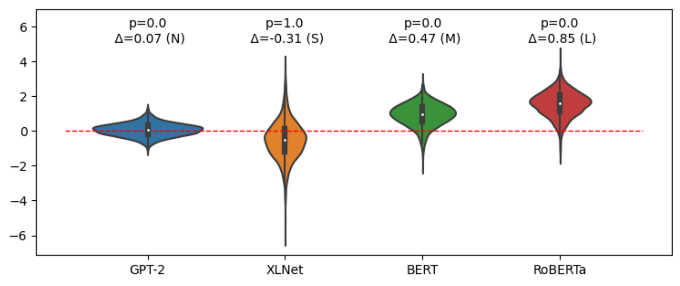


Figure 2: Surprisal contrasts between (8a) and (8b). Note XLNet is not extremely relevant as it failed on Task 0.

Testing & Results

- Same frames, test and scoring method as in Task 0. Here we expect (8a) to be more surprising than (8b).
- **Contrast found in 2/3 models (the bidirectional ones) which succeeded in Task 0, with medium to large effect sizes**.
- This suggests some models prefer a gap in an adjunct island *when it is parasitic on a matrix gap*, as opposed to when it is not.

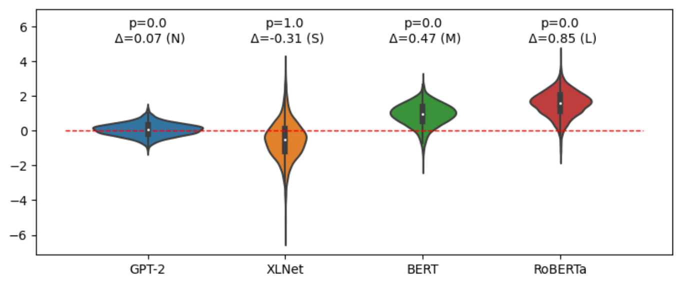


Figure 2: Surprisal contrasts between (8a) and (8b). Note XLNet is not extremely relevant as it failed on Task 0.

Task 2: testing PG-specificity at the word-level

Why look at word-level surprisals?

- Task 0 and 1 measured global grammaticality scores in the form of (normalized) sentence surprisal.
- Even if the sentences tested were minimal pairs, given the complex architecture of modern LLMs (especially bidirectional ones!), **it is hard to tell if the minimally differing elements really drive the surprisal contrasts**...
- Let's investigate the *processing* of (8a) vs. (8b):

*(8a)	What	did	you	V_1	**pro**	prep	V_2-ing	**?**
(8b)	What	did	you	V_1	**prep**		V_2-ing	**?**

- A human subject would be more puzzled reading:
  - **pro** (=*it*, *this*, or *that*) after a matrix strongly transitive V as in (8a), as opposed to reading **prep** (=*before*, *after*, or *without*), as in (8b).
  - a final **?** following the verb (suggesting an illicit gap) as in (8a), as opposed to the same **?** (suggesting a licensed PG) as in (8b).

Why look at word-level surprisals?

- Task 0 and 1 measured global grammaticality scores in the form of (normalized) sentence surprisal.
- Even if the sentences tested were minimal pairs, given the complex architecture of modern LLMs (especially bidirectional ones!), **it is hard to tell if the minimally differing elements really drive the surprisal contrasts**...
- Let's investigate the *processing* of (8a) vs. (8b):

*(8a)	What	did	you	V_1	**pro**	prep	V_2-ing	**?**
(8b)	What	did	you	V_1	**prep**		V_2-ing	**?**

- A human subject would be more puzzled reading:
  - **pro** (=*it*, *this*, or *that*) after a matrix strongly transitive V as in (8a), as opposed to reading **prep** (=*before*, *after*, or *without*), as in (8b).
  - a final **?** following the verb (suggesting an illicit gap) as in (8a), as opposed to the same **?** (suggesting a licensed PG) as in (8b).

Why look at word-level surprisals?

- Task 0 and 1 measured global grammaticality scores in the form of (normalized) sentence surprisal.
- Even if the sentences tested were minimal pairs, given the complex architecture of modern LLMs (especially bidirectional ones!), **it is hard to tell if the minimally differing elements really drive the surprisal contrasts**...
- Let's investigate the *processing* of (8a) vs. (8b):

*(8a)	What	did	you	V_1	**pro**	prep	V_2-ing	**?**
(8b)	What	did	you	V_1	**prep**		V_2-ing	**?**

- A human subject would be more puzzled reading:
  - **pro** (=*it*, *this*, or *that*) after a matrix strongly transitive V as in (8a), as opposed to reading **prep** (=*before*, *after*, or *without*), as in (8b).
  - a final **?** following the verb (suggesting an illicit gap) as in (8a), as opposed to the same **?** (suggesting a licensed PG) as in (8b).

Why look at word-level surprisals?

- Task 0 and 1 measured global grammaticality scores in the form of (normalized) sentence surprisal.
- Even if the sentences tested were minimal pairs, given the complex architecture of modern LLMs (especially bidirectional ones!), **it is hard to tell if the minimally differing elements really drive the surprisal contrasts**...
- Let's investigate the *processing* of (8a) vs. (8b):

*(8a)	What	did	you	V_1	**pro**	prep	V_2-ing	**?**
(8b)	What	did	you	V_1	**prep**		V_2-ing	**?**

- A human subject would be more puzzled reading:
  - **pro** (=*it*, *this*, or *that*) after a matrix strongly transitive V as in (8a), as opposed to reading **prep** (=*before*, *after*, or *without*), as in (8b).
  - a final **?** following the verb (suggesting an illicit gap) as in (8a), as opposed to the same **?** (suggesting a licensed PG) as in (8b).

Why look at word-level surprisals?

- Task 0 and 1 measured global grammaticality scores in the form of (normalized) sentence surprisal.
- Even if the sentences tested were minimal pairs, given the complex architecture of modern LLMs (especially bidirectional ones!), **it is hard to tell if the minimally differing elements really drive the surprisal contrasts**...
- Let's investigate the *processing* of (8a) vs. (8b):

*(8a)	What	did	you	V_1	**pro**	prep	V_2-ing	**?**
(8b)	What	did	you	V_1	**prep**		V_2-ing	**?**

- A human subject would be more puzzled reading:
  - **pro** (=*it*, *this*, or *that*) after a matrix strongly transitive V as in (8a), as opposed to reading **prep** (=*before*, *after*, or *without*), as in (8b).
  - a final **?** following the verb (suggesting an illicit gap) as in (8a), as opposed to the same **?** (suggesting a licensed PG) as in (8b).

Why look at word-level surprisals?

- Task 0 and 1 measured global grammaticality scores in the form of (normalized) sentence surprisal.
- Even if the sentences tested were minimal pairs, given the complex architecture of modern LLMs (especially bidirectional ones!), **it is hard to tell if the minimally differing elements really drive the surprisal contrasts**...
- Let's investigate the *processing* of (8a) vs. (8b):

*(8a)	What	did	you	V_1	**pro**	prep	V_2-ing	**?**
(8b)	What	did	you	V_1	**prep**		V_2-ing	**?**

- A human subject would be more puzzled reading:
  - **pro** (=*it*, *this*, or *that*) after a matrix strongly transitive V as in (8a), as opposed to reading **prep** (=*before*, *after*, or *without*), as in (8b).
  - a final **?** following the verb (suggesting an illicit gap) as in (8a), as opposed to the same **?** (suggesting a licensed PG) as in (8b).

Testing & Results

- Word-level surprisal scores were computed using `minicons`.
- Differences in the scores of **pro**$_{\mathbf{(8a)}}$ vs. **prep**$_{\mathbf{(8b)}}$, and **?**$_{\mathbf{(8a)}}$ vs. **?**$_{\mathbf{(8b)}}$ were assessed using Wilcoxon tests:
  - The two left-to-right models (GPT-2 and XLNet) do not show significant effects...
  - While the bidirectional models (BERT and RoBERTa) tend to show the expected contrasts.

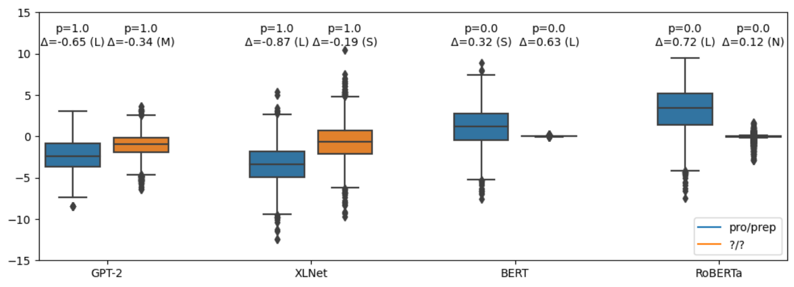


Figure 3: Differences in surprisal between critical words of (8a) and (8b)

Testing & Results

- Word-level surprisal scores were computed using `minicons`.
- Differences in the scores of **pro**$_{\mathbf{(8a)}}$ vs. **prep**$_{\mathbf{(8b)}}$, and **?**$_{\mathbf{(8a)}}$ vs. **?**$_{\mathbf{(8b)}}$ were assessed using Wilcoxon tests:
  - The two left-to-right models (GPT-2 and XLNet) do not show significant effects...
  - While the bidirectional models (BERT and RoBERTa) tend to show the expected contrasts.

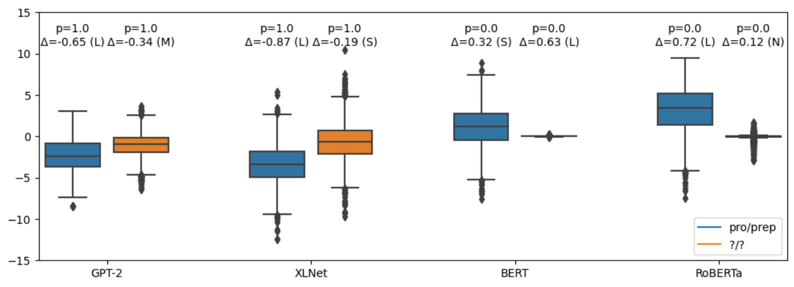


Figure 3: Differences in surprisal between critical words of (8a) and (8b)

Testing & Results

- Word-level surprisal scores were computed using `minicons`.
- Differences in the scores of **pro**$_{(8a)}$ vs. **prep**$_{(8b)}$, and **?**$_{(8a)}$ vs. **?**$_{(8b)}$ were assessed using Wilcoxon tests:
  - The two left-to-right models (GPT-2 and XLNet) do not show significant effects...
  - While the bidirectional models (BERT and RoBERTa) tend to show the expected contrasts.

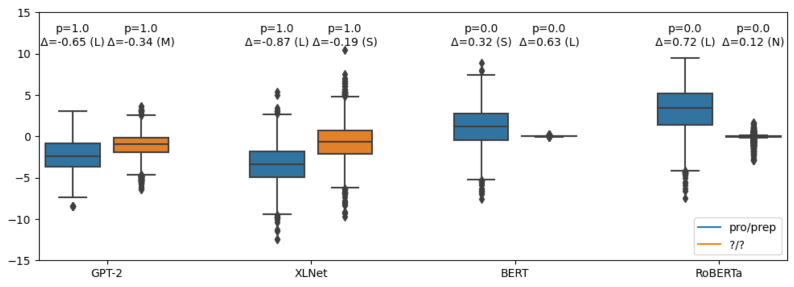


Figure 3: Differences in surprisal between critical words of (8a) and (8b)

Testing & Results

- Word-level surprisal scores were computed using `minicons`.
- Differences in the scores of **pro**$_{\textbf{(8a)}}$ vs. **prep**$_{\textbf{(8b)}}$, and **?**$_{\textbf{(8a)}}$ vs. **?**$_{\textbf{(8b)}}$ were assessed using Wilcoxon tests:
  - The two left-to-right models (GPT-2 and XLNet) do not show significant effects...
  - While the bidirectional models (BERT and RoBERTa) tend to show the expected contrasts.

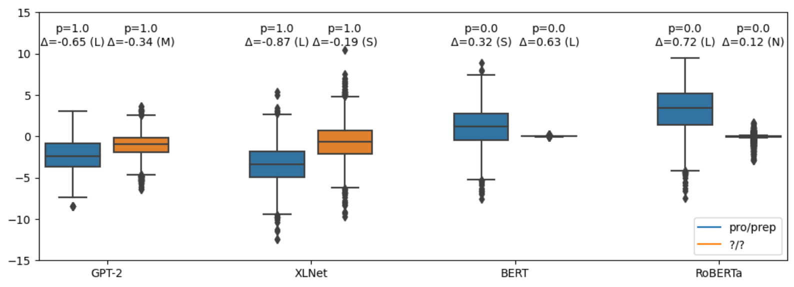


Figure 3: Differences in surprisal between critical words of (8a) and (8b)

Testing & Results

- Word-level surprisal scores were computed using `minicons`.
- Differences in the scores of **pro**$_{(8a)}$ vs. **prep**$_{(8b)}$, and **?**$_{(8a)}$ vs. **?**$_{(8b)}$ were assessed using Wilcoxon tests:
  - The two left-to-right models (GPT-2 and XLNet) do not show significant effects...
  - While the bidirectional models (BERT and RoBERTa) tend to show the expected contrasts.

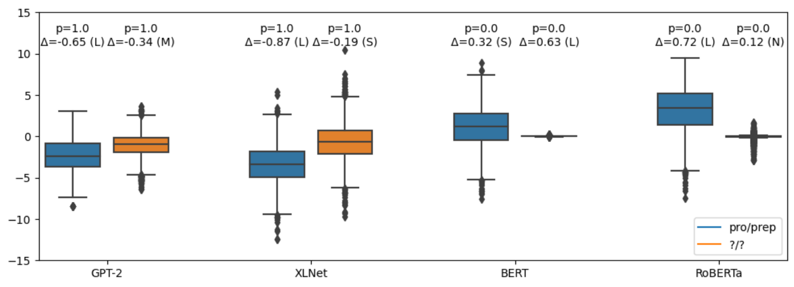
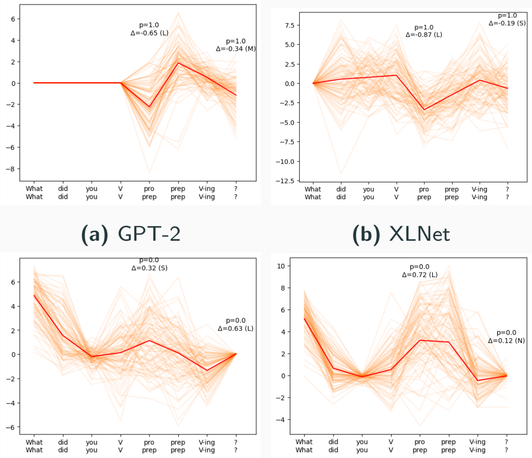


Figure 3: Differences in surprisal between critical words of (8a) and (8b)

What about the other words in the sentences?

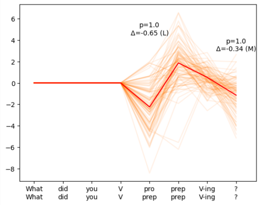


(a) GPT-2 **(b)** XLNet **(c)** BERT **(d)** RoBERTa

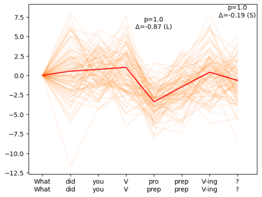
Figure 4: Paired surprisal differences between the words of (8a) vs. (8b). 100 samples (orange lines). Red lines represent averages over the whole dataset.

- Left-to-right models are "unsurprised" by the object pronoun, yet GPT-2 is more surprised to see a preposition after it...
- Bidirectional models "spread" the surprisal across different items in the sentence; the *wh*-word in particular!
- No model (except BERT perhaps) exhibits a notable peak of surprisal towards the end of the sentence.

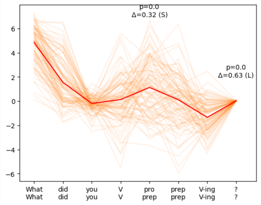
What about the other words in the sentences?



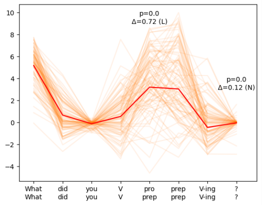
(a) GPT-2



(b) XLNet



(c) BERT

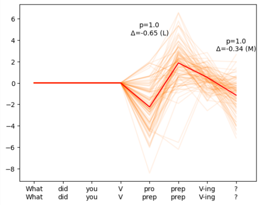


(d) RoBERTa

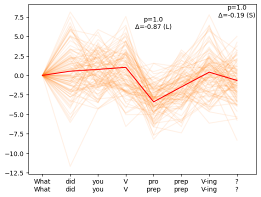
Figure 4: Paired surprisal differences between the words of (8a) vs. (8b). 100 samples (orange lines). Red lines represent averages over the whole dataset.

- Left-to-right models are "unsurprised" by the object pronoun, yet GPT-2 is more surprised to see a preposition after it...
- Bidirectional models "spread" the surprisal across different items in the sentence; the *wh*-word in particular!
- No model (except BERT perhaps) exhibits a notable peak of surprisal towards the end of the sentence.

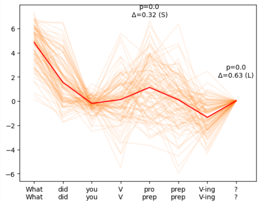
What about the other words in the sentences?



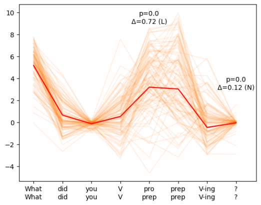
(a) GPT-2



(b) XLNet



(c) BERT



(d) RoBERTa

Figure 4: Paired surprisal differences between the words of (8a) vs. (8b). 100 samples (orange lines). Red lines represent averages over the whole dataset.

- Left-to-right models are "unsurprised" by the object pronoun, yet GPT-2 is more surprised to see a preposition after it...
- Bidirectional models "spread" the surprisal across different items in the sentence; the *wh*-word in particular!
- No model (except BERT perhaps) exhibits a notable peak of surprisal towards the end of the sentence.

Discussion

- The fact that left-to-right LLMs, which intuitively, are closer to human readers, did not succeed in capturing the expected processing contrasts is **puzzling at first blush**.
- On the other hand, bidirectional LLMs may in principle use the information contained in the adjunct clauses in (8b) and (8a) to compute the probability of resp. a gap and a pronoun in the matrix clause...**which may reinforce the surprisal contrasts in that position**.
    - This kind of behavior might be compared to human **backtracking** when processing syntactic dependencies.
    - This may also explain why bidirectional LLMs "found" the presence of an **initial wh-word** so puzzling in (8a) as opposed to (8b): it is not binding any legit gap!
    - Finally, this might partly explain **why the sentence-final contrasts are somewhat weak**: bidirectional LLMs may prefer to "blame" the *wh*-word instead of the gap present in the adjunct clause.

Conclusion

- PGs are a kind of empirically rare syntactic dependency which had not been previously investigated in the context of LLMs before.
- We showed, using island-violating structures as a baseline, that **some but not all recent LLMs distinguish PGs from regular gaps**.
- Yet the specific representation that LLMs assign to PGs remains unclear.
- Future work may involve:
  - using **intransitive matrix verbs as controls**, as opposed to saturated strongly transitive ones;
  - testing if LLMs understand PGs as a proper dependency, or as some sort of contextually-determined covert pronoun, by testing contrasts like those in (9), in which the **PG precedes the actual gap**.

(9) a. *Which girl did [the rumor about **her**] annoy __?

b. Which girl did [the rumor about $\_\_{pg}$] annoy __?

Conclusion

- PGs are a kind of empirically rare syntactic dependency which had not been previously investigated in the context of LLMs before.
- We showed, using island-violating structures as a baseline, that **some but not all recent LLMs distinguish PGs from regular gaps**.
- Yet the specific representation that LLMs assign to PGs remains unclear.
- Future work may involve:
  - using **intransitive matrix verbs as controls**, as opposed to saturated strongly transitive ones;
  - testing if LLMs understand PGs as a proper dependency, or as some sort of contextually-determined covert pronoun, by testing contrasts like those in (9), in which the **PG precedes the actual gap**.

(9) a. * Which girl did [the rumor about **her**] annoy __?

b. Which girl did [the rumor about $\_\__{pg}$] annoy __?

Conclusion

- PGs are a kind of empirically rare syntactic dependency which had not been previously investigated in the context of LLMs before.
- We showed, using island-violating structures as a baseline, that **some but not all recent LLMs distinguish PGs from regular gaps**.
- Yet the specific representation that LLMs assign to PGs remains unclear.
- Future work may involve:
  - using **intransitive matrix verbs as controls**, as opposed to saturated strongly transitive ones;
  - testing if LLMs understand PGs as a proper dependency, or as some sort of contextually-determined covert pronoun, by testing contrasts like those in (9), in which the **PG precedes the actual gap**.

(9) a. * Which girl did [the rumor about **her**] annoy __?

b. Which girl did [the rumor about $\_\__{pg}$] annoy __?

Conclusion

- PGs are a kind of empirically rare syntactic dependency which had not been previously investigated in the context of LLMs before.
- We showed, using island-violating structures as a baseline, that **some but not all recent LLMs distinguish PGs from regular gaps**.
- Yet the specific representation that LLMs assign to PGs remains unclear.
- Future work may involve:
  - using **intransitive matrix verbs as controls**, as opposed to saturated strongly transitive ones;
  - testing if LLMs understand PGs as a proper dependency, or as some sort of contextually-determined covert pronoun, by testing contrasts like those in (9), in which the **PG precedes the actual gap**.

(9) a. * Which girl did [the rumor about **her**] annoy __?
    b. Which girl did [the rumor about $\_\_{pg}$] annoy __?

Thank you !

Selected references i

Engdahl, E. (1983). Parasitic gaps. *Linguistics and Philosophy*, *6*(1), 5–34. https://doi.org/10.1007/bf00868088

Hale, J. (2001). A probabilistic earley parser as a psycholinguistic model. *Second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies 2001 - NAACL '01*. https://doi.org/10.3115/1073336.1073357

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177. https://doi.org/10.1016/j.cognition.2007.05.006

Linzen, T., Dupoux, E., & Goldberg, Y. (2016). Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, *4*, 521–535. https://doi.org/10.1162/tacl_a_00115

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *CoRR, abs/1706.03762*.

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR, abs/1810.04805*. http://arxiv.org/abs/1810.04805

Selected references ii

Gulordava, K., Bojanowski, P., Grave, E., Linzen, T., & Baroni, M. (2018). Colorless green recurrent networks dream hierarchically. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 1195–1205. https://doi.org/10.18653/v1/N18-1108

Wilcox, E., Levy, R., Morita, T., & Futrell, R. (2018). What do RNN language models learn about filler–gap dependencies? *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 211–221. https://doi.org/10.18653/v1/W18-5423

Futrell, R., Wilcox, E., Morita, T., Qian, P., Ballesteros, M., & Levy, R. (2019). Neural language models as psycholinguistic subjects: Representations of syntactic state. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 32–42. https://doi.org/10.18653/v1/N19-1004

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized BERT pretraining approach. *CoRR*, *abs/1907.11692*. http://arxiv.org/abs/1907.11692

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners.

Selected references iii

Yang, Z., Dai, Z., Yang, Y., Carbonell, J. G., Salakhutdinov, R., & Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. In H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems 32: Annual conference on neural information processing systems 2019, neurips 2019, december 8-14, 2019, vancouver, bc, canada* (pp. 5754–5764). https://proceedings.neurips.cc/paper/2019/hash/dc6a7e655d7e5840e66733e9ee67cc69-Abstract.html

Kobzeva, A., Arehalli, S., Linzen, T., & Kush, D. (2022). Lstms can learn basic wh- and relative clause dependencies in norwegian. *44th Annual Meeting of the Cognitive Science Society: Cognitive Diversity, CogSci 2022*.

Misra, K. (2022). Minicons: Enabling flexible behavioral and representational analyses of transformer language models. *arXiv preprint arXiv:2203.13112*.

Wilcox, E. G., Futrell, R., & Levy, R. (2023). Using Computational Models to Test Syntactic Learnability. *Linguistic Inquiry*, 1–44. https://doi.org/10.1162/ling_a_00491