Yale

Meaning Beyond Truth Conditions

Evaluating Discourse Level Understanding via Anaphora Accessibility

Xiaomeng Zhu*, Zhenghao Zhou*, Simon Charlow, Robert Frank
Department of Linguistics, Yale University
ACL 2025 @ Vienna, Austria

ACL 2025 VIENNA

*Equal Contributions

Example

🟧 Discourse entity 🟨 Anaphora

Example

- A farmer worked in his field. He dreamed of the harvest.

🟧 Discourse entity 🟨 Anaphora

Example

- A farmer worked in his field. He dreamed of the harvest.

🟧 Discourse entity 🟨 Anaphora

Example

- A farmer worked in his field. He dreamed of the harvest.

🟧 Discourse entity 🟨 Anaphora

Example

- A farmer worked in his field. He dreamed of the harvest.

🟧 Discourse entity 🟨 Anaphora

Example

- A farmer worked in his field. He dreamed of the harvest.
- Every farmer worked in his field.   He dreamed of the harvest.

🟧 Discourse entity   🟨 Anaphora

Example

- A farmer worked in his field. He dreamed of the harvest.
- Every farmer worked in his field. He dreamed of the harvest.

🟧 Discourse entity 🟨 Anaphora

Example

- A farmer worked in his field. He dreamed of the harvest.
- Every farmer worked in his field.   He dreamed of the harvest.

🟧 Discourse entity   🟨 Anaphora

Example

- A farmer worked in his field. He dreamed of the harvest.
- Every farmer worked in his field. He dreamed of the harvest.

🟧 Discourse entity 🟨 Anaphora

Example

- A farmer worked in his field. He dreamed of the harvest.
- Every farmer worked in his field.# He dreamed of the harvest.

🟧 Discourse entity 🟨 Anaphora

Anaphora Accessibility

Dynamic Semantics

- **A** farmer worked in his field. He dreamed of the harvest.

- **Every** farmer worked in his field.# He dreamed of the harvest.

🟧 Discourse entity ◼ **Quantifier** scope 🟨 Anaphora

Anaphora Accessibility

Dynamic Semantics

- Pronominal anaphora (i.e. using pronouns to refer back to discourse referents introduced earlier) is influenced by the semantic scope of the antecedent.

- **A** farmer worked in his field. He dreamed of the harvest.

- **Every** farmer worked in his field.# He dreamed of the harvest.

🟧 Discourse entity ◼ **Quantifier** scope 🟨 Anaphora

Anaphora Accessibility

Dynamic Semantics

- Pronominal anaphora (i.e. using pronouns to refer back to discourse referents introduced earlier) is influenced by the semantic scope of the antecedent.

- **A** farmer worked in his field. He dreamed of the harvest.

- **Every** farmer worked in his field.# He dreamed of the harvest.

🟧 Discourse entity ◼ **Quantifier** scope 🟨 Anaphora

Anaphora Accessibility

Dynamic Semantics

- Pronominal anaphora (i.e. using pronouns to refer back to discourse referents introduced earlier) is influenced by the semantic scope of the antecedent.

- **A** farmer worked in his field. He dreamed of the harvest.

- **Every** farmer worked in his field.# He dreamed of the harvest.

🟧 Discourse entity ■ **Quantifier** scope 🟨 Anaphora

Anaphora Accessibility

Dynamic Semantics

- Pronominal anaphora (i.e. using pronouns to refer back to discourse referents introduced earlier) is influenced by the semantic scope of the antecedent.

- **A** farmer worked in his field. He dreamed of the harvest.

- **Every** farmer worked in his field.# He dreamed of the harvest.

🟧 Discourse entity ◼ **Quantifier** scope 🟨 Anaphora

Anaphora Accessibility

Dynamic Semantics

- Pronominal anaphora (i.e. using pronouns to refer back to discourse referents introduced earlier) is influenced by the semantic scope of the antecedent.

- **A** farmer worked in his field. He dreamed of the harvest.

- **Every** farmer worked in his field.# He dreamed of the harvest.

🟧 Discourse entity ◼ **Quantifier** scope 🟨 Anaphora

Anaphora Accessibility

Dynamic Semantics

- Pronominal anaphora (i.e. using pronouns to refer back to discourse referents introduced earlier) is influenced by the semantic scope of the antecedent.

- **A** farmer worked in his field. He dreamed of the harvest.

- **Every** farmer worked in his field.# He dreamed of the harvest.

🟧 Discourse entity ■ **Quantifier** scope 🟨 Anaphora

Anaphora Accessibility

Dynamic Semantics

- Pronominal anaphora (i.e. using pronouns to refer back to discourse referents introduced earlier) is influenced by the semantic scope of the antecedent.

- **A** farmer worked in his field. He dreamed of the harvest.

- **Every** farmer worked in his field.# He dreamed of the harvest.

🟧 Discourse entity ◼ **Quantifier** scope 🟨 Anaphora

Anaphora Accessibility

Dynamic Semantics

- Pronominal anaphora (i.e. using pronouns to refer back to discourse referents introduced earlier) is influenced by the semantic scope of the antecedent.

- **A** farmer worked in his field. He dreamed of the harvest.

- **Every** farmer worked in his field.# He dreamed of the harvest.

🟧 Discourse entity ■ **Quantifier** scope 🟨 Anaphora

Anaphora Accessibility

Dynamic Semantics

- Pronominal anaphora (i.e. using pronouns to refer back to discourse referents introduced earlier) is influenced by the semantic scope of the antecedent.

- **A** farmer worked in his field. He dreamed of the harvest.

- **Every** farmer worked in his field.# He dreamed of the harvest.

🟧 Discourse entity ■ **Quantifier** scope 🟨 Anaphora

Anaphora Accessibility

Dynamic Semantics

- Pronominal anaphora (i.e. using pronouns to refer back to discourse referents introduced earlier) is influenced by the semantic scope of the antecedent.

- **A** farmer worked in his field. He dreamed of the harvest.

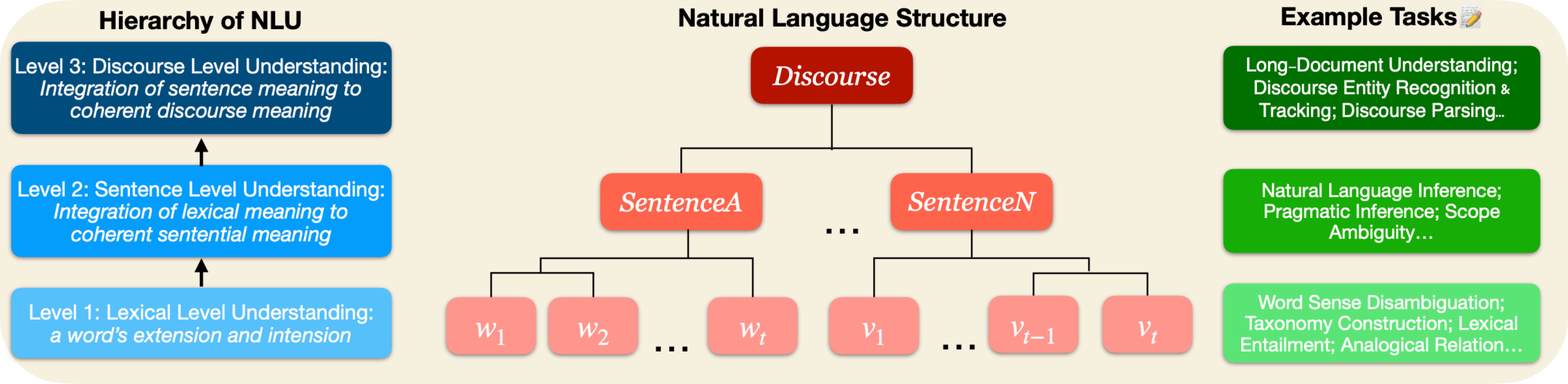
- **Every** farmer worked in his field.# He dreamed of the harvest.

🟧 Discourse entity ■ **Quantifier** scope 🟨 Anaphora

- Formalized in ‘dynamic’ variants of formal semantics, where utterances update the discourse state (e.g. Heim, 1983; Groenendijk and Stokhof, 1991; Kamp et al., 2010)

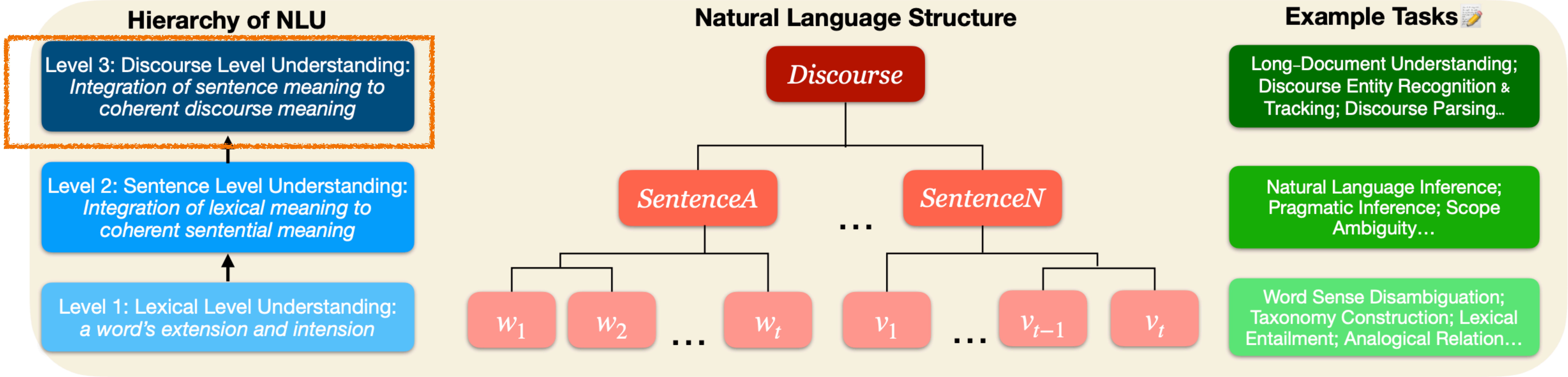
Hierarchy of NLU Abilities

There is a gap in previous tasks accessing LLM NLU abilties at the discourse level.


Hierarchy of NLU
Level 3: Discourse Level Understanding: *Integration of sentence meaning to coherent discourse meaning*
Level 2: Sentence Level Understanding: *Integration of lexical meaning to coherent sentential meaning*
Level 1: Lexical Level Understanding: *a word's extension and intension*
Natural Language Structure
Discourse
SentenceA
…
SentenceN
w_1
w_2
…
w_t
v_1
…
v_{t-1}
v_t
Example Tasks 📝
Long-Document Understanding; Discourse Entity Recognition & Tracking; Discourse Parsing…
Natural Language Inference; Pragmatic Inference; Scope Ambiguity…
Word Sense Disambiguation; Taxonomy Construction; Lexical Entailment; Analogical Relation…

Hierarchy of NLU Abilities

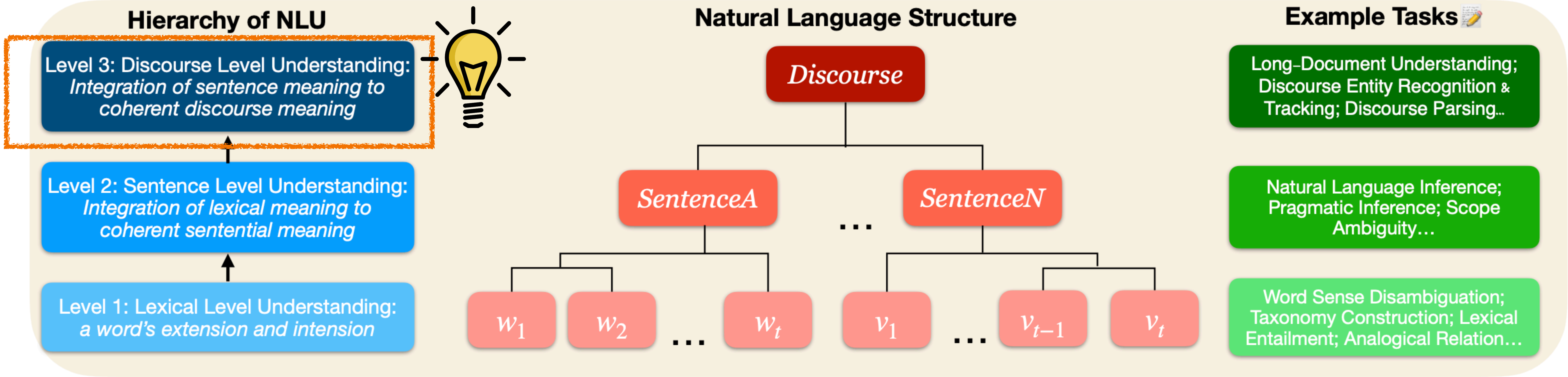
There is a gap in previous tasks accessing LLM NLU abilties at the discourse level.


Hierarchy of NLU
Level 3: Discourse Level Understanding: *Integration of sentence meaning to coherent discourse meaning*
Level 2: Sentence Level Understanding: *Integration of lexical meaning to coherent sentential meaning*
Level 1: Lexical Level Understanding: *a word's extension and intension*
Natural Language Structure
Discourse
SentenceA
…
SentenceN
w_1
w_2
…
w_t
v_1
…
v_{t-1}
v_t
Example Tasks📝
Long-Document Understanding; Discourse Entity Recognition & Tracking; Discourse Parsing…
Natural Language Inference; Pragmatic Inference; Scope Ambiguity…
Word Sense Disambiguation; Taxonomy Construction; Lexical Entailment; Analogical Relation…

Hierarchy of NLU Abilities

There is a gap in previous tasks accessing LLM NLU abilties at the discourse level.

Research Question: Do LLMs know anaphora accessibility?



Hierarchy of NLU
Level 3: Discourse Level Understanding: *Integration of sentence meaning to coherent discourse meaning*
Level 2: Sentence Level Understanding: *Integration of lexical meaning to coherent sentential meaning*
Level 1: Lexical Level Understanding: *a word's extension and intension*
Natural Language Structure
Discourse
SentenceA
...
SentenceN
w_1 w_2 ... w_t
v_1 ... v_{t-1} v_t
Example Tasks
Long-Document Understanding; Discourse Entity Recognition & Tracking; Discourse Parsing...
Natural Language Inference; Pragmatic Inference; Scope Ambiguity...
Word Sense Disambiguation; Taxonomy Construction; Lexical Entailment; Analogical Relation...

Methodology

Models & Metric

Methodology

Models & Metric

- **Open-sourse models (logit-based):**
  - Llama3-1-{8B, 8B-Instruct}, Llama3-2-{1B, 3B}; GPT3: babbage-002, davinci-002;
  - <u>Metric</u>: accessing the surprisal (negative log probability) on parts of the sentences:

$$surprisal(w_i) = \log \frac{1}{P(w_i \mid w_1, \cdots, w_{i-1})}$$

Methodology

Models & Metric

- **Open-sourse models (logit-based):**
  - Llama3-1-{8B, 8B-Instruct}, Llama3-2-{1B, 3B}; GPT3: babbage-002, davinci-002;
  - <u>Metric</u>: accessing the surprisal (negative log probability) on parts of the sentences:

$$surprisal(w_i) = \log \frac{1}{P(w_i \mid w_1, \cdots, w_{i-1})}$$

Methodology

Models & Metric

- **Open-sourse models (logit-based):**
  - Llama3-1-{8B, 8B-Instruct}, Llama3-2-{1B, 3B}; GPT3: babbage-002, davinci-002;
  - <u>Metric</u>: accessing the surprisal (negative log probability) on parts of the sentences:

$$surprisal(w_i) = \log \frac{1}{P(w_i \,|\, w_1, \cdots, w_{i-1})}$$

- **Closed-source models (prompting-based)**
  - GPT-4o;
  - <u>Metric</u>: accuracy of the model's output choice.

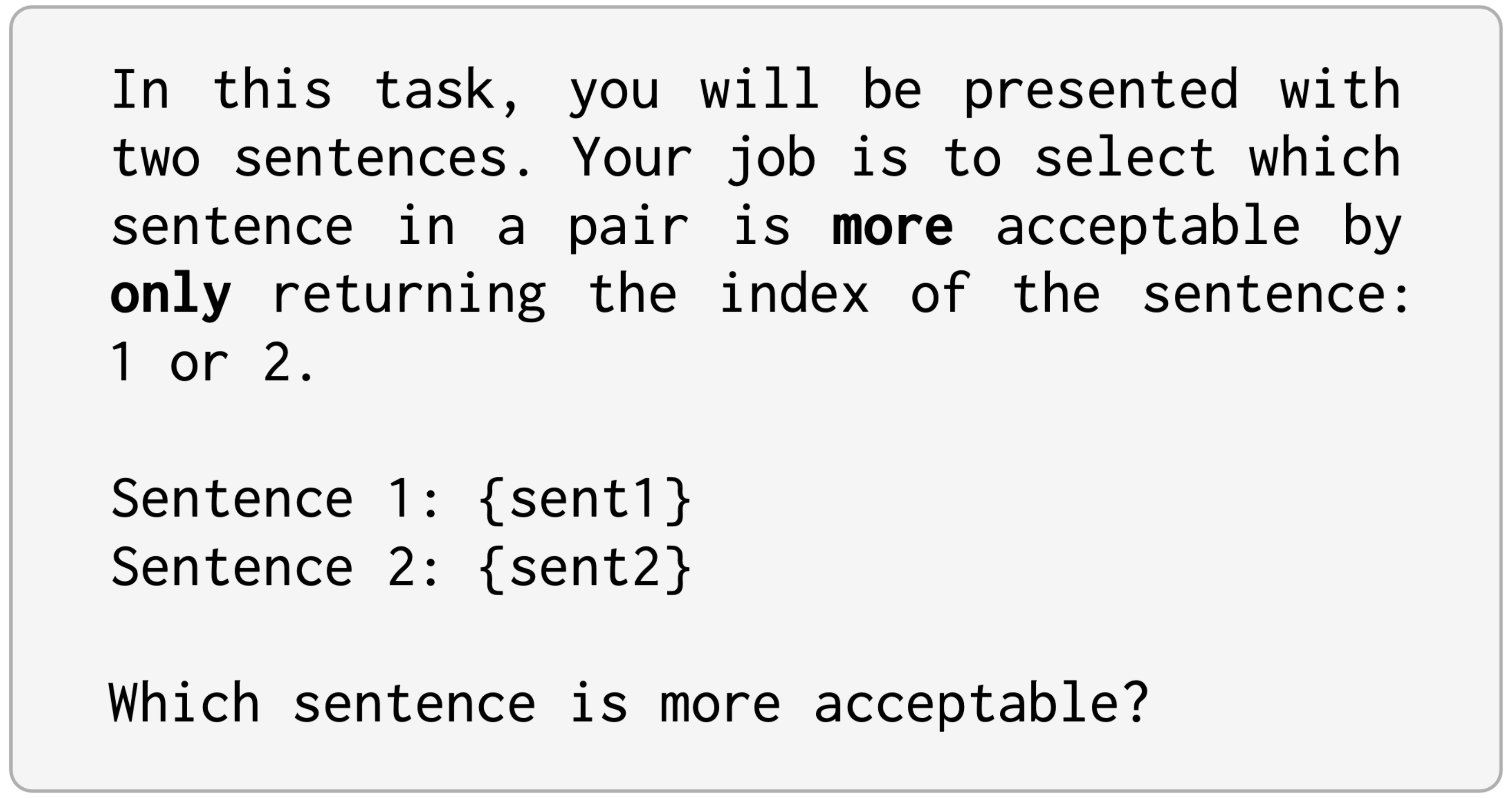
Methodology

Models & Metric

- **Open-sourse models (logit-based):**
  - Llama3-1-{8B, 8B-Instruct}, Llama3-2-{1B, 3B}; GPT3: babbage-002, davinci-002;
  - Metric: accessing the surprisal (negative log probability) on parts of the sentences:

$$surprisal(w_i) = \log \frac{1}{P(w_i \mid w_1, \cdots, w_{i-1})}$$

- **Closed-source models (prompting-based)**
  - GPT-4o;
  - Metric: accuracy of the model’s output choice.



In this task, you will be presented with two sentences. Your job is to select which sentence in a pair is **more** acceptable by **only** returning the index of the sentence: 1 or 2.

Sentence 1: {sent1}
Sentence 2: {sent2}

Which sentence is more acceptable?

Methodology

Models & Metric

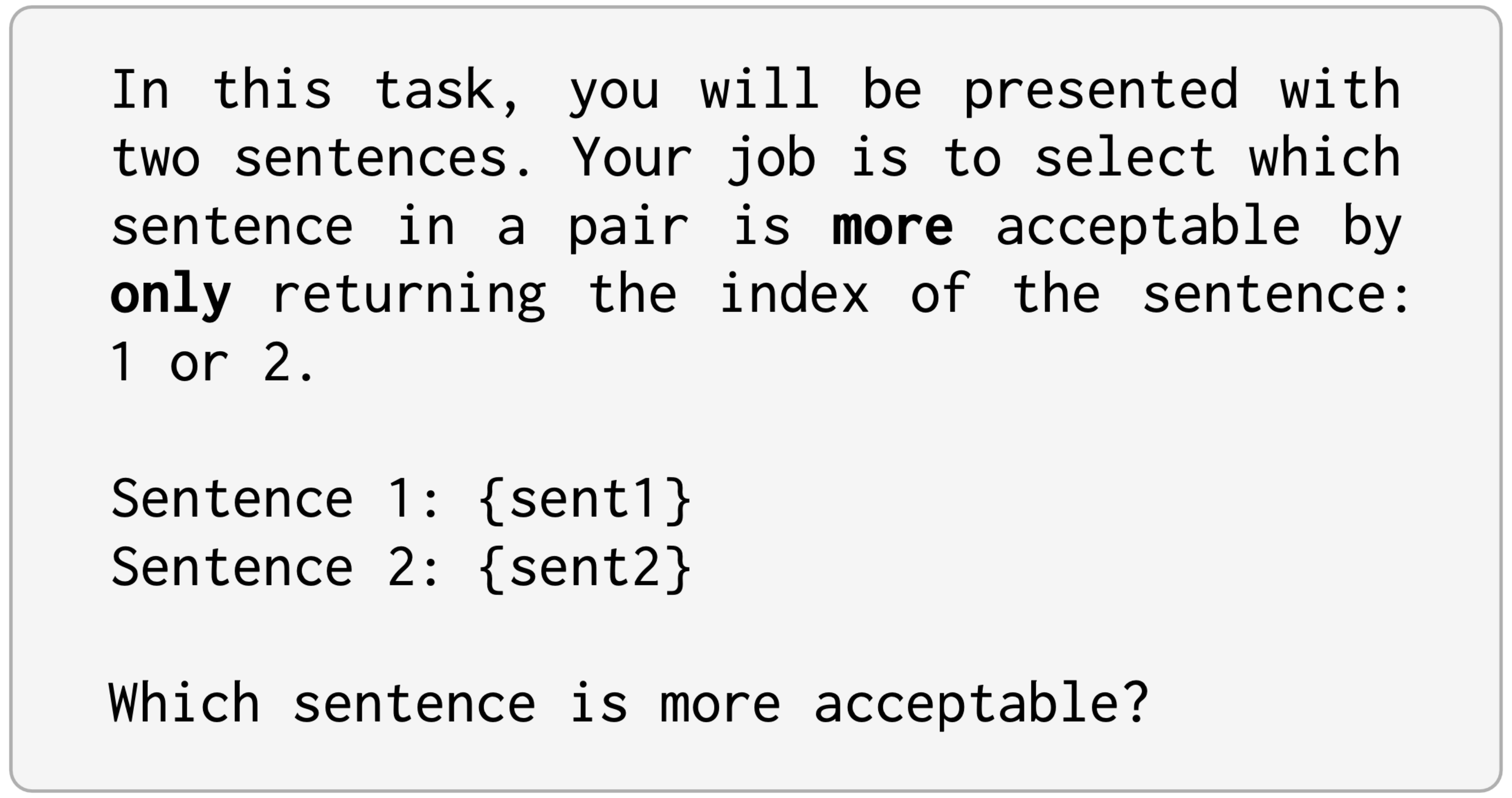
LLaMA

ChatGPT-4o

- **Open-sourse models (logit-based):**
  - Llama3-1-{8B, 8B-Instruct}, Llama3-2-{1B, 3B}; GPT3: babbage-002, davinci-002;
  - Metric: accessing the surprisal (negative log probability) on parts of the sentences:

$$surprisal(w_i) = \log \frac{1}{P(w_i \mid w_1, \cdots, w_{i-1})}$$

- **Closed-source models (prompting-based)**
  - GPT-4o;
  - Metric: accuracy of the model's output choice.



```
In this task, you will be presented with
two sentences. Your job is to select which
sentence in a pair is more acceptable by
only returning the index of the sentence:
1 or 2.

Sentence 1: {sent1}
Sentence 2: {sent2}

Which sentence is more acceptable?
```

- **Corpus**
  - 9816 sentences, synthetically generated by filling context words into structural templates;
  - Context words inspired by GPT-4o and curated for semantic plausibility by linguists.

Methodology - Cont.

Human Baseline

Methodology - Cont.

Human Baseline

- We conducted online human experiments (104 participants) to get a human baseline on the set of comparisons in the dataset.

Methodology - Cont.

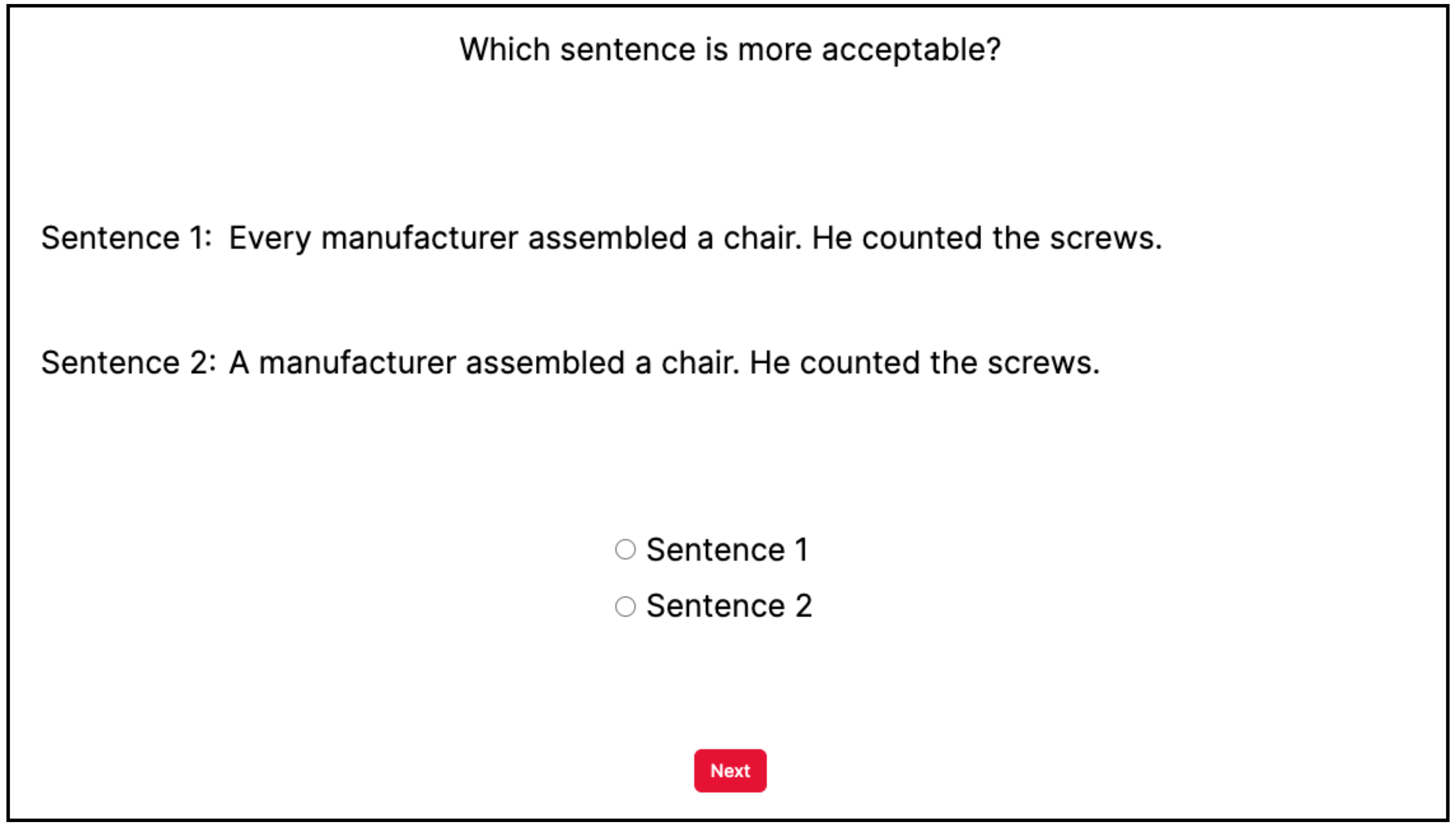
Human Baseline

- We conducted online human experiments (104 participants) to get a human baseline on the set of comparisons in the dataset.
- Forced-choice paradigm on pairs of sentences, aligning with the prompt we use on GPT-4o.

Methodology - Cont.

Human Baseline

- We conducted online human experiments (104 participants) to get a human baseline on the set of comparisons in the dataset.
- Forced-choice paradigm on pairs of sentences, aligning with the prompt we use on GPT-4o.



Which sentence is more acceptable?

Sentence 1: Every manufacturer assembled a chair. He counted the screws.

Sentence 2: A manufacturer assembled a chair. He counted the screws.

○ Sentence 1
○ Sentence 2

Results

Exp1. Universal Quantifiers
Exp2. Negation
Exp3. Disjunction

Exp1. Existential vs. Universal

- **A** farmer worked in his field. He dreamed of the harvest.

- **Every** farmer worked in his field.# He dreamed of the harvest.

- **EXISTENTIAL (∃)**: A farmer worked in the field.

- **EVERY (∀)**: Every farmer worked in the field.

- **CONTINUATION**: He dreamed of the harvest.

Exp1. Existential vs. Universal

- **A** farmer worked in his field. He dreamed of the harvest.
- **Every** farmer worked in his field.# He dreamed of the harvest.

- **EXISTENTIAL (∃)**: A farmer worked in the field.
- **EVERY (∀)**: Every farmer worked in the field.
- **CONTINUATION**: He dreamed of the harvest.

Exp1. Existential vs. Universal

- **A** farmer worked in his field. He dreamed of the harvest.
- **Every** farmer worked in his field.# He dreamed of the harvest.

- **EXISTENTIAL (∃)**: A farmer worked in the field.
- **EVERY (∀)**: Every farmer worked in the field.
- **CONTINUATION**: He dreamed of the harvest.

$$p(cont \mid \exists) > p(cont \mid \forall)$$

Exp1. Existential vs. Universal (Cont.)

Donkey Conditionals

- The farmer owns a donkey, and he beats it. It is a big one.

- If the farmer owns a donkey, he beats it.#It is a big one.

- **EXISTENTIAL (∃)**: The farmer owns a donkey, and he beats it.
- **CONDITIONAL (∀)**: {If, Whenever} the farmer owns a donkey, he beats it.
- **CONTINUATION**: It is a big one.

Exp1. Existential vs. Universal (Cont.)

Donkey Conditionals

- The farmer owns a donkey, and he beats it. It is a big one.
- If the farmer owns a donkey, he beats it.#It is a big one.

- **Existential (∃)**: The farmer owns a donkey, and he beats it.
- **Conditional (∀)**: {If, Whenever} the farmer owns a donkey, he beats it.
- **Continuation**: It is a big one.

Exp1. Existential vs. Universal (Cont.)

Donkey Conditionals

- The farmer owns a donkey, and he beats it. It is a big one.
- If the farmer owns a donkey, he beats it.#It is a big one.

- **Existential (∃)**: The farmer owns a donkey, and he beats it.
- **Conditional (∀)**: {If, Whenever} the farmer owns a donkey, he beats it.
- **Continuation**: It is a big one.

Exp1. Existential vs. Universal (Cont.)

Donkey Conditionals

- The farmer owns a donkey, and he beats it. It is a big one.

- If the farmer owns a donkey, he beats it.#It is a big one.

- **Existential (∃)**: The farmer owns a donkey, and he beats it.
- **Conditional (∀)**: {If, Whenever} the farmer owns a donkey, he beats it.
- **Continuation**: It is a big one.

Exp1. Existential vs. Universal (Cont.)

Donkey Conditionals

- The farmer owns a donkey, and he beats it. It is a big one.

- If the farmer owns a donkey, he beats it.#It is a big one.

- **Existential (∃)**: The farmer owns a donkey, and he beats it.
- **Conditional (∀)**: {If, Whenever} the farmer owns a donkey, he beats it.
- **Continuation**: It is a big one.

Exp1. Existential vs. Universal (Cont.)

Donkey Conditionals

- The farmer owns a donkey, and he beats it. It is a big one.

- If the farmer owns a donkey, he beats it.#It is a big one.

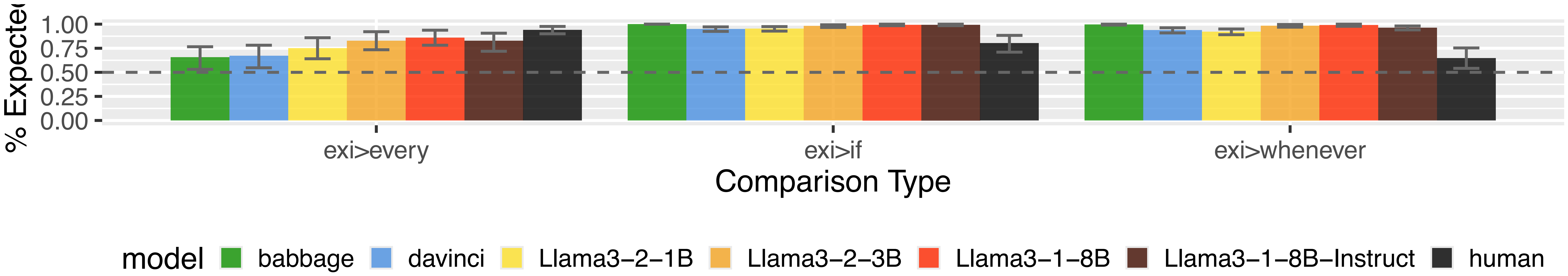
- **Existential (∃):** The farmer owns a donkey, and he beats it.
- **Conditional (∀):** {If, Whenever} the farmer owns a donkey, he beats it.
- **Continuation:** It is a big one.

$$p(cont \mid \exists) > p(cont \mid \forall)$$

Exp1. Existential vs. Universal (Cont.)

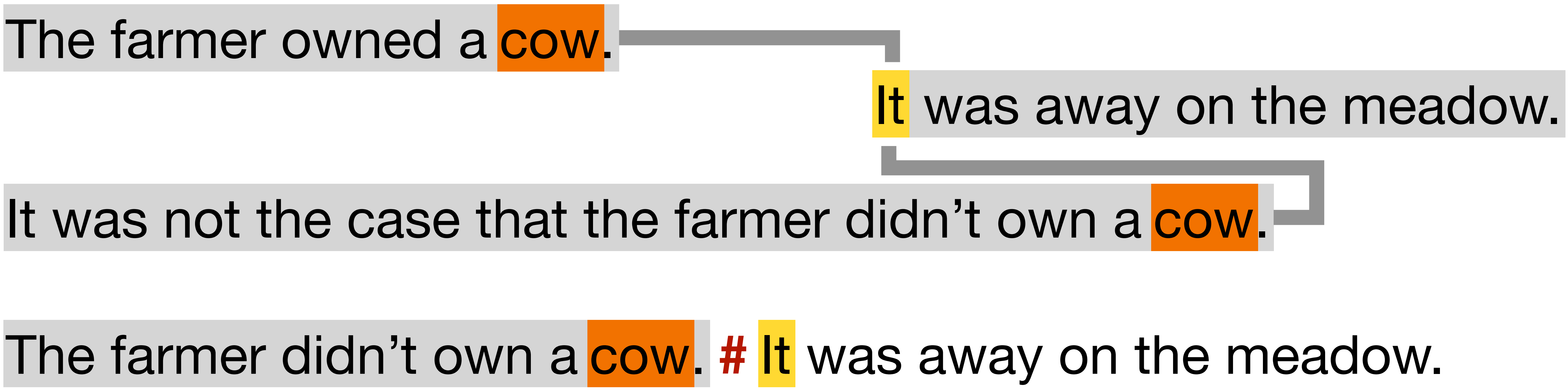
Results

- All models show above chance performance for the expected inequality.



- **Takeaway**: the LLMs examined know the scope of the discourse entity introduced within the universal quantifier and that it is infelicitous to refer back to such entities outside of the scope.

Exp2. Negation



The farmer owned a cow.

It was away on the meadow.

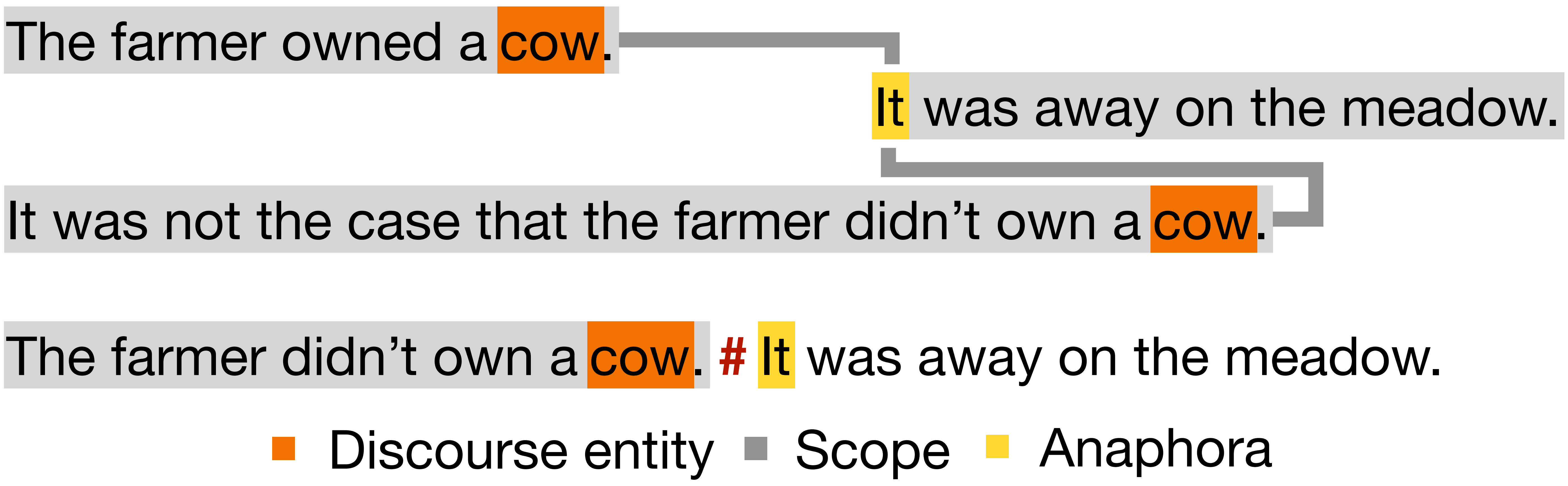
It was not the case that the farmer didn't own a cow.

The farmer didn't own a cow. # It was away on the meadow.

🟧 Discourse entity ■ Scope 🟨 Anaphora

- **Existential (∃)**: The farmer owned a cow.
- **Neg (¬)**: The farmer didn't own a cow.
- **DoubleNegation (DN):** It was not the case that the farmer didn't own a cow.
- **Continuation**: It was (just) away on the meadow.

Exp2. Negation



The farmer owned a cow.

It was away on the meadow.

It was not the case that the farmer didn’t own a cow.

The farmer didn’t own a cow. # It was away on the meadow.

🟧 Discourse entity ◼ Scope 🟨 Anaphora

- **EXISTENTIAL (∃)**: The farmer owned a cow.

$$p(Cont\,|\,\exists) > p(Cont\,|\,\neg)$$

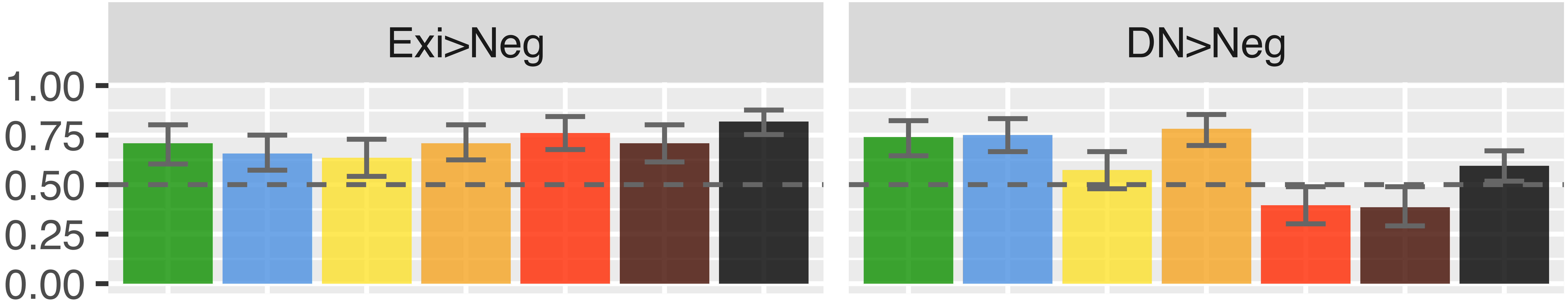
- **NEG (¬)**: The farmer didn’t own a cow.

$$p(Cont\,|\,DN) > p(Cont\,|\,\neg)$$

- **DOUBLENEGATION (DN):** It was not the case that the farmer didn’t own a cow.
- **CONTINUATION**: It was (just) away on the meadow.

Exp2. Negation (Cont.)

Results



- All models succeed in Exi > Neg; three models struggle with DN > Neg.

Exp2. Negation (Cont.)

Lexical

Exp2. Negation (Cont.)

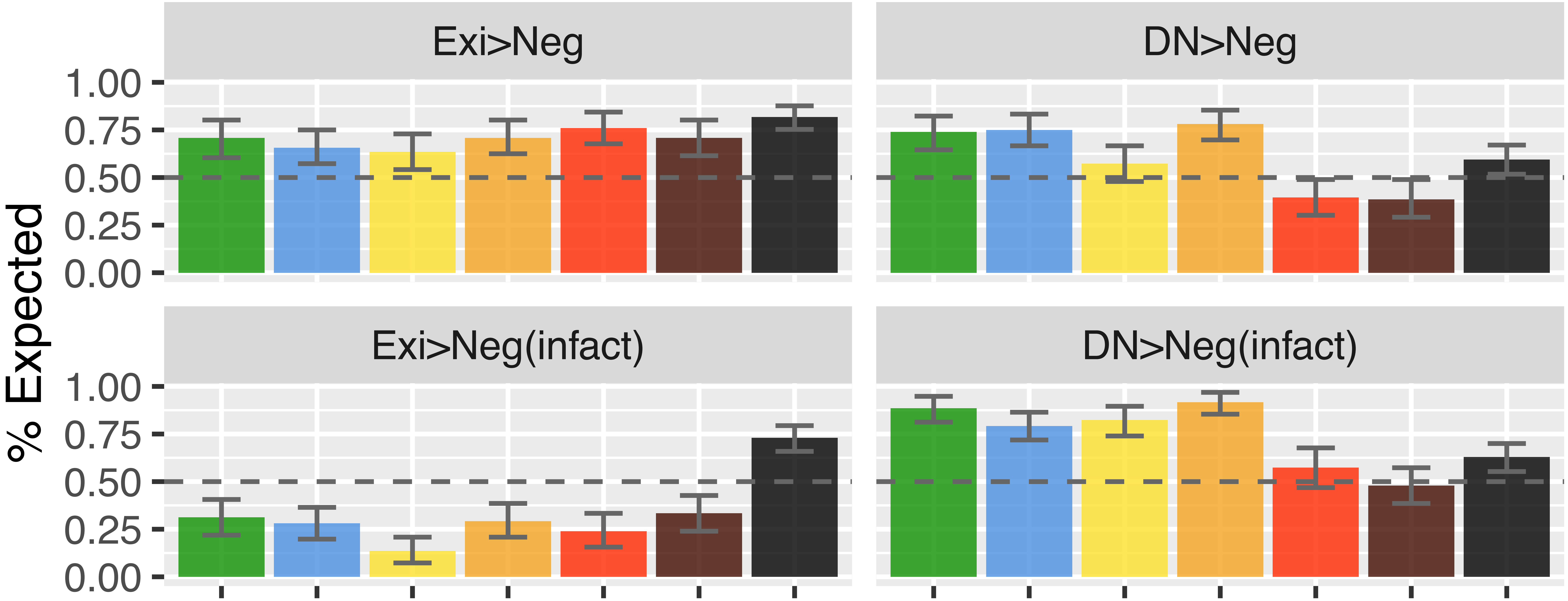
Lexical

- **CONTINUATION**: **In fact,** it was (just) away on the meadow.

Exp2. Negation (Cont.)

Lexical

- **CONTINUATION**: **In fact,** it was (just) away on the meadow.



Exp3. Disjunction vs. Conjunction

With negation, disjunction is felicitous, while conjunction is not.

- (∨): Either there was no manuscript, **or** it was hidden by the librarian.

Exp3. Disjunction vs. Conjunction

With negation, disjunction is felicitous, while conjunction is not.

- ($\vee$): Either there was no manuscript, **or** it was hidden by the librarian.

(Either) There was no manuscript

Exp3. Disjunction vs. Conjunction

With negation, disjunction is felicitous, while conjunction is not.

- (∨): Either there was no manuscript, **or** it was hidden by the librarian.

disjunction
either… or…

(Either) There was no manuscript

Exp3. Disjunction vs. Conjunction

With negation, disjunction is felicitous, while conjunction is not.

- (∨): Either there was no manuscript, **or** it was hidden by the librarian.

disjunction
either… or…

(Either) There was no manuscript

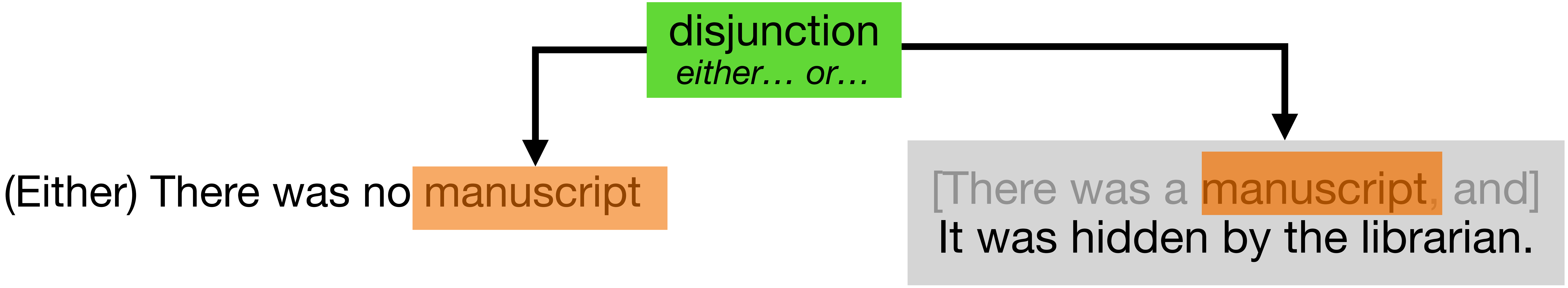
[There was a manuscript, and]
It was hidden by the librarian.

■ **Quantifier** scope

Exp3. Disjunction vs. Conjunction

With negation, disjunction is felicitous, while conjunction is not.

- (⋁): Either there was no manuscript, **or** it was hidden by the librarian.



disjunction
either… or…

(Either) There was no manuscript

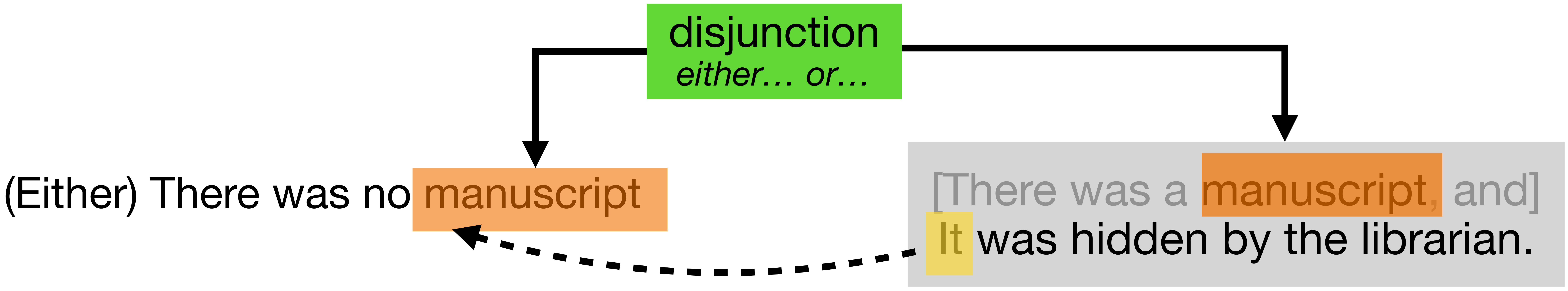
[There was a manuscript, and]
It was hidden by the librarian.

🟧 Discourse entity ■ **Quantifier** scope

Exp3. Disjunction vs. Conjunction

With negation, disjunction is felicitous, while conjunction is not.

- (∨): Either there was no manuscript, **or** it was hidden by the librarian.



disjunction

either… or…

(Either) There was no manuscript

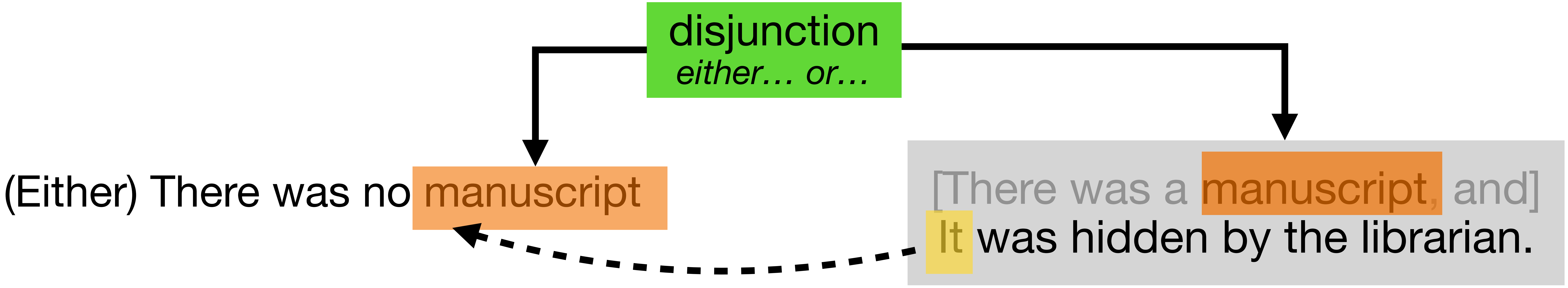
[There was a manuscript, and] It was hidden by the librarian.

🟧 Discourse entity ◼ **Quantifier** scope 🟨 Anaphora

Exp3. Disjunction vs. Conjunction

With negation, disjunction is felicitous, while conjunction is not.

- (∨): Either there was no manuscript, **or** it was hidden by the librarian.



disjunction
either… or…

(Either) There was no manuscript

[There was a manuscript, and] It was hidden by the librarian.

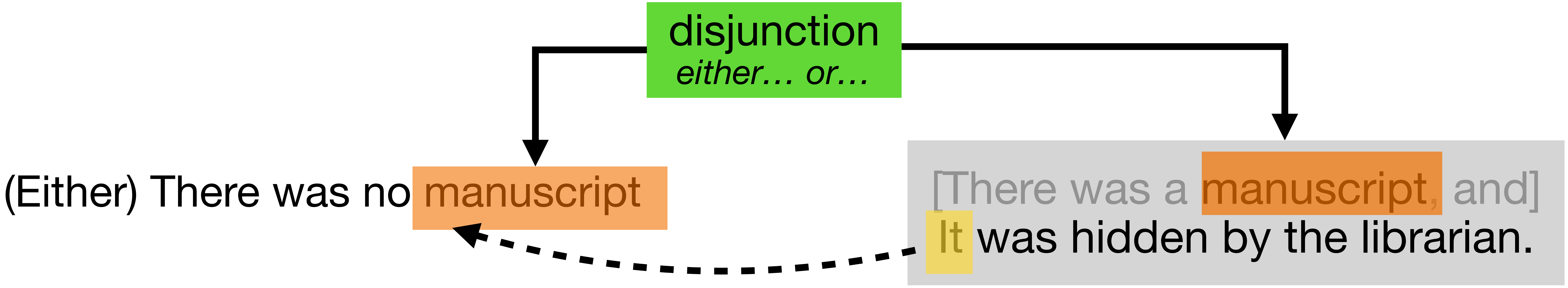
not possible for conjunction…

🟧 Discourse entity ■ **Quantifier** scope 🟨 Anaphora

Exp3. Disjunction vs. Conjunction

With negation, disjunction is felicitous, while conjunction is not.

- (∨): Either there was no manuscript, **or** it was hidden by the librarian.



disjunction
either… or…

(Either) There was no manuscript

[There was a manuscript, and]
It was hidden by the librarian.

not possible for conjunction…

🟧 Discourse entity　▪ **Quantifier** scope　🟨 Anaphora

- (∧): # There was no manuscript, **and** it was hidden by the librarian.

Exp3. Disjunction vs. Conjunction Cont.

Conditions and Predictions

Exp3. Disjunction vs. Conjunction Cont.

Conditions and Predictions

- Conditions
  - **EitherOr:** Either there was no manuscript, **or** it was hidden by the librarian.
  - **EitherPosOr**: # Either there was a manuscript, **or** it was hidden by the librarian.
  - **Conjunction**: # There was no manuscript, **and** it was hidden by the librarian.

Exp3. Disjunction vs. Conjunction Cont.

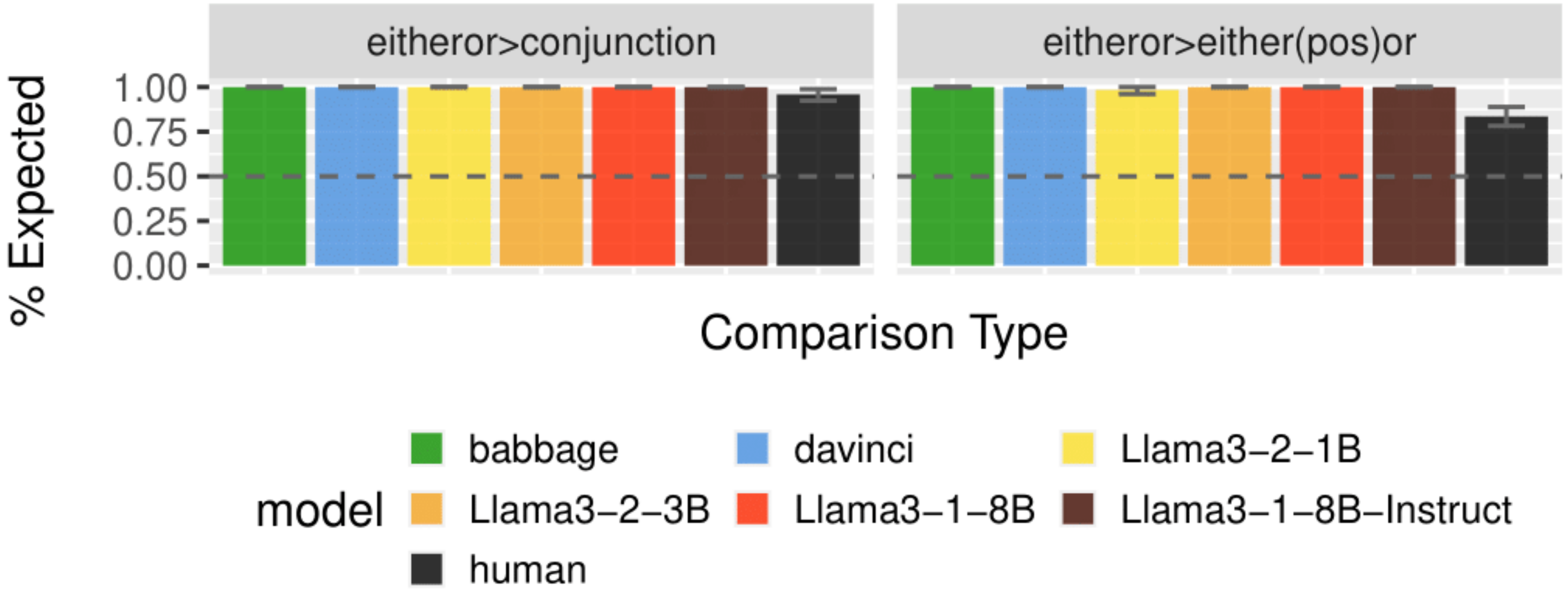
Conditions and Predictions

- Conditions
  - **EITHEROR**: Either there was no manuscript, **or** it was hidden by the librarian.
  - **EITHERPOSOR**: # Either there was a manuscript, **or** it was hidden by the librarian.
  - **CONJUNCTION**: # There was no manuscript, **and** it was hidden by the librarian.
- Predictions:
  - SLOR(**EITHEROR**) > SLOR(**CONJUNCTION**)
  - SLOR(**EITHEROR**) > SLOR(**EITHERPOSOR**)

* Syntactic Log-Odds Ratio, $SLOR(s) = \frac{\log p_m(s) - \sum_{w \in s} \log p_u(w)}{|s|}$, is a metric on sentence well-formedness (Lau et al. 2017).

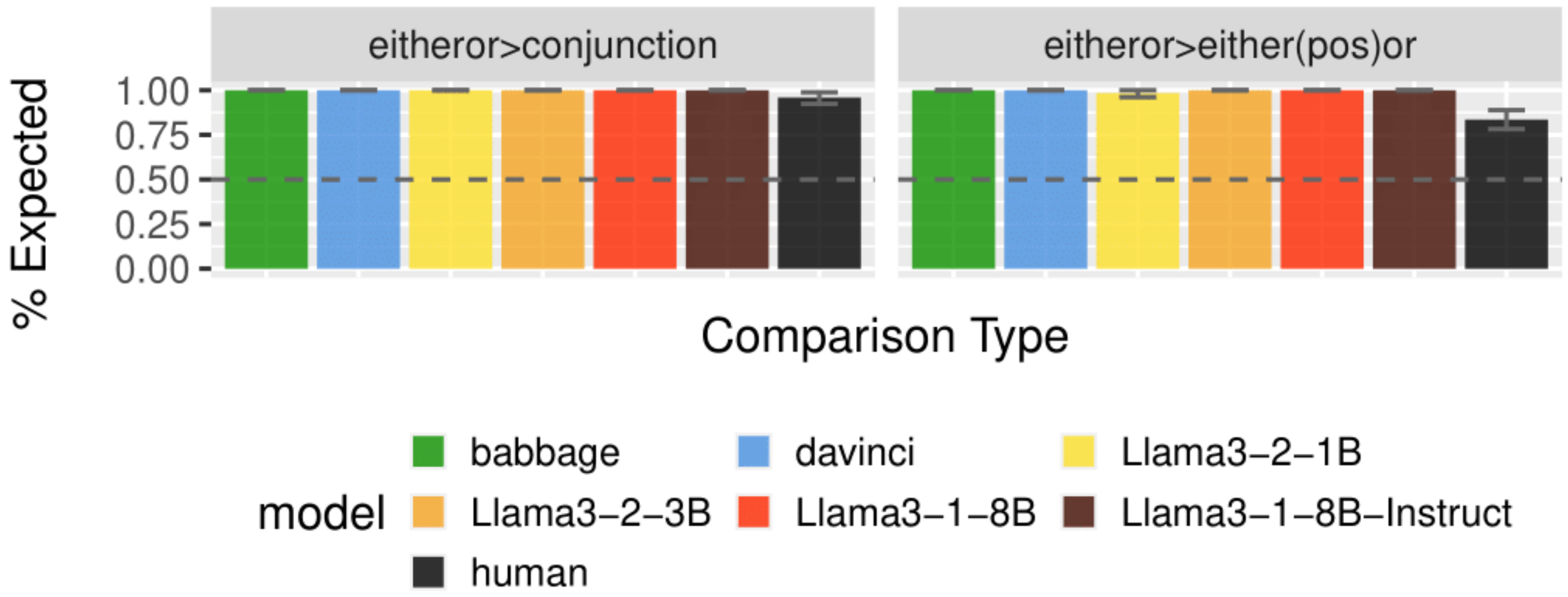
Exp3. Disjunction vs. Conjunction Cont.

Results



Exp3. Disjunction vs. Conjunction Cont.

Results



- All models robustly favored the felicitous disjunction sentences over the infelicitous ones.

Exp3. Disjunction vs. Conjunction Cont.

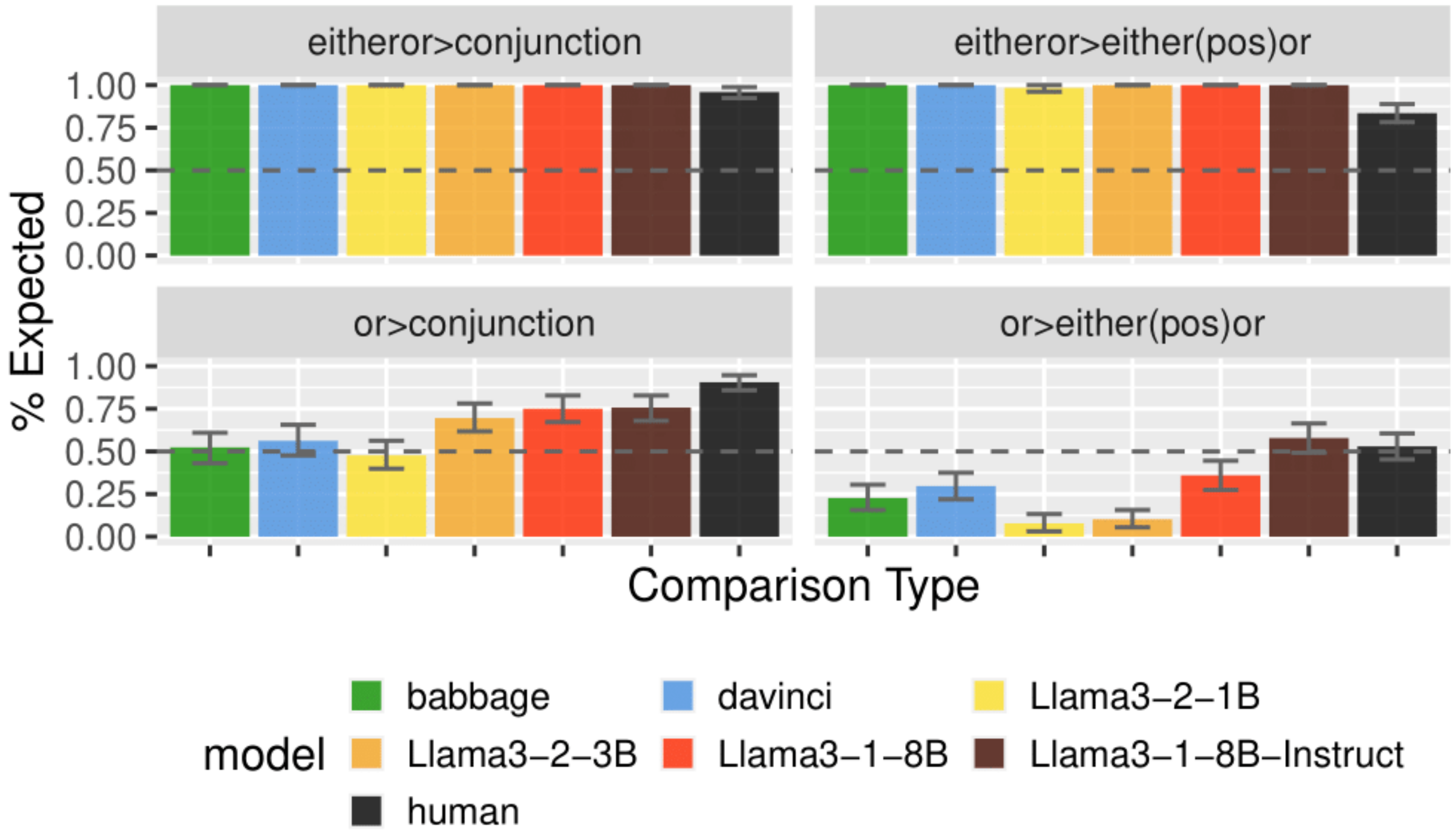
Sensitivity to Lexical Impacts

- What if we get rid of “either”, and only use “or” for disjunction?
  - **OR**: ~~Either~~ There was no manuscript, **or** it was hidden by the librarian.

Exp3. Disjunction vs. Conjunction Cont.

Sensitivity to Lexical Impacts

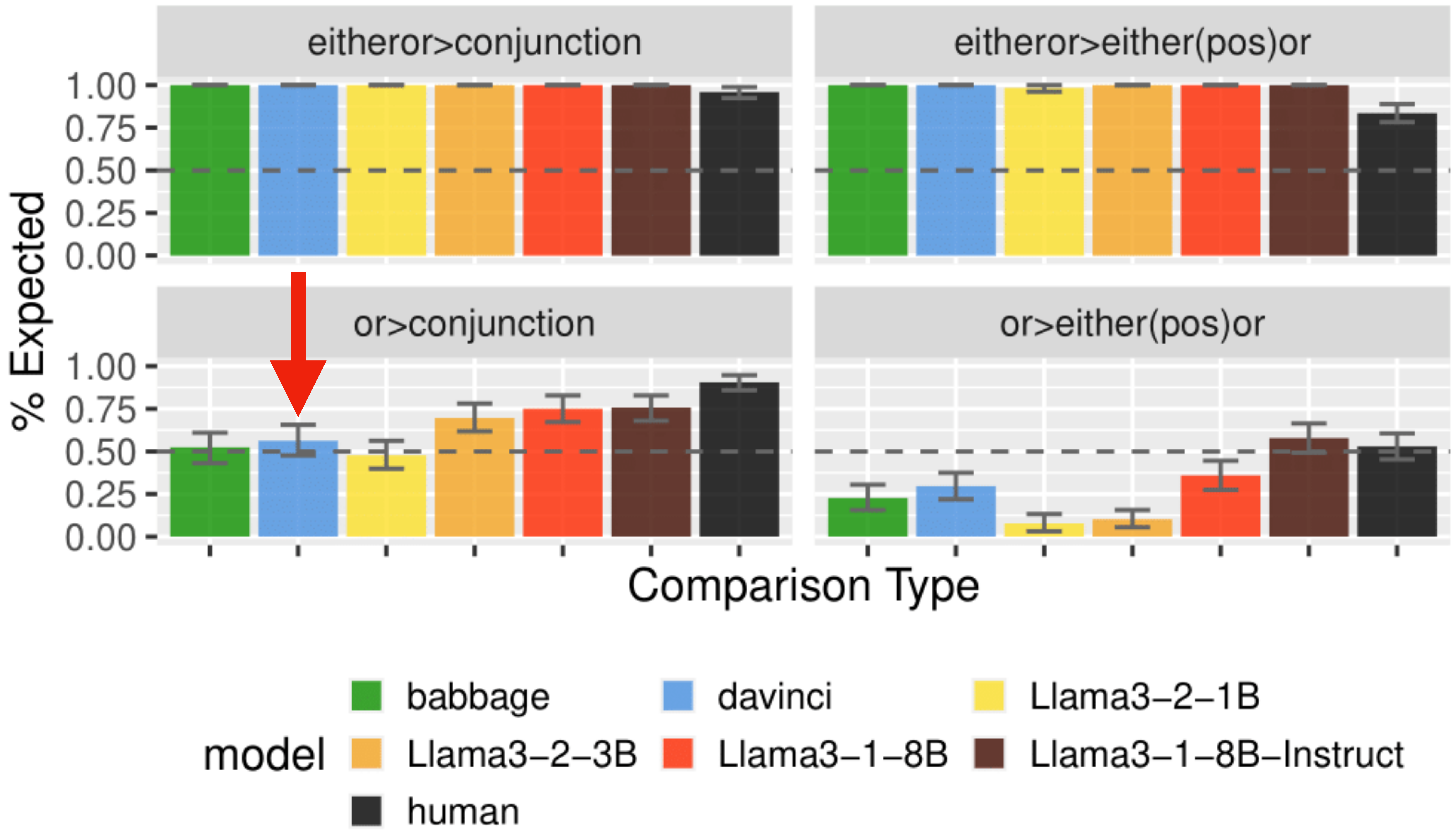
- What if we get rid of “either”, and only use “or” for disjunction?
  - **Or**: ~~Either~~ There was no manuscript, **or** it was hidden by the librarian.



Exp3. Disjunction vs. Conjunction Cont.

Sensitivity to Lexical Impacts

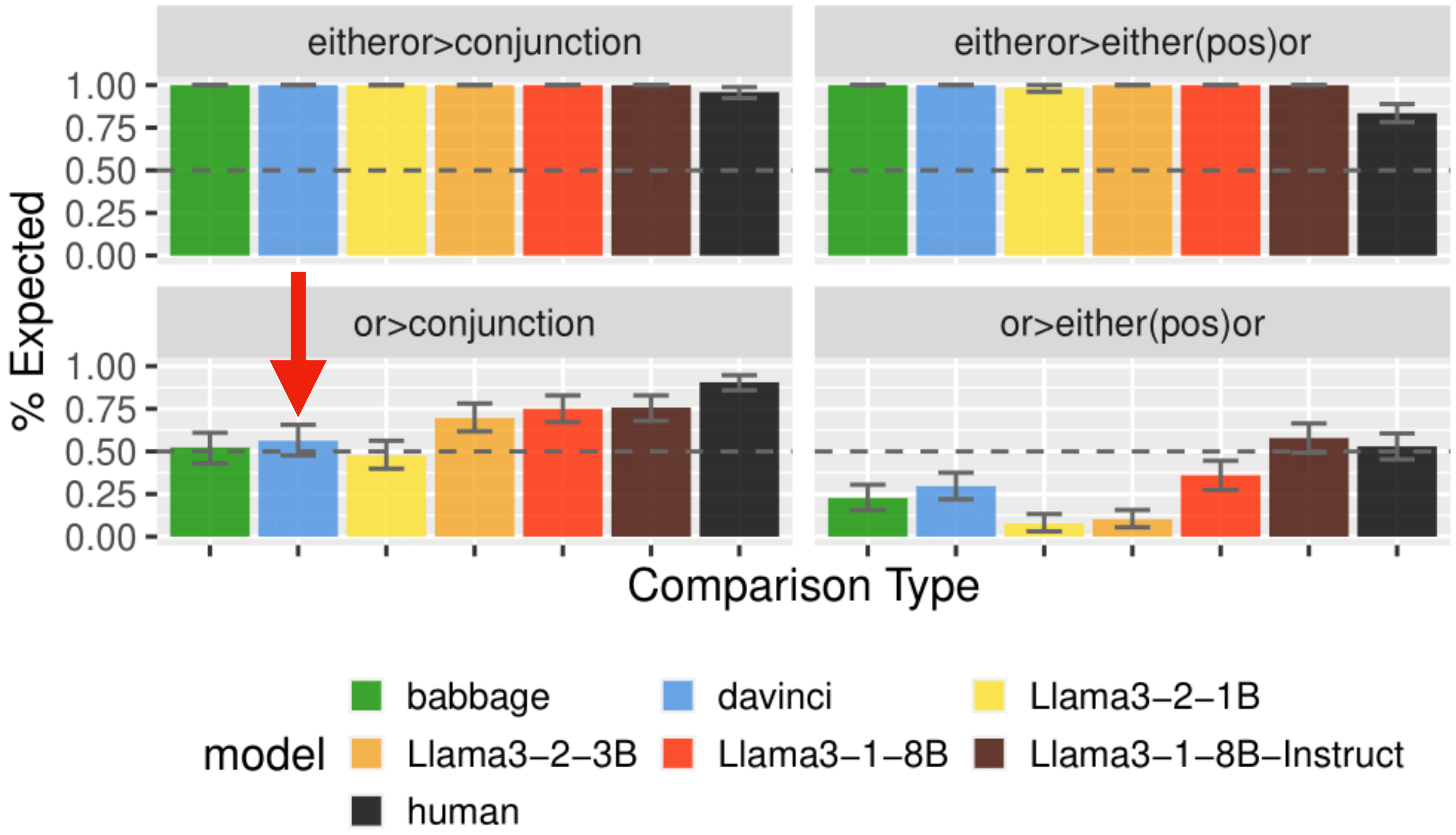
- What if we get rid of “either”, and only use “or” for disjunction?
    - **OR**: ~~Either~~ There was no manuscript, **or** it was hidden by the librarian.



Exp3. Disjunction vs. Conjunction Cont.

Sensitivity to Lexical Impacts

- What if we get rid of “either”, and only use “or” for disjunction?
    - **OR**: ~~Either~~ There was no manuscript, **or** it was hidden by the librarian.

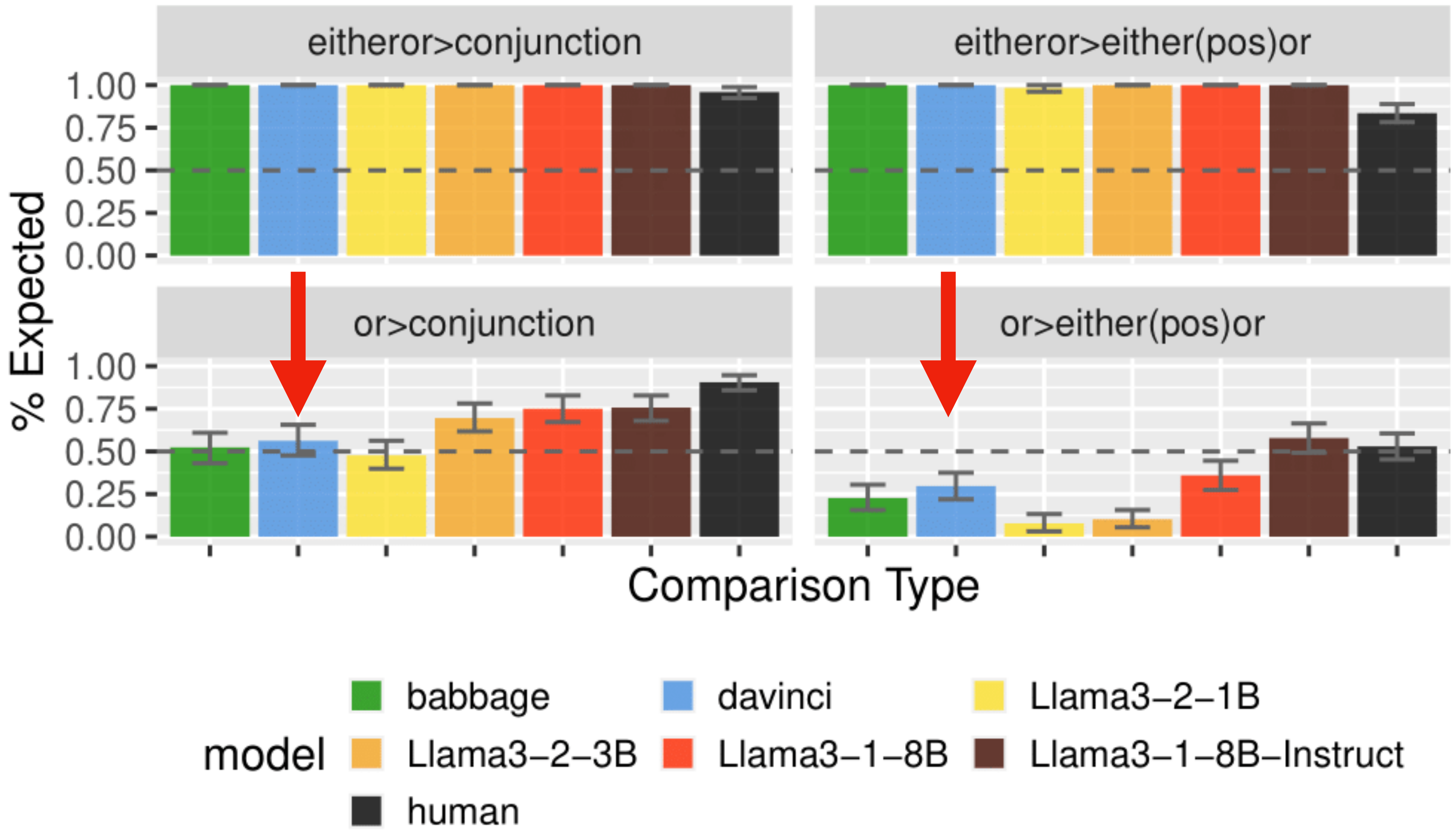


- For almost all models, accuracies dropped to or below chance-level.

Exp3. Disjunction vs. Conjunction Cont.

Sensitivity to Lexical Impacts

- What if we get rid of “either”, and only use “or” for disjunction?
  - **OR**: ~~Either~~ There was no manuscript, **or** it was hidden by the librarian.

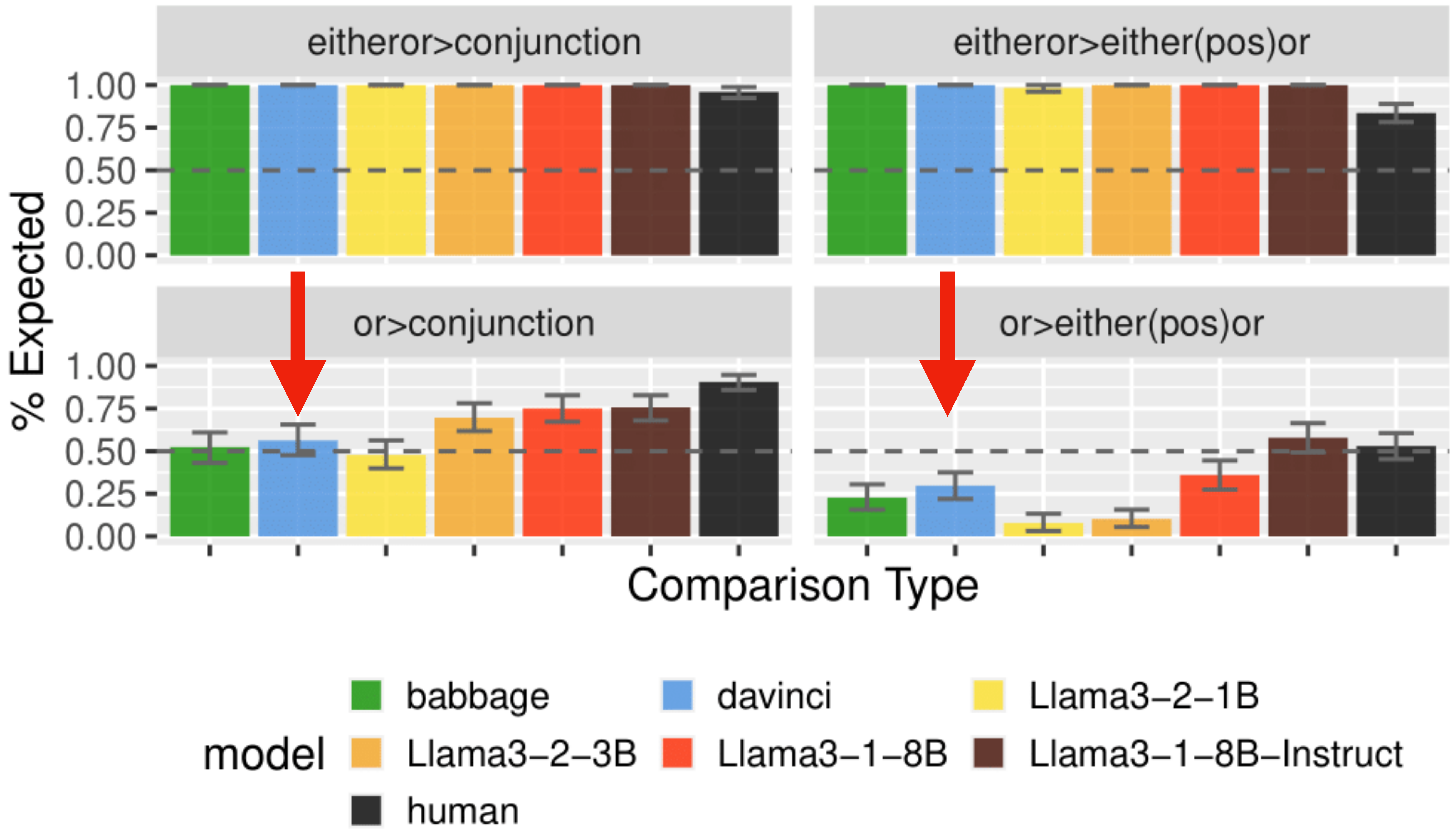


- For almost all models, accuracies dropped to or below chance-level.

Exp3. Disjunction vs. Conjunction Cont.

Sensitivity to Lexical Impacts

- What if we get rid of “either”, and only use “or” for disjunction?
  - **OR**: ~~Either~~ There was <u>no</u> manuscript, **or** it was hidden by the librarian.

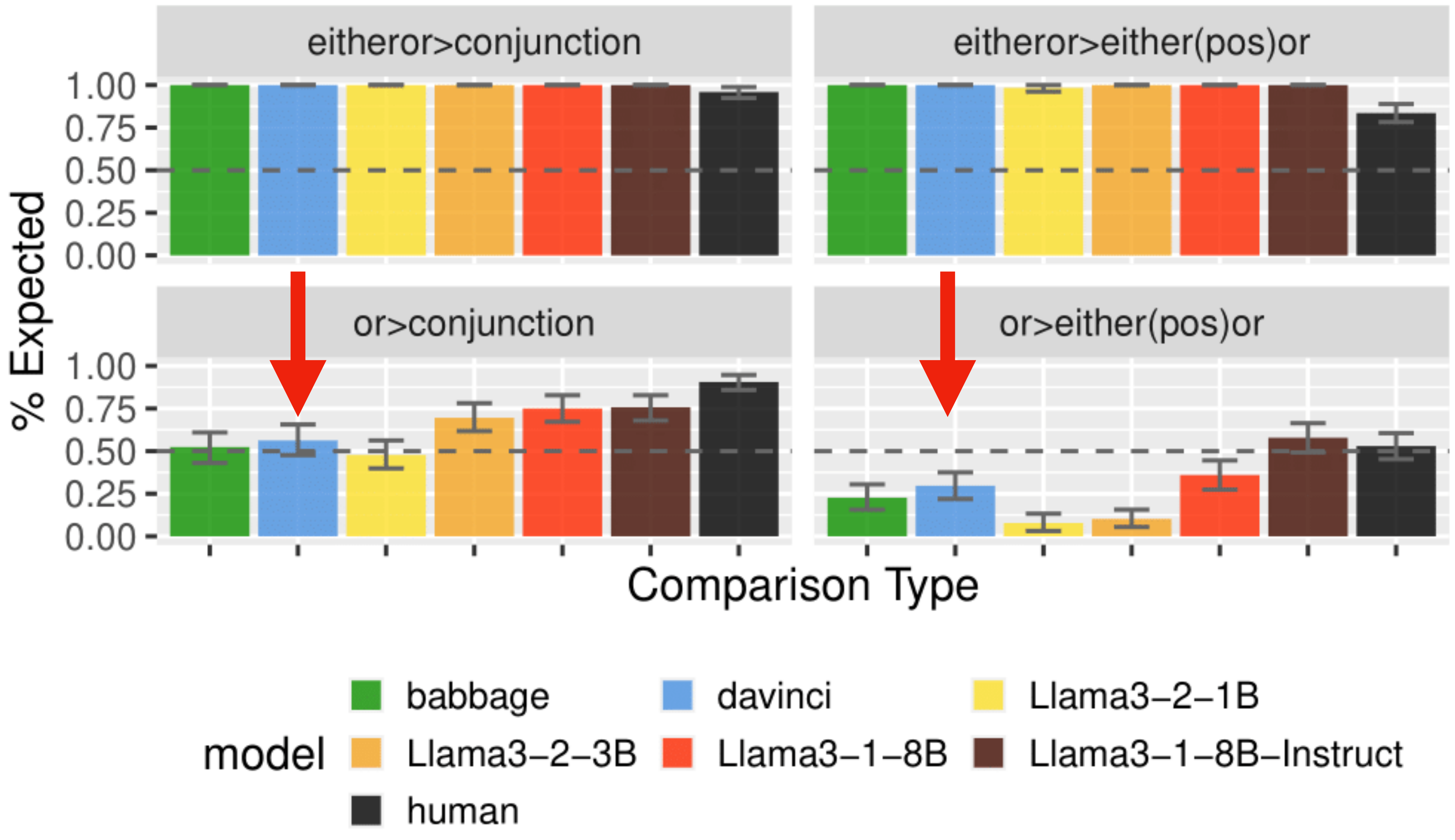


- For almost all models, accuracies dropped to or below chance-level.
- Human preference stayed robust in Or>Conjunction but dropped to chance for Or>Either(Pos)Or.

Exp3. Disjunction vs. Conjunction Cont.

Sensitivity to Lexical Impacts

- What if we get rid of “either”, and only use “or” for disjunction?
  - **Or**: ~~Either~~ There was no manuscript, **or** it was hidden by the librarian.



- For almost all models, accuracies dropped to or below chance-level.
- Human preference stayed robust in Or>Conjunction but dropped to chance for Or>Either(Pos)Or.
- Another example of LMs’ lexical-sensitivity modulating anaphora accessibility.

Conclusion

Conclusion

- We filled the gap of evaluating LLM natural language understanding abilities at the discourse level, motivated by the dynamic semantics framework.

Conclusion

- We filled the gap of evaluating LLM natural language understanding abilities at the discourse level, motivated by the dynamic semantics framework.
- We constructed a hand-curated dataset focusing on anaphora accessibility, and we used it to evaluate the discourse / entity tracking ability with natural language sentences.

Conclusion

- We filled the gap of evaluating LLM natural language understanding abilities at the discourse level, motivated by the dynamic semantics framework.
- We constructed a hand-curated dataset focusing on anaphora accessibility, and we used it to evaluate the discourse / entity tracking ability with natural language sentences.
- We found places of both convergence and divergence between LLMs and human performance, where LLMs rely on specifical lexical cues but humans don’t.

Thank you for listening!

Acknowledgment

We thank the anonymous reviewers from the ARR Feb 2025 cycles and SCiL 2025. We also thank the Yale Linguistics Deparment, especially the members of the Computational Linguistics at Yale Lab, for helpful feedback.

Paper Link:

ACL 2025
VIENNA