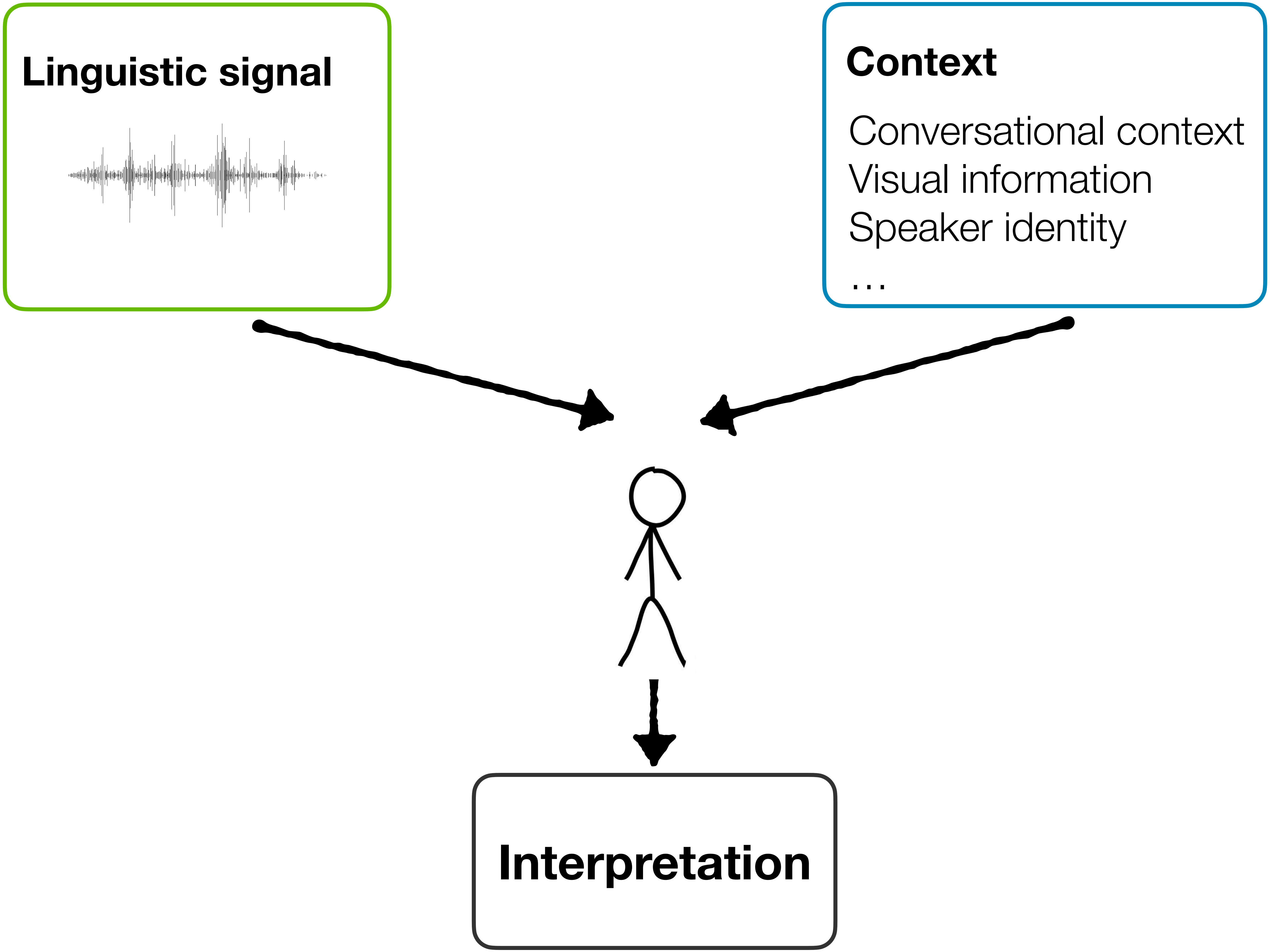
How contextual are contextual language models?

NYU Text-as-Data Series
March 31, 2022

Sebastian Schuster


Linguistic signal
Context
Conversational context
Visual information
Speaker identity
…
Interpretation

A: What on earth happened to the roast beef?

B: The dog is looking very happy

⇝ The dog likely ate the roast beef

Levinson (1983)

Linguistic signal | **Context** | **World knowledge**

Partitive constructions make scalar inferences more likely

Joe ate *some* cookies.

Joe ate *some* **of the** cookies.

e.g., Horn (1997) , Degen (2015)

Supportive or unsupportive contexts for presuppositions

Chet never became a lawyer, he didn't *finish law school*.

—> Chet went to law school.

Chet just finished med school, he didn't *finish law school*.

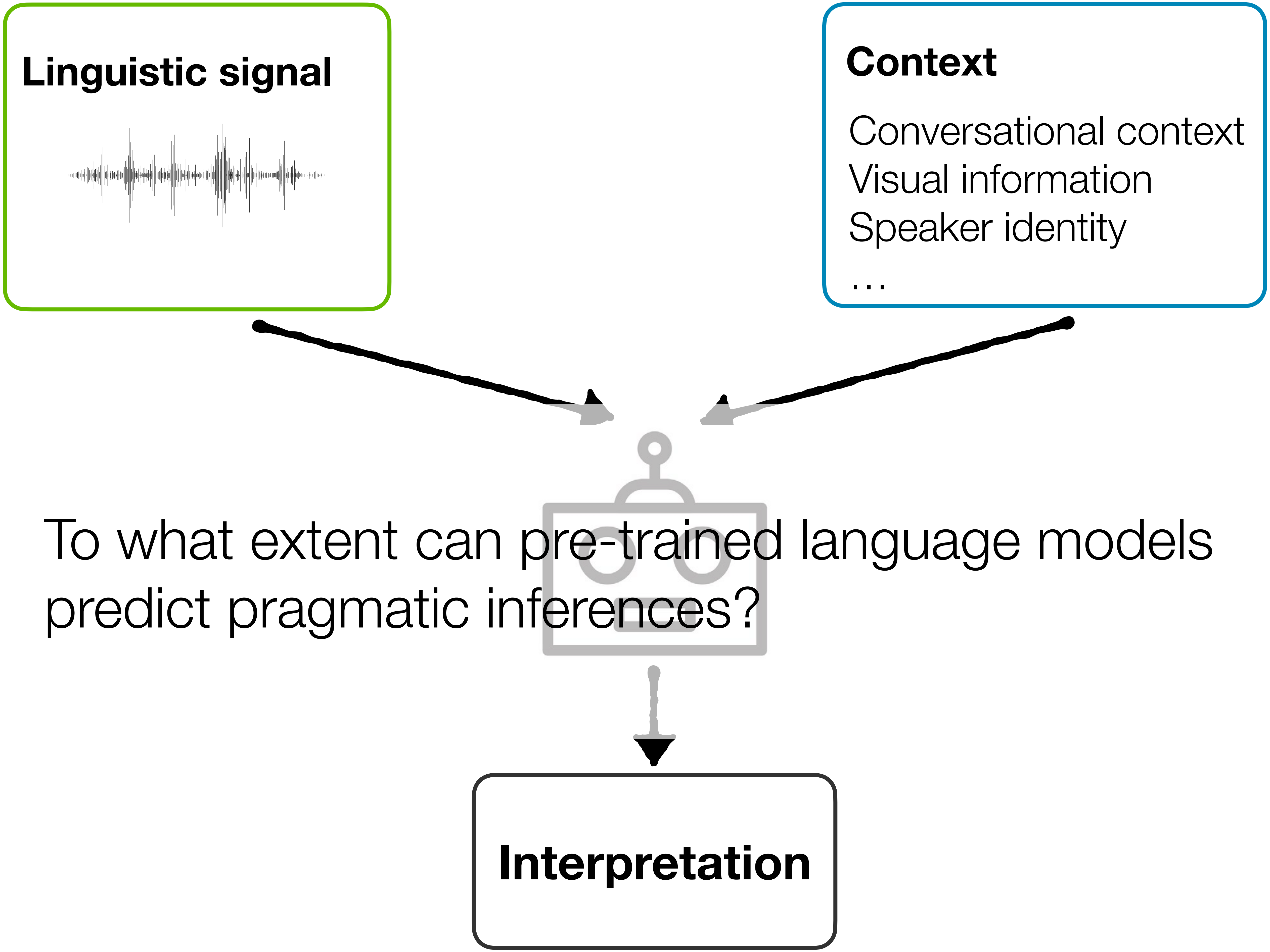
-/-> Chet went to law school.

Indefinite noun phrases embedded under positive implicatives are more likely to introduce discourse entities

Sue **managed** to find *a marble*.

Sue **failed** to find *a marble*.

e.g., Karttunen (1976)



To what extent can pre-trained language models predict pragmatic inferences?

Plan for today

1. To what extent can BERT learn to predict context-sensitive inferences from **“*some*” to “*some but not all*”**? {Schuster, Chen}, and Degen, 2020

2. To what extent can NLI models based on RoBERTa/DeBERTa predict **presuppositions**? {Parrish, Schuster, Warstadt}, et al., 2021

3. To what extent can GPT-2 and GPT-3 track **discourse entities**? Schuster and Linzen, under review

Plan for today

1. To what extent can BERT learn to predict context-sensitive inferences from **"*some*" to "*some but not all*"**? {Schuster, Chen}, and Degen, 2020

2. To what extent can NLI models based on RoBERTa/DeBERTa predict **presuppositions**? {Parrish, Schuster, Warstadt}, et al., 2021

3. To what extent can GPT-2 and GPT-3 track **discourse entities**? Schuster and Linzen, under review

Scalar inferences with *some*

I ate ***some*** of the cookies

✓

✗

Scalar inferences with *some*

I ate ***some*** of the cookies
all

Grice, 1975; Horn, 1987; Hirschberg, 1989

Scalar inferences with *some*

I ate ***some*** of the cookies

~~***all***~~

Grice, 1975; Horn, 1987; Hirschberg, 1989

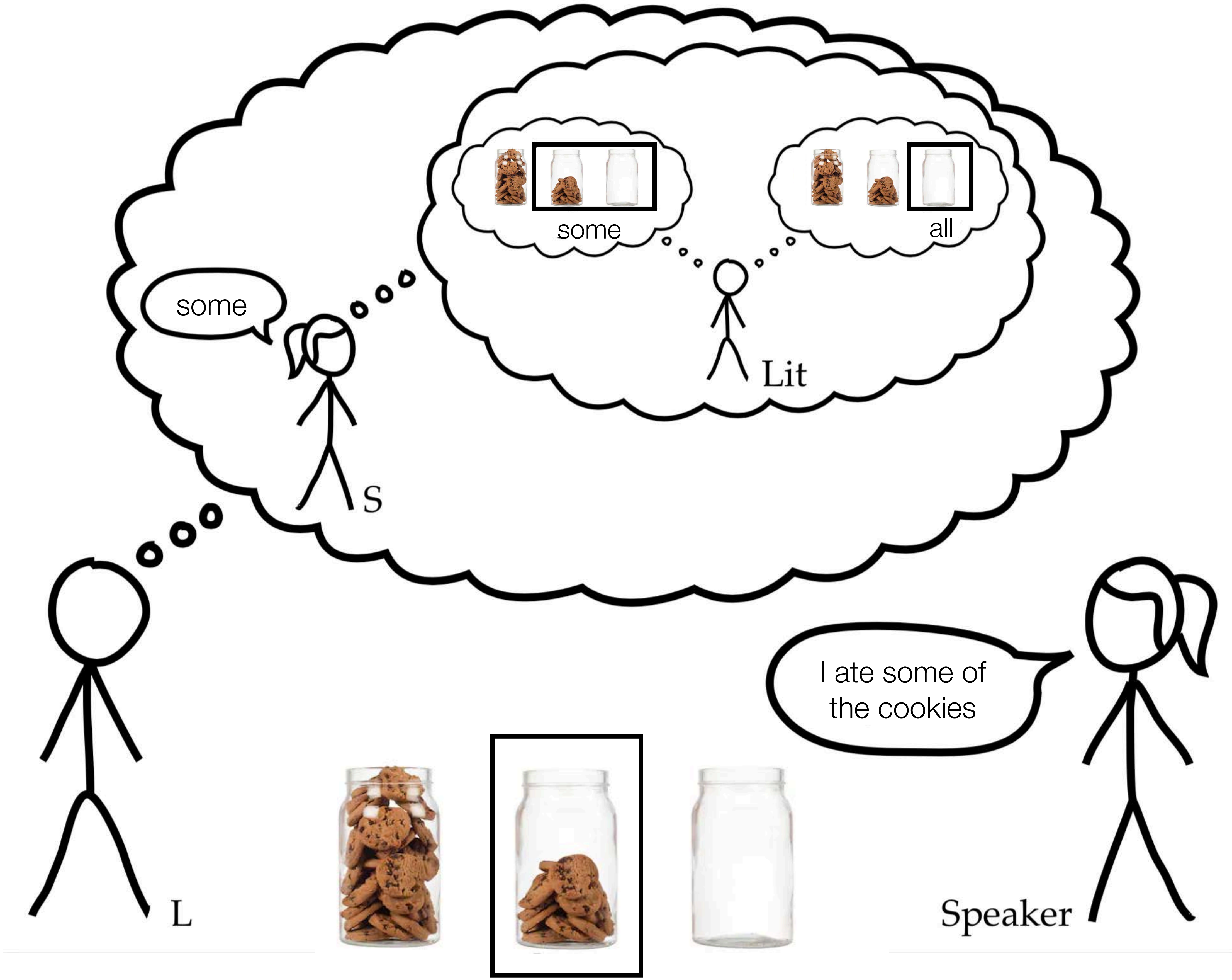
Scalar inferences with *some*

I ate ***some*** of the cookies

~~***all***~~

I ate **some but not all** of the cookies

Grice, 1975; Horn, 1987; Hirschberg, 1989



Goodman & Frank, 2016; Franke & Jäger, 2016

Rational Speech Act Framework

- **It does not scale**
  The model requires a pre-defined **set of possible utterances** and their mapping to a truth-conditional semantics

Contextual variation in scalar implicatures

scalar inference strength

I like **some country music**.

Intended inference? I like some, but not all, country music

It would certainly help them to appreciate **some of the things we have here**.

Intended inference? …to appreciate some, but not all…

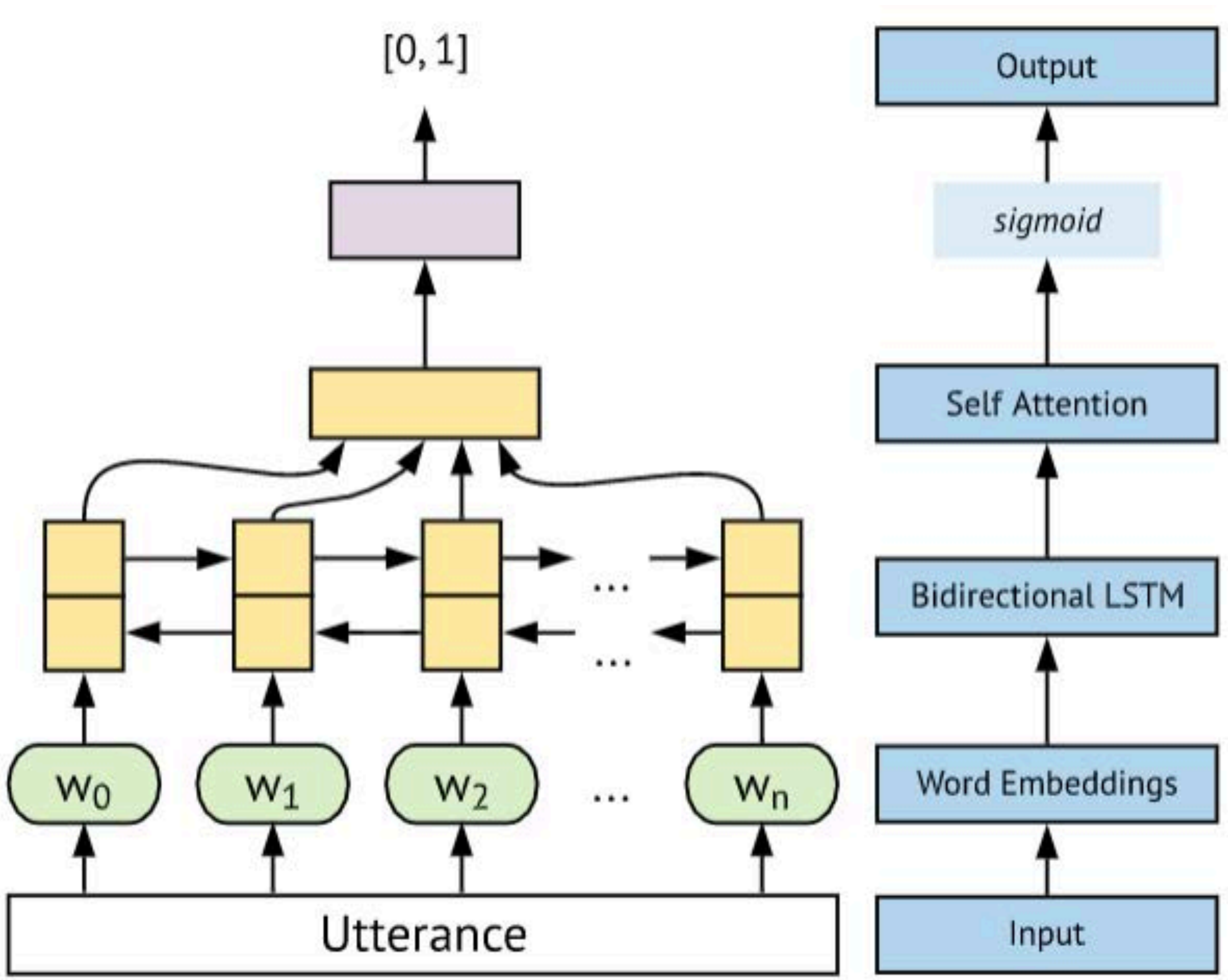
You sound like you have **some small ones** in the background.

Intended inference? … some, but not all small ones…

to what extent can neural network sentence encoders learn to predict scalar inference strength?

Schuster, Chen, and Degen (2020)

Neural sentence encoders



i like some country music

~= Lin et al. (2017)

Data

1,390 sentences containing *some* from the Switchboard corpus of spoken American English

Degen (2015)

Corpus study

Speaker A: i mean, they just have beautiful, beautiful homes and they have everything. the kids only wear name brand things to school and it's one of these things,

Speaker B: oh me. well that makes it hard for you, doesn't it.

Speaker A: well it does, you know. it really does because i'm a single mom and i have a thirteen year old now and uh, you know, it does.

Speaker B: oh, me.

Speaker A: i mean, we do it to a point but uh, not to where she feels different ,

Speaker B: yeah.

Speaker A:
but some of them are very rich

but **some, but not all** of them are very rich

How similar is the statement with 'some, but not all' (green) to the statement with 'some' (red)?

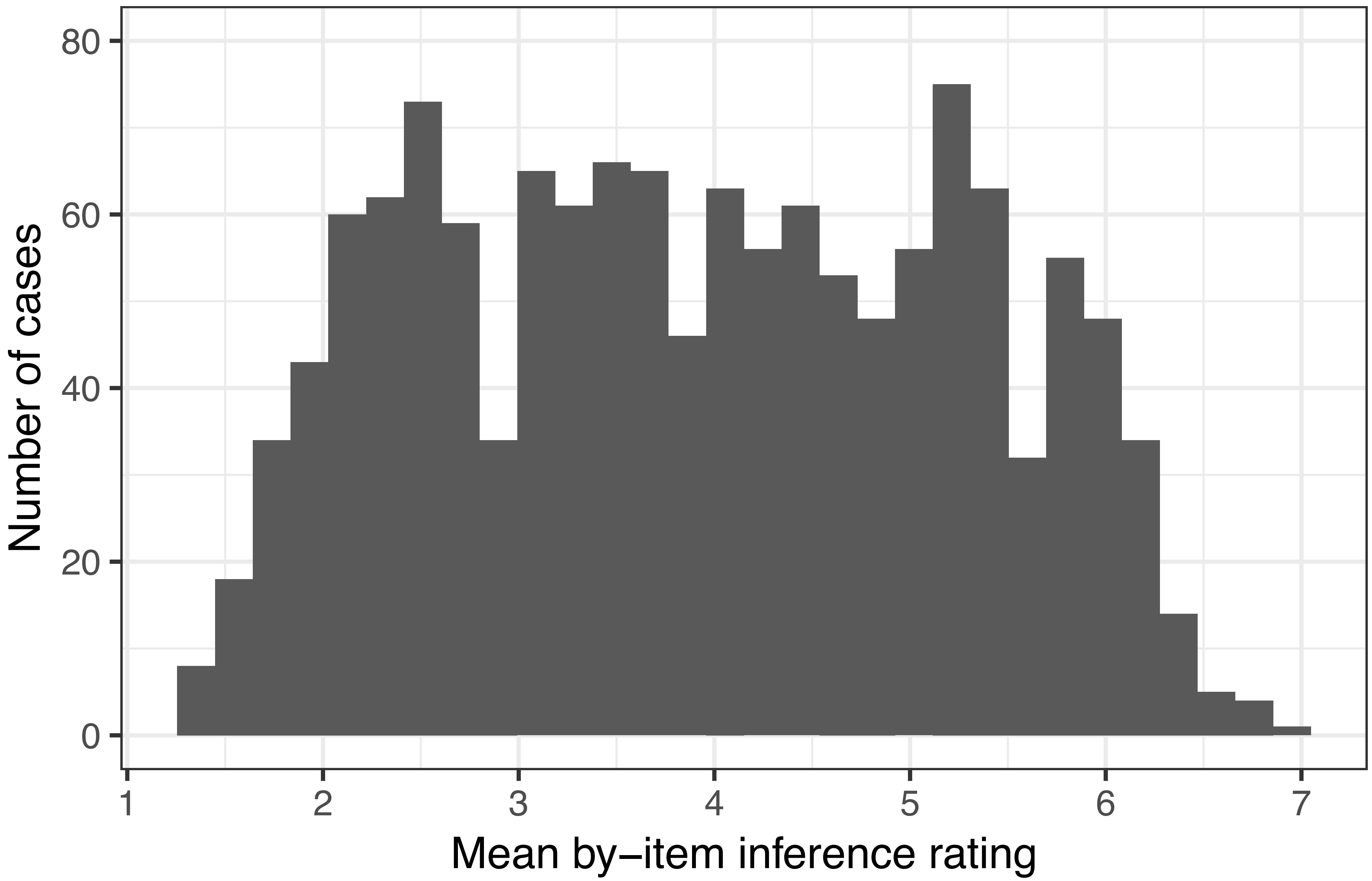
Very different meaning ○ ○ ○ ○ ○ ○ ○ Same meaning

1 2 3 4 5 6 7

Continue

Degen (2015)

Results

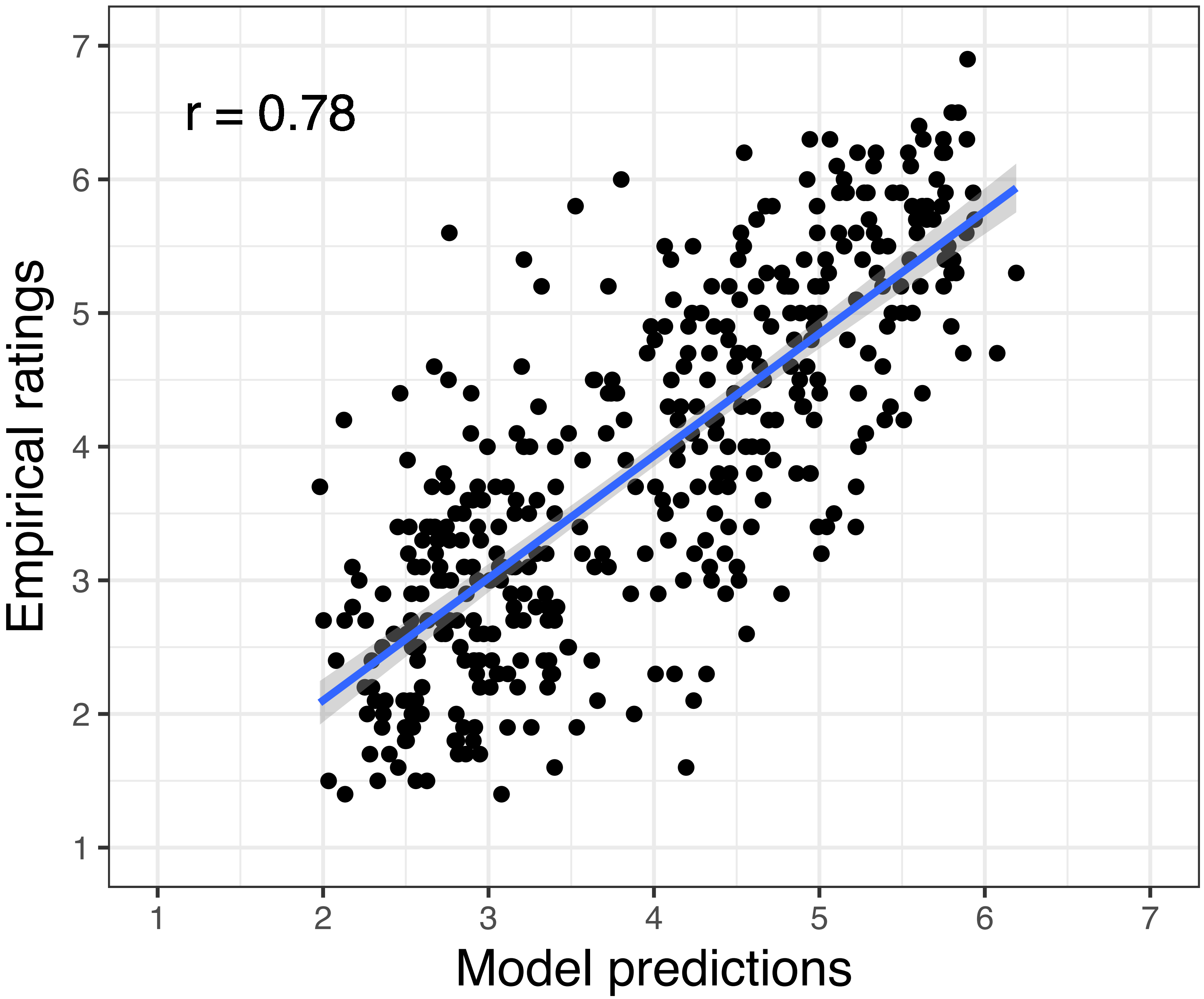


Degen (2015)

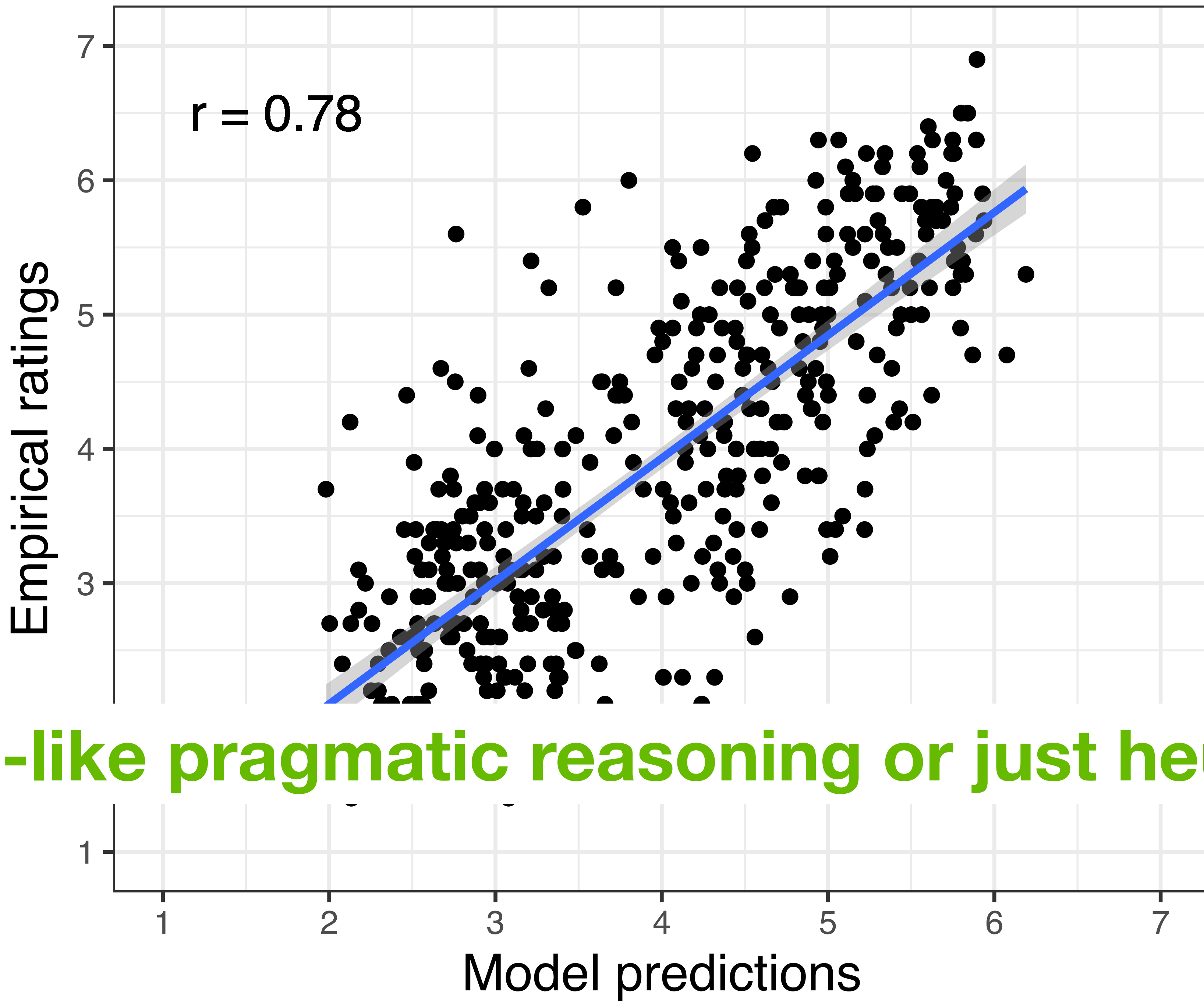
Why might neural language models exhibit pragmatic behavior?

- Pre-trained neural language models predict a lot of **complex human behavior at the level of syntax**:
  - long-distance **subject-verb agreement**: e.g., Goldberg, 2019; Warstadt et. al, 2020
  - **filler-gap dependencies**: e.g., Da Costa and Chaves, 2020
  - structurally sensitive syntactic **transformations**: e.g., Warstadt et al., 2020; Mueller et al., 2022
- Models are trained on **naturalistic texts** that were written by humans, i.e., **pragmatic agents**.

Held-out test set predictions

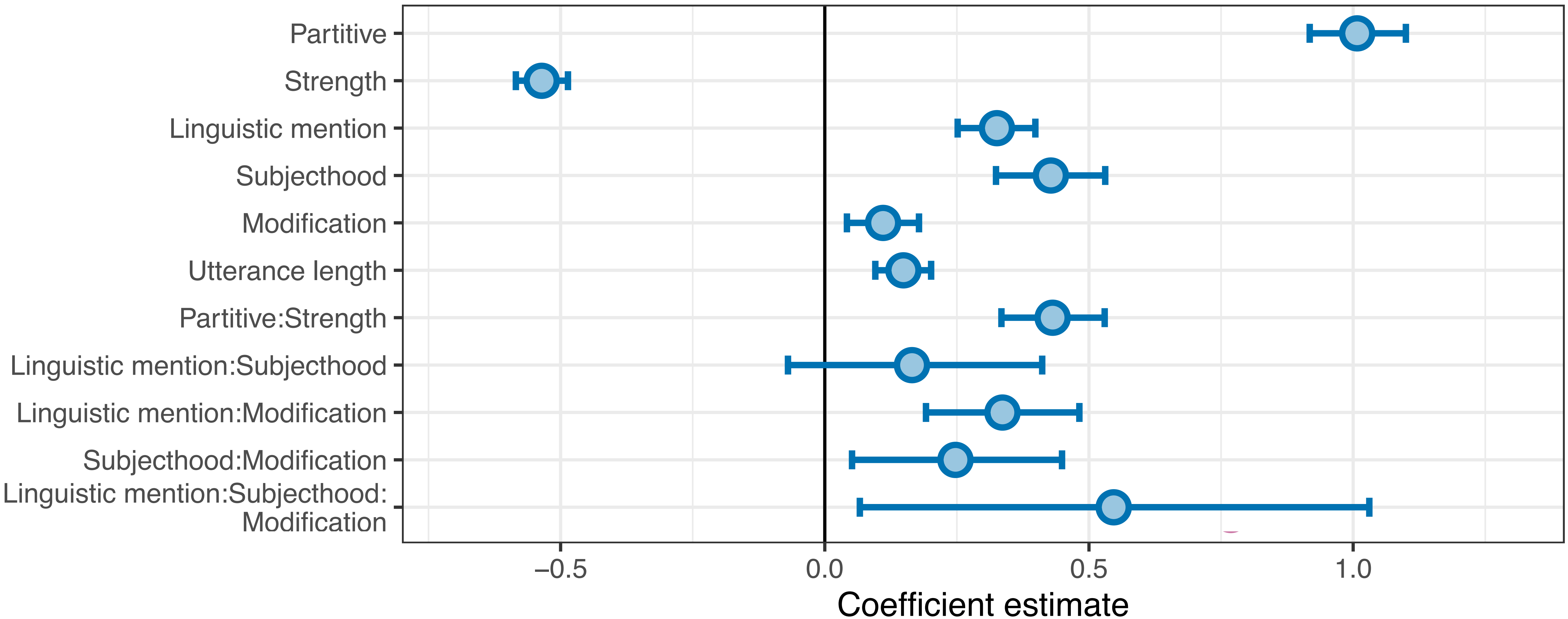


Held-out test set predictions



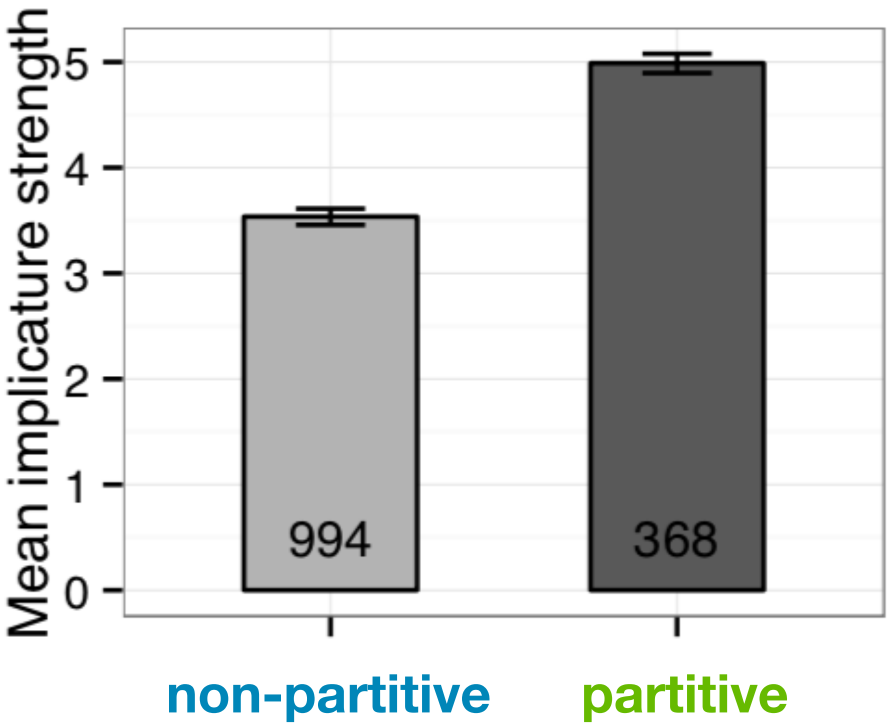
Human-like pragmatic reasoning or just heuristics?

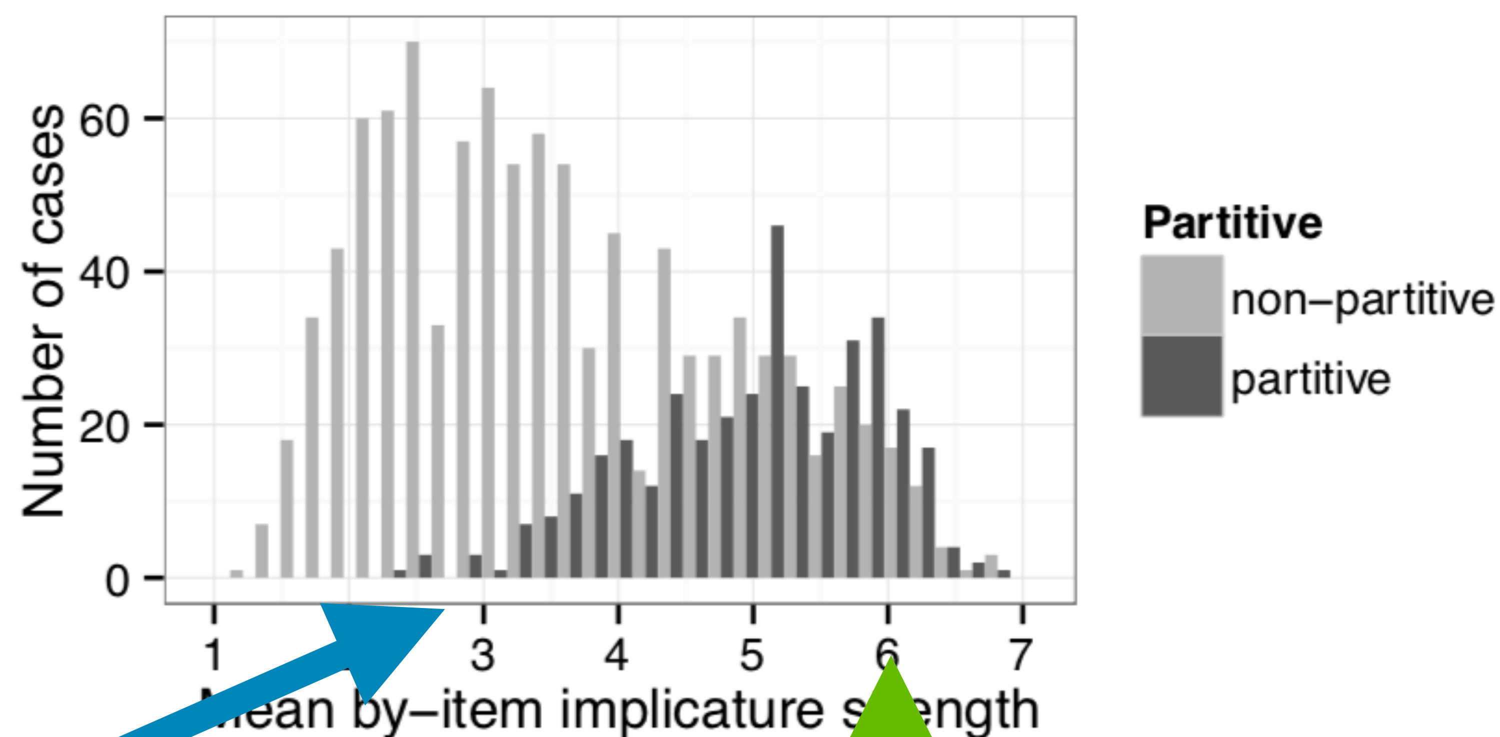
Features influencing pragmatic inference



Stronger inferences …

… with partitive some-NPs





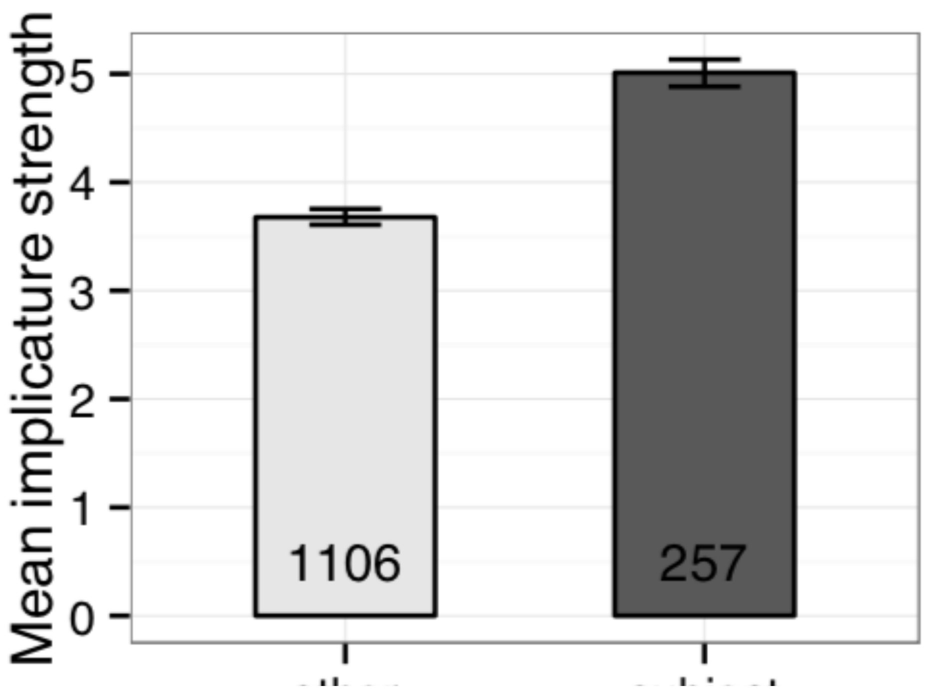
*I've seen **some of them** on repeats*

*so you ha-, you have been to **some family reunions**, perhaps.*

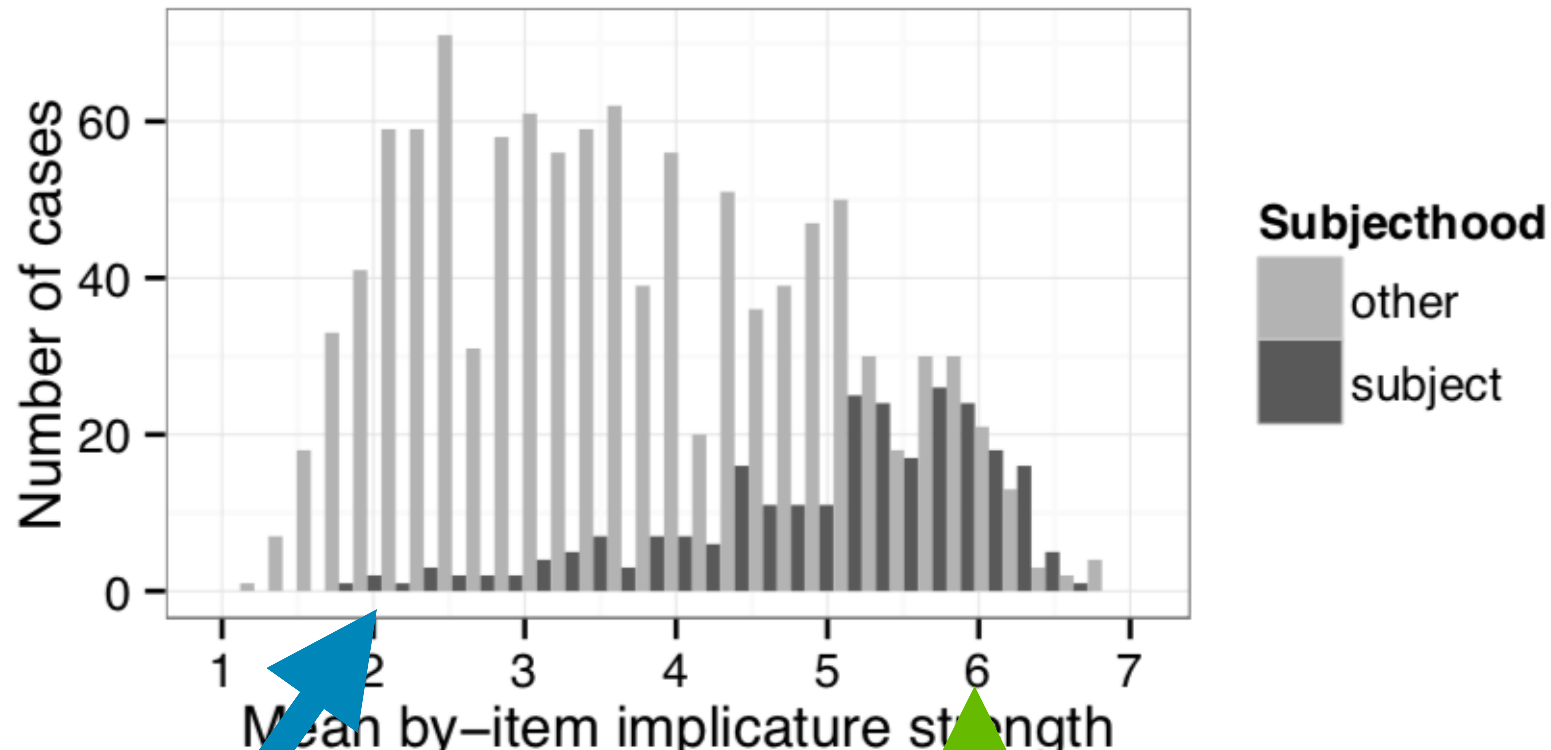
Degen (2015)

Stronger inferences …

… when some-NP is in subject position



other **subject**



***Some kids** are really having it.*

it would certainly help them to appreciate
some of the things …

Degen (2015)

is the model sensitive to these factors?

- **Minimal pair analysis:**
  (e.g., Marvin & Linzen, 2018; Futrell et al., 2019; Wilcox et al., 2019)
  Does the model make expected predictions on minimal sentence pairs varying along particular features?

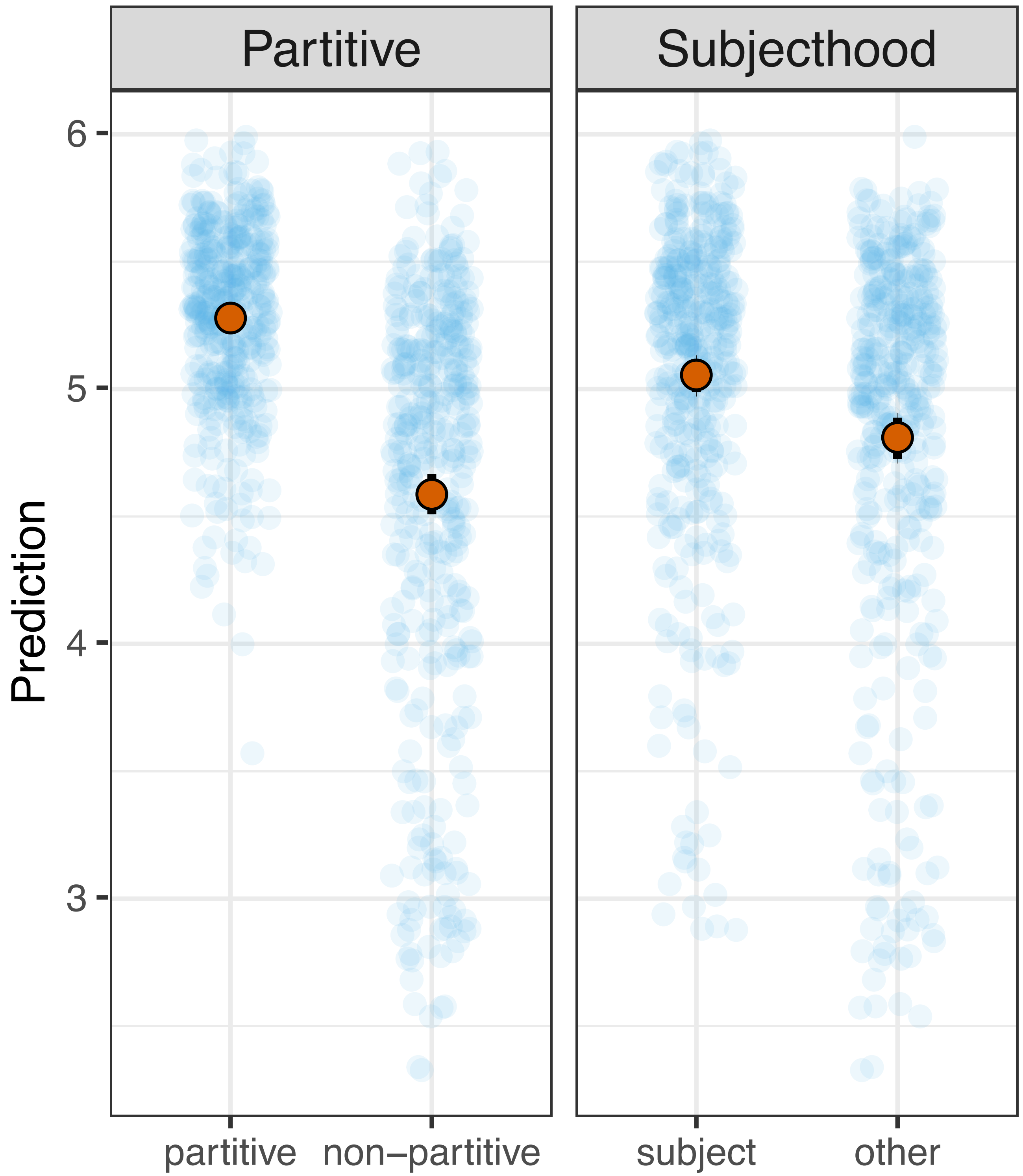
Minimal pair analysis

Manually constructed sentences that cross several linguistic factors, including subjecthood and partitive

1. **Some (of the) bakers** kneaded the dough.
2. The dough was kneaded by **some (of the) bakers**.
3. The bakers kneaded **some (of the) dough**.
4. **Some (of the) dough** was kneaded by the bakers.

25 items, 32 variants of each item = 800 sentences

Minimal pair analysis



Model predicts effects of linguistic features on artificial data set of minimal pairs!

Context

Speaker A: i mean, they just have beautiful, beautiful homes and they have everything. the kids only wear name brand things to school and it's one of these things,

Speaker B: oh me. well that makes it hard for you, doesn't it.

Speaker A: well it does, you know. it really does because i'm a single mom and i have a thirteen year old now and uh, you know, it does.

Speaker B: oh, me.

Speaker A: i mean, we do it to a point but uh, not to where she feels different ,

Speaker B: yeah.

Speaker A:
but some of them are very rich

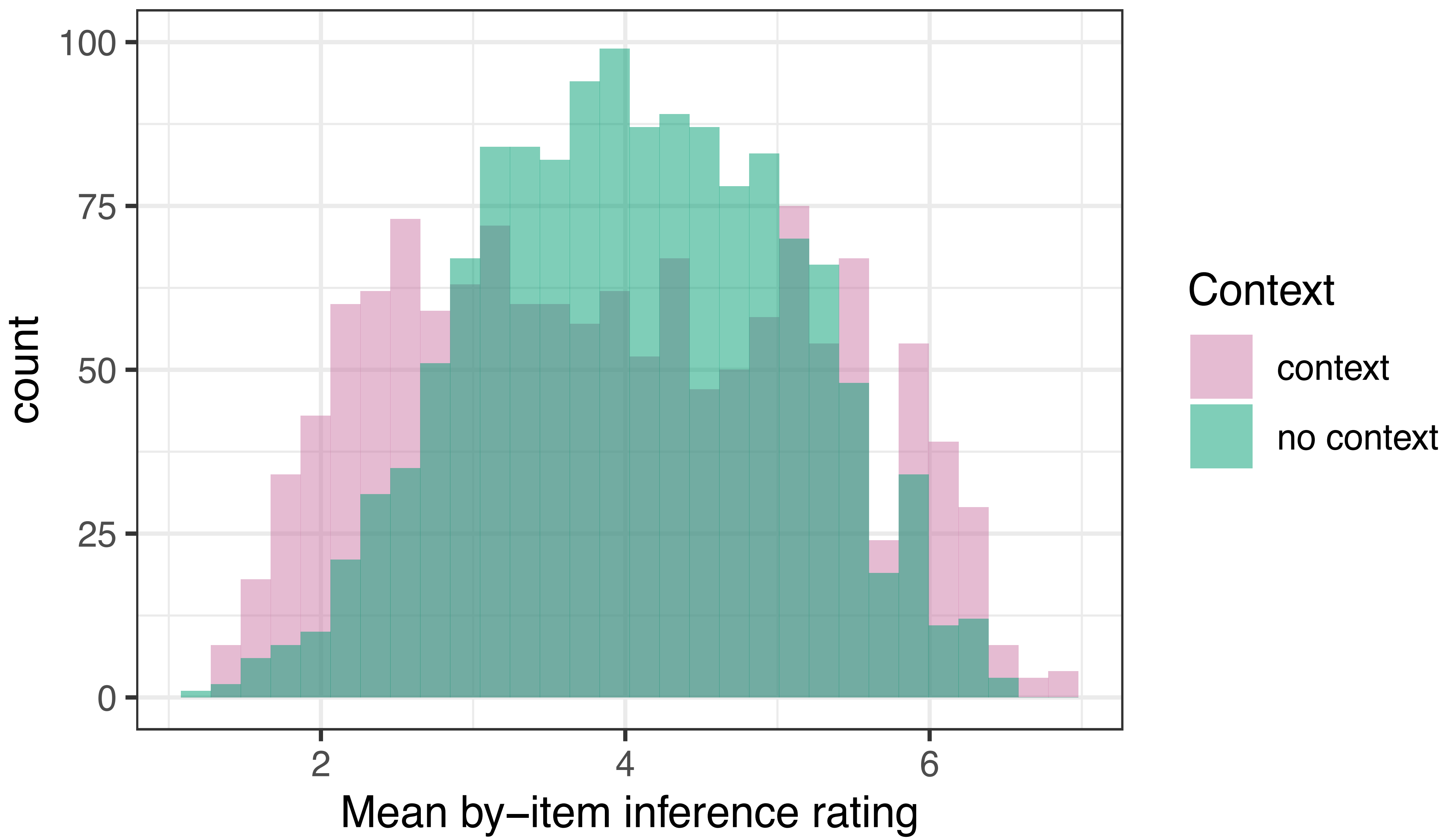
but **some, but not all** of them are very rich

How similar is the statement with 'some, but not all' (green) to the statement with 'some' (red)?

Very different meaning ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 Same meaning

Continue

Context



100
75
50
25
0
count
2
4
6
Mean by–item inference rating
Context
context
no context

Interesting context-sensitive "some" examples

- A: i took, uh, cammy to a … oh, it was a preschool day-care type of thing
- B: oh, uh-huh.
- A: but i kind of, i liked it some ways …
- **and some ways i didn't**.

no context: 2.80, context: 5.7, model prediction 4.4

Interim takeaways

- There exists **considerable variability** in the strength of scalar inferences across contexts
- Superficially, the **model can to a large extent learn** to closely predict human scalar inference strength for *some*
- Predictions primarily seem to be based on **associations between linguistic features and inference strength**
- Cannot make use of **larger conversational context**

Plan for today

1. To what extent can BERT learn to predict context-sensitive inferences from **“*some*” to “*some but not all*”**? {Schuster, Chen}, and Degen, 2020

2. To what extent can NLI models predict **presuppositions**? {Parrish, Schuster, Warstadt}, et al., 2021

3. To what extent can GPT-2 and GPT-3 track **discourse entities**? Schuster and Linzen, under review

Presuppositions

Chet finished law school

directly asserts

presupposes

Chet finished law school

Chet attended law school

Presuppositions project out of negation

Chet finished law school

Chet **didn't** finish law school

presupposes

presupposes

Chet attended law school

Presuppositions show context sensitivity

Chet never became a lawyer,
he **didn't** finish law school

presupposes

Chet just finished med school,
he **didn't** finish law school

doesn't presuppose

Chet attended law school

Presuppositions are gradient

Chet finished law school

Chet finished the last year of law school

presupposes weakly

presupposes strongly

Chet attended law school

Research Questions

- How much does context affect projection out of negation for a wide range of presupposition triggers?

- How well can natural language inference models predict (context-sensitive) presuppositions?

Presupposition datasets

- Existing datasets either
  - lack naturalistic contexts (e.g., *MegaVeridicality*, White et al., 2018, *ImpPres*, Jeretič et al., 2020)
  - focus on one trigger type (e.g., CommitmentBank, de Marneffe et al., 2019; Ross and Pavlick, 2019)
- NOPE provides examples with naturalistic contexts for a range of trigger types

Trigger types

Lexical triggers:

- Change of state (*appear, melt*)
- Aspectual verbs (*stop, start*)
- Embedded questions (*know why, see how*)
- Clause embed. verbs (*realize, regret*)
- Implicatives (*manage to, fail to*)
- Numeric determiners (*both, the three*)
- ‘Re-’ prefixed verbs (*rebuild, retell*)
- Temporal adverbs (*before, after*)

Syntactic triggers:

- Clefts (*It’s the X that Y*)
- Comparatives (*X is a Y-er Z than …*)

Example construction

Sentence from COCA:

Kmart declined to comment.

Expert negated sentence:

Kmart did not decline to comment.

Expert-written presupposition:

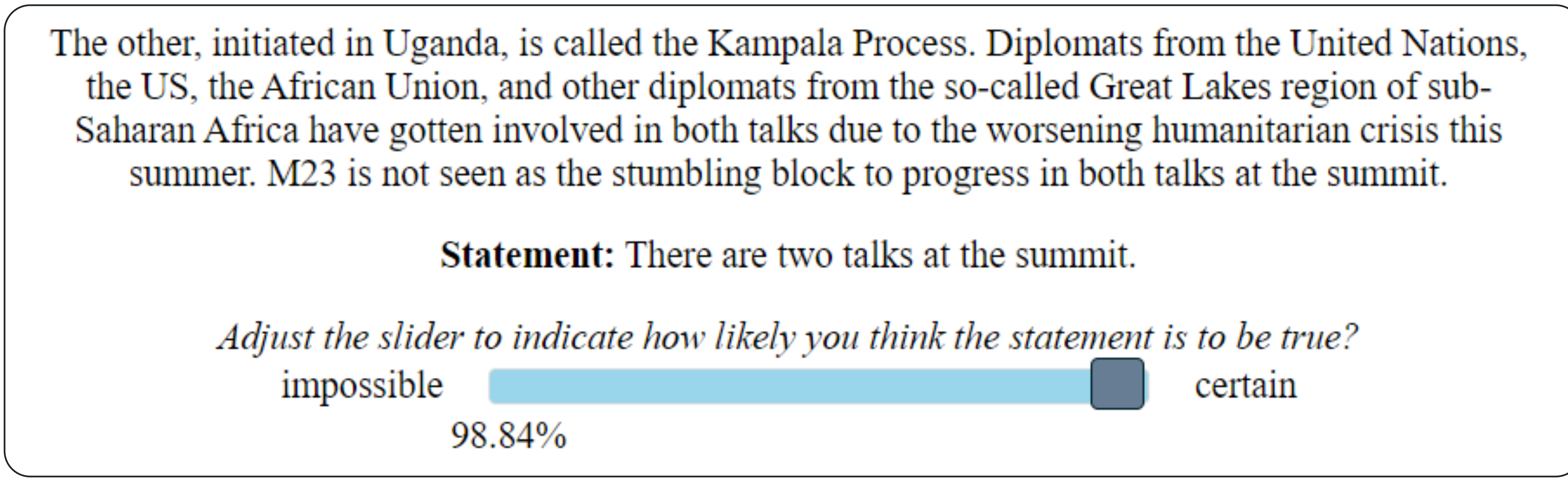
Kmart was asked to comment.

Context from COCA (2 preceding sentences):
In the Noels' case, the foundation contacted Kmart. Within a few months the company revised its insurance to cover up to $500,000 annually for inpatient and outpatient care combined.

Human Experiments & Results

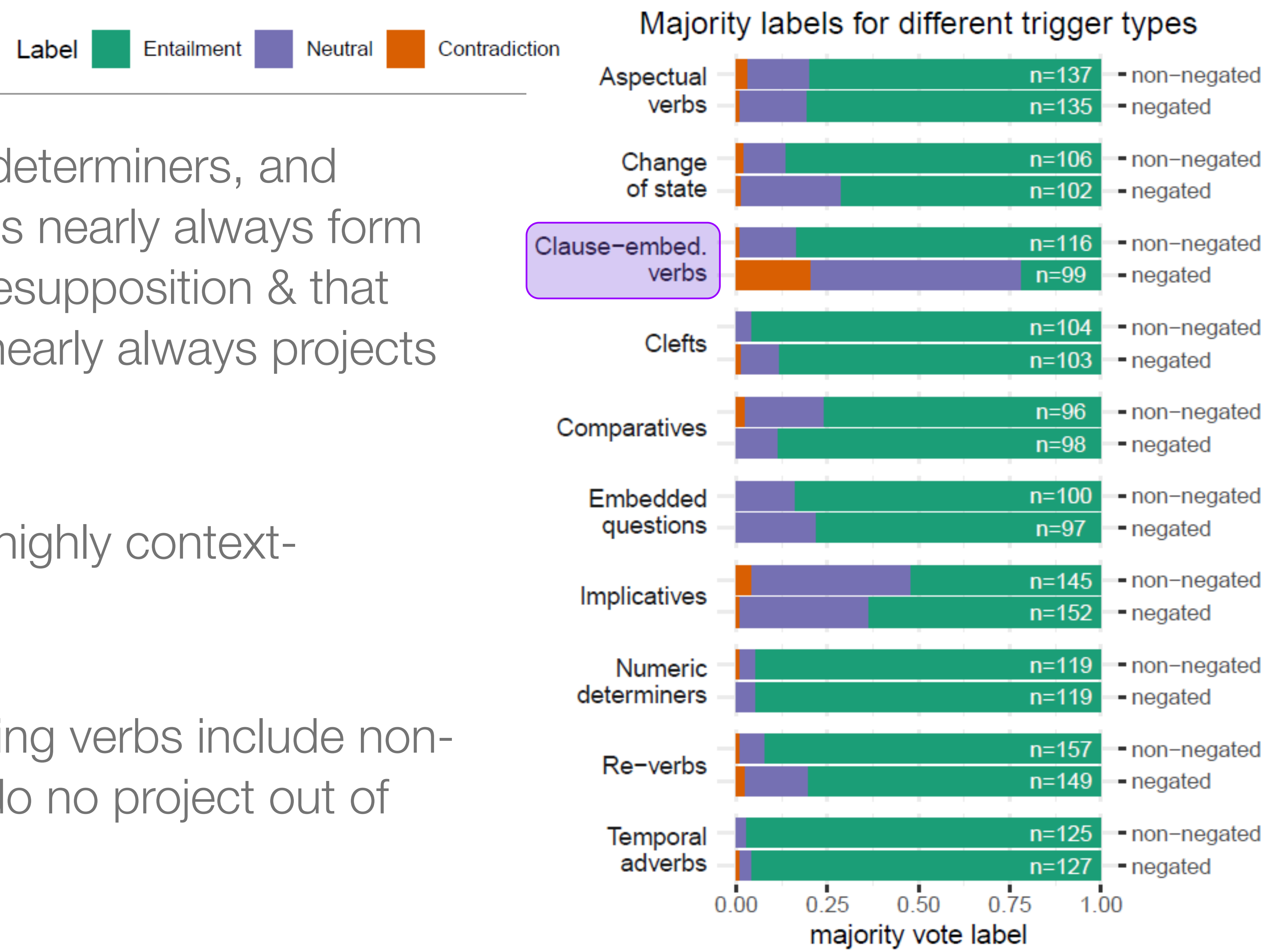
Task description

- Qualified MTurk annotators used a slider to rate how likely a statement is



- Map to NLI labels

Results



- Clefts, numeric determiners, and temporal adverbs nearly always form the expected presupposition & that presupposition nearly always projects out of negation
- Implicatives are highly context-dependent
- Clause-embedding verbs include non-triggers, which do no project out of negation

Modeling Experiments & Results

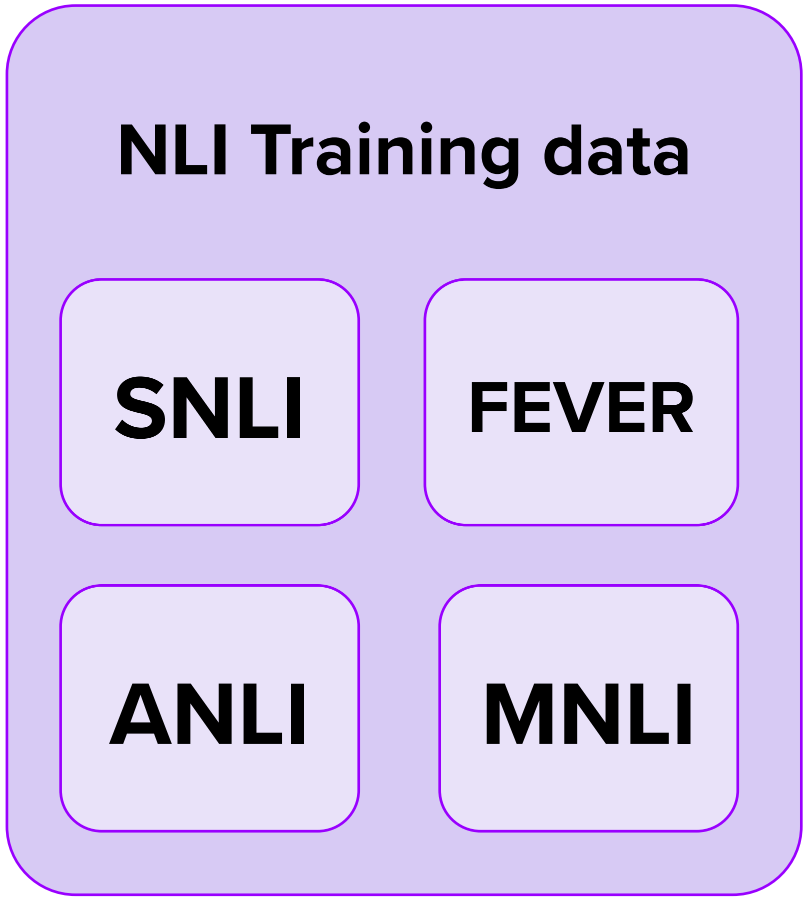
Models & Training

Pretrained Models

- RoBERTa-large
- DeBERTa-V2-XL

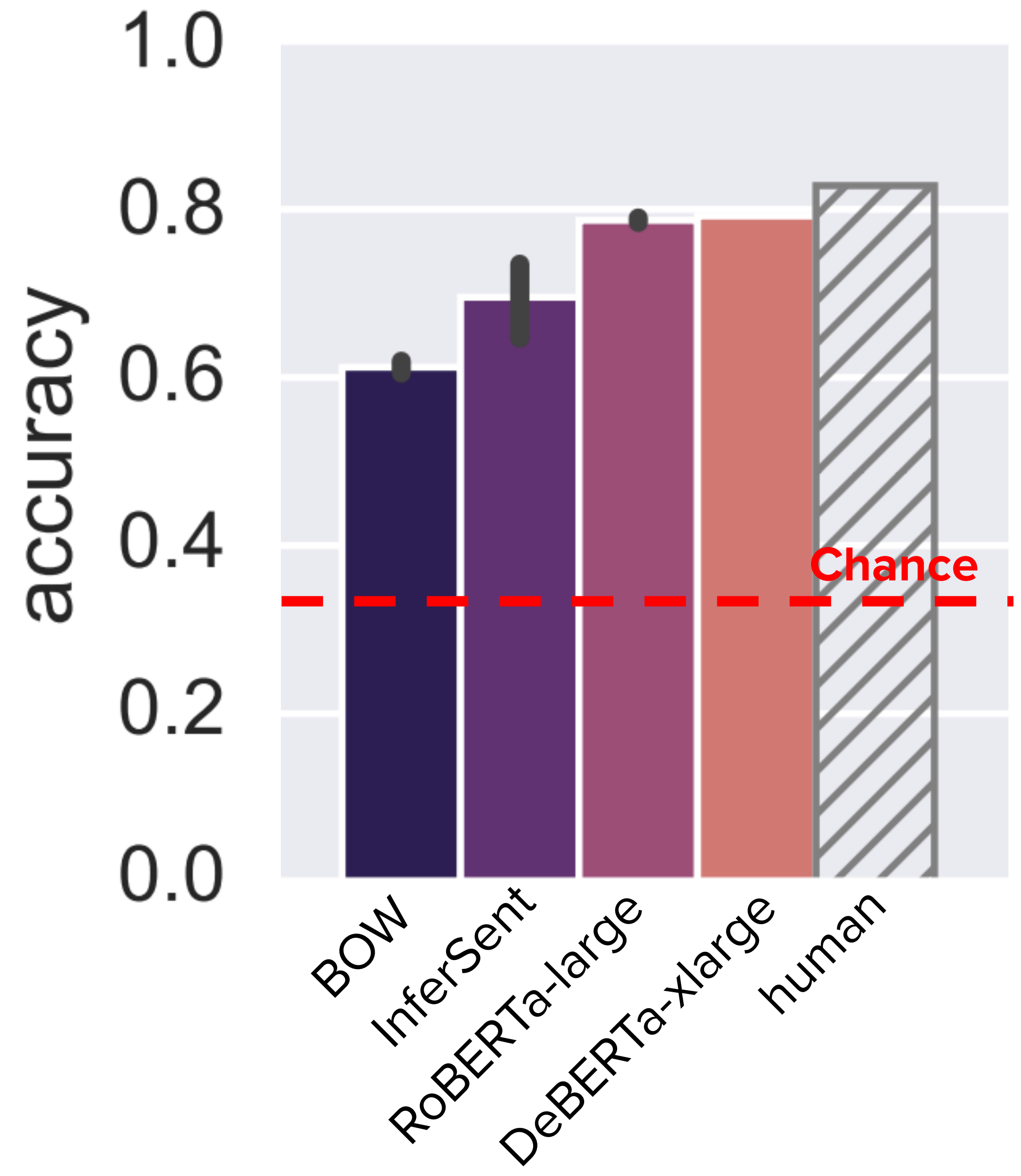
Baselines (only NLI training)

- BoW (FastText)
- InferSent

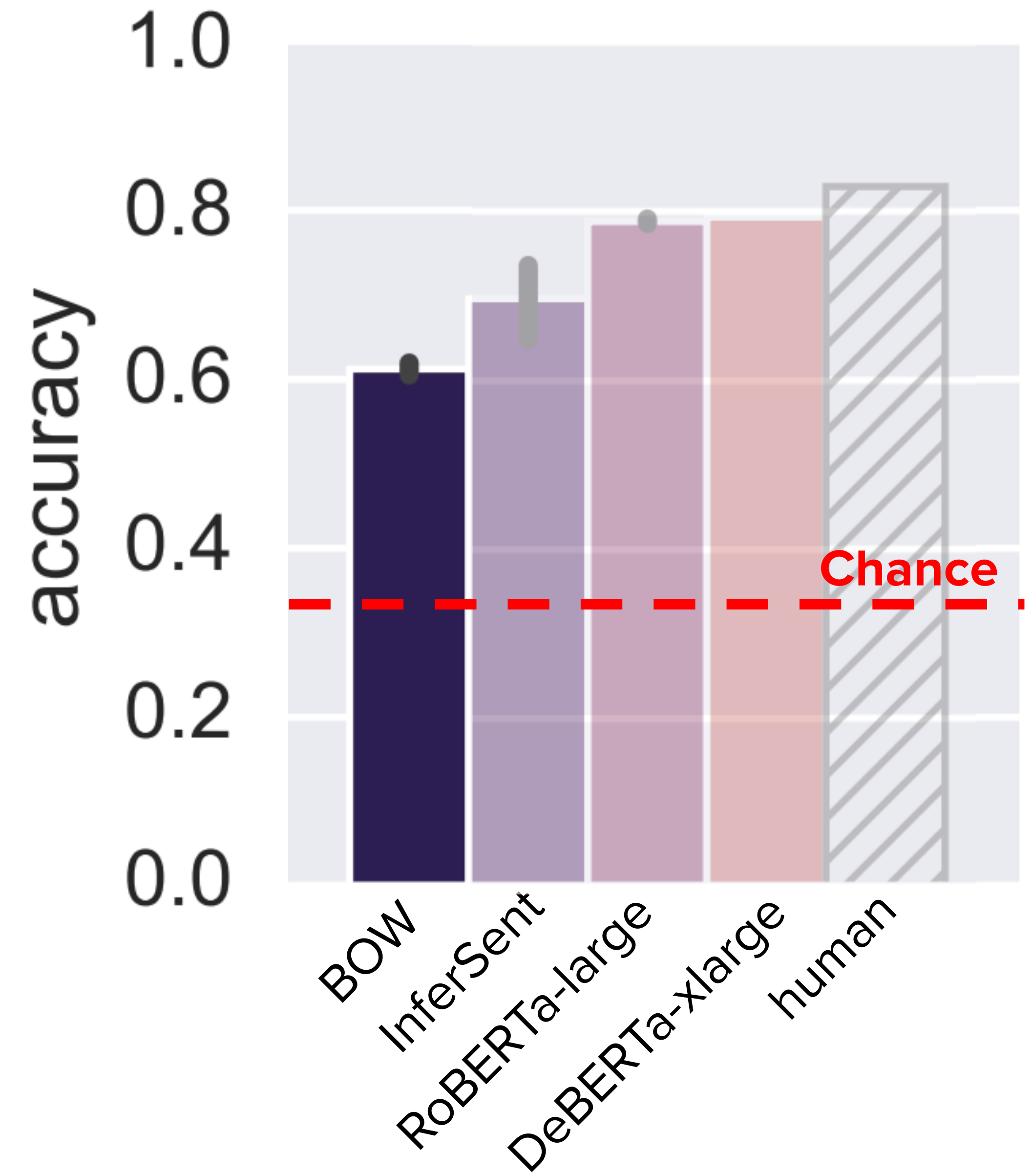


Main results

- Human performance is % of responses that agree with majority.
- Baselines performs well above chance.
- Transformers have strongest performance, near-human level.



Shallow heuristics?



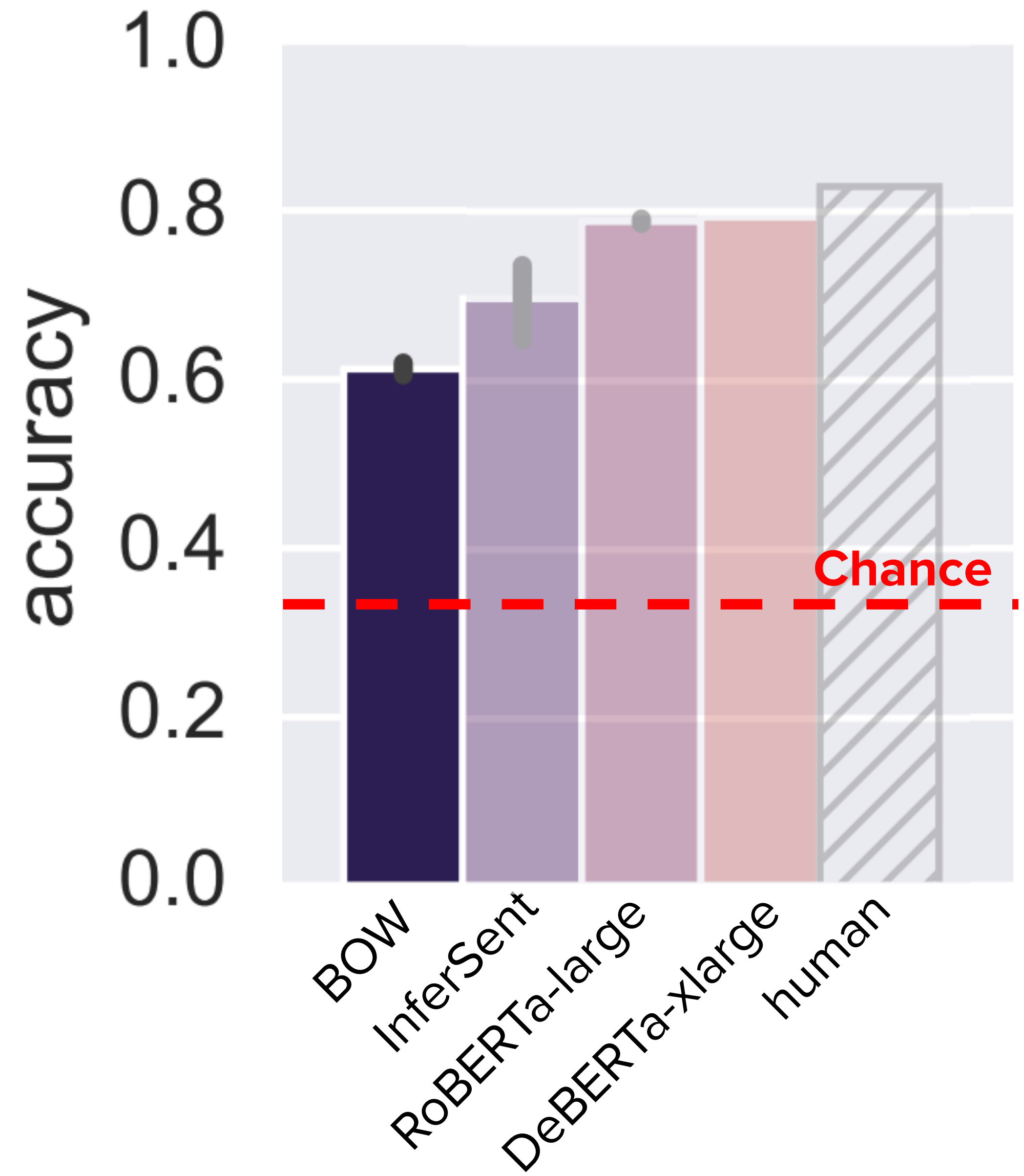
Shallow heuristics?

Trigger sentence

Women from **both sides** of town formed a mothers group.

Presupposition sentence

There are two sides of town.



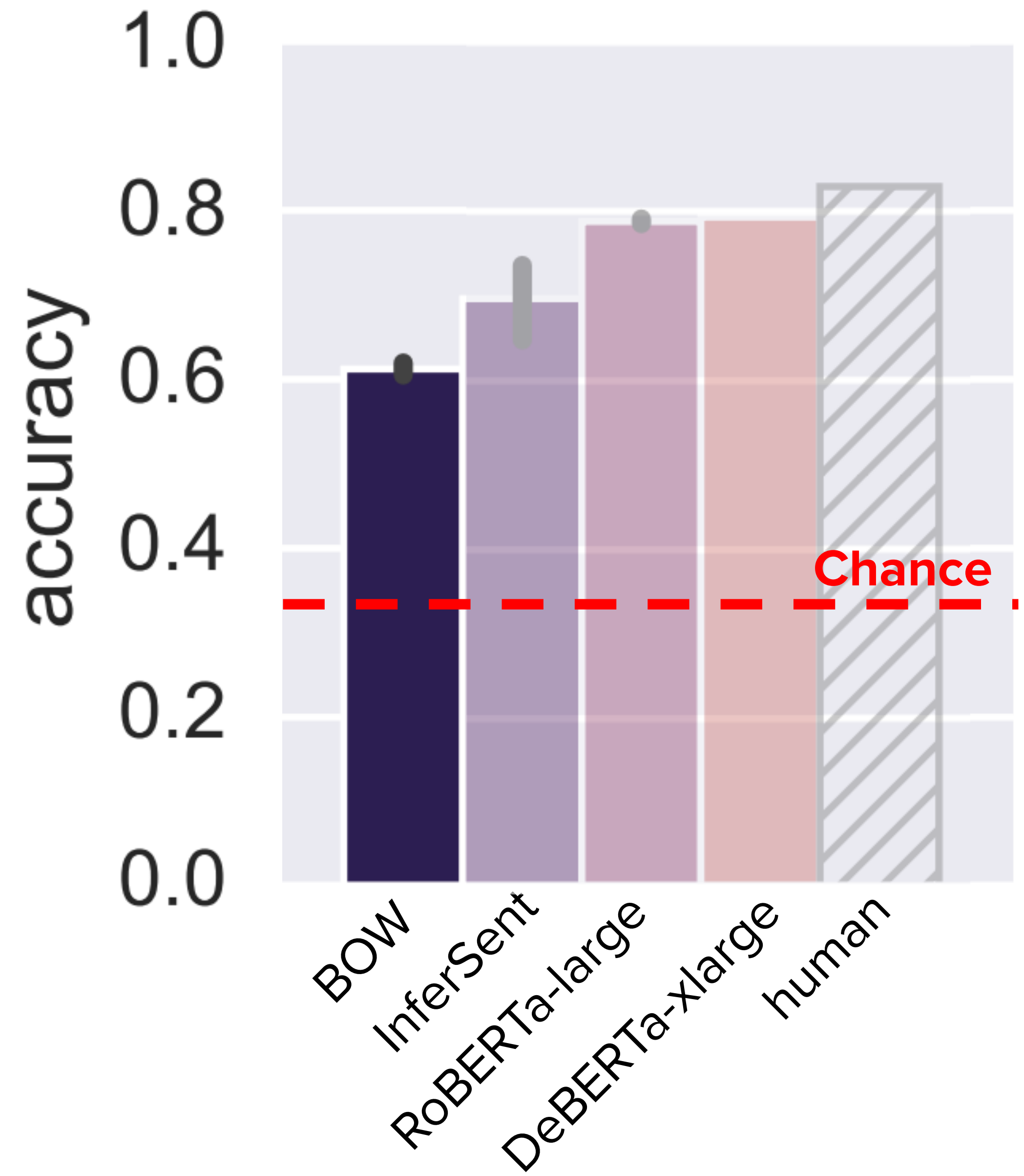
Shallow heuristics?

Trigger sentence

Women from **both sides** of town formed a mothers group.

Presupposition sentence

There are two sides of town.



Shallow heuristics?

Trigger sentence

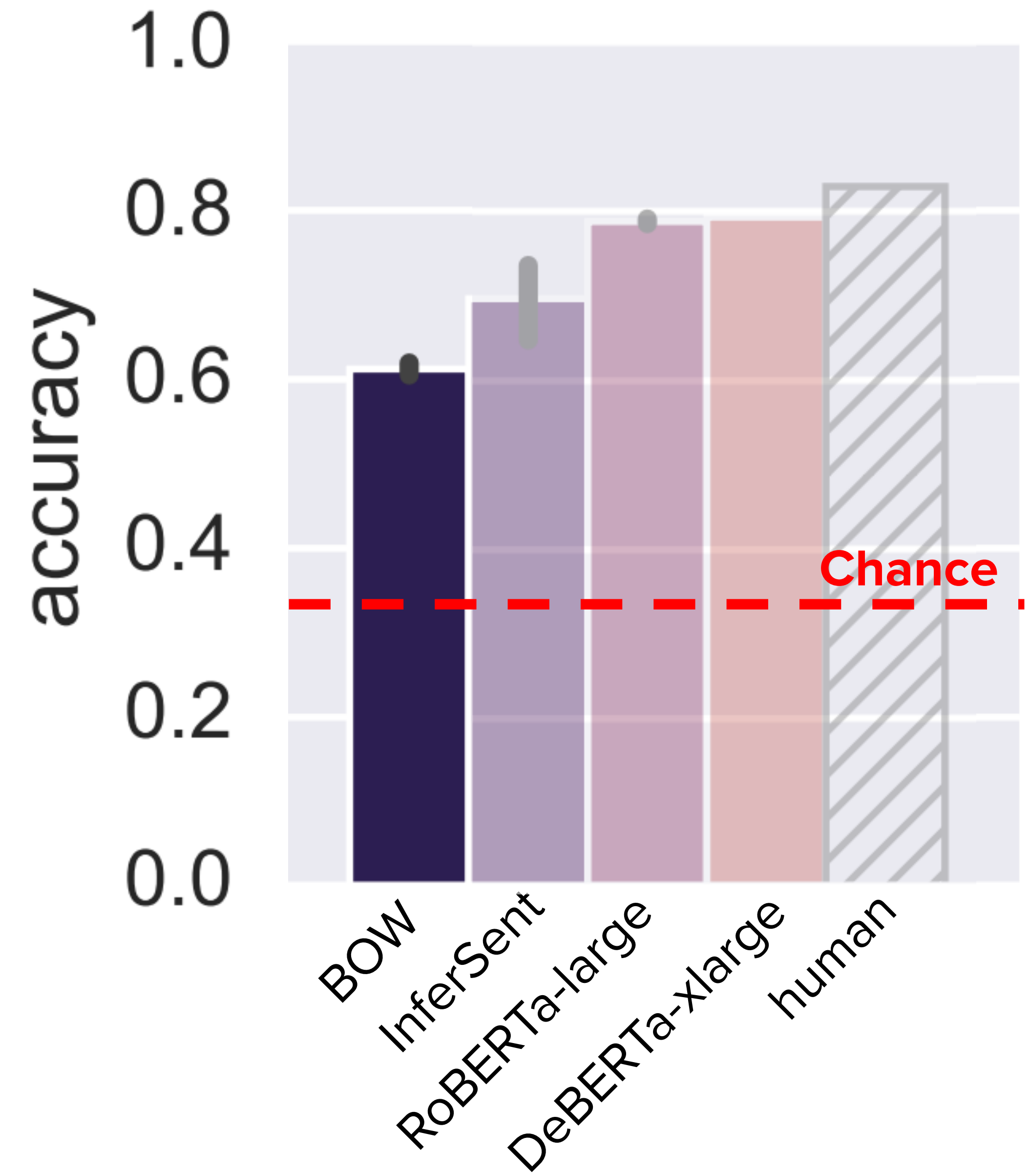
Women from **both sides** of town formed a mothers group.

Presupposition sentence

There are two sides of town.

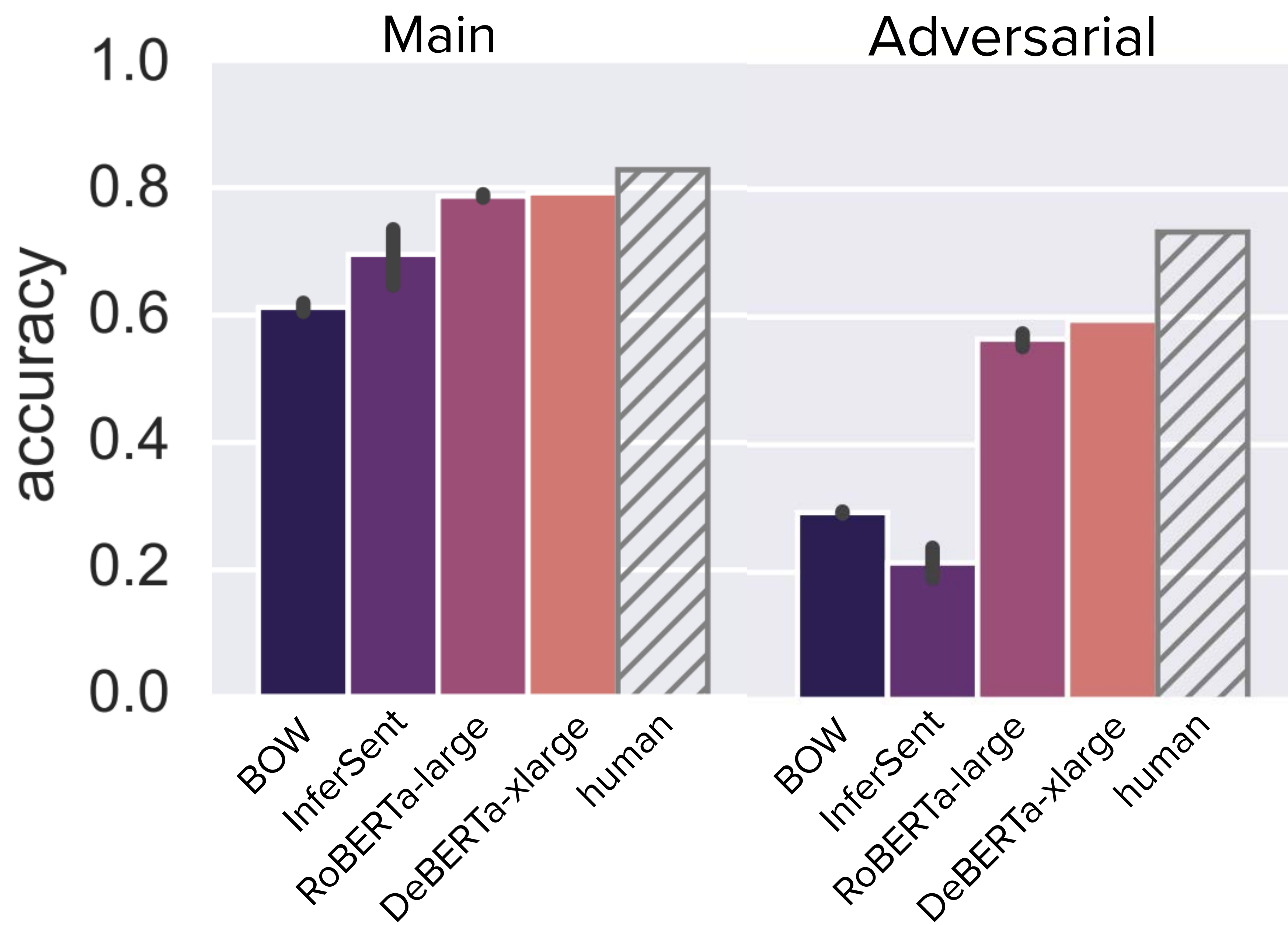
Adversarial sentence

There are three **sides of town**.



Adversarial results

- Human performance is not strongly affected by adversarial perturbation.
- Baseline models are reduced to chance accuracy or worse.
- Pre-trained transformers are slightly affected but still perform way above chance



Context sensitivity?

Model	E ⟶ {N,C} nonneg	neg
RoBERTa		
DeBERTa		

Premise
Chet just finished med school… he finished law school.

Hypothesis
Chet attended law school.

Context sensitivity?

Model	E nonneg
RoBERTa	80.6
DeBERTa	81.8

Premise
Chet just finished med school... He finished law school.

Hypothesis
Chet attended law school.

Context sensitivity?

Model	E ⟶ {N,C} nonneg	neg
RoBERTa	80.6	32.7
DeBERTa	81.8	32.1

Premise
Chet just finished med school...
He **didn't** finish law school.

Hypothesis
Chet attended law school.

Conclusions

- Presupposition triggers are “real”, but so is cancellability and gradience.
- Pretrained Transformers learn some of the basic characteristics of presuppositions like projection, but **do not show human-like context-sensitivity and variability**

Plan for today

1. To what extent can BERT learn to predict context-sensitive inferences from **"*some*" to "*some but not all*"**? {Schuster, Chen}, and Degen, 2020

2. To what extent can NLI models predict **presuppositions**? {Parrish, Schuster, Warstadt}, et al., 2021

3. To what extent can GPT-2 and GPT-3 track **discourse entities**? Schuster and Linzen, under review

Why may language models struggle with larger conversational context?

1	2
John owns(2)	🐶 is-owned-by(1)

John owns a dog.

Heim (1982)

To what extent can language models keep track of discourse entities?

To what extent are language models sensitive to contextual factors that modulate whether an indefinite noun phrase introduces a discourse entity?

Schuster and Linzen (under review)

The phenomenon

- Indefinite noun phrases generally introduce discourse entities…
  - John owns **a dog.** ***It*** *has a red collar.*
  - Sarah managed to buy **a car.** ***It*** *gets really good mileage.*
  - I know that Carol built **a house**. ***It*** *is very spacious.*

e.g., Karttunen (1976), Heim (1981)

The phenomenon

- …. but not always (with lots of additional caveats):
  - John doesn't own **a dog.** # ***It*** *has a red collar.*
  - Sue failed to write **a book. #** ***It*** *is a real page-turner.*
  - I doubt that Michael baked **a pie**. # ***It*** *was delicious*.
  - Sarah wants to knit **a hat**. # ***It*** *is very colorful.*

e.g., Karttunen (1976), Heim (1981)

Methodology

	Referential: It has a red collar	**Non-referential:** It's not a big deal
A: John **owns** a dog	✅	✅
B: John **doesn't own** a dog	❌	✅

Expected language model behavior

	Referential: It has a red collar	**Non-referential:** It's not a big deal
A: John **owns** a dog	0.2	0.2
B: John **doesn't own** a dog	0.001	0.2

$$\frac{P(\text{Ref} \mid A)}{P(\text{Non-Ref} \mid A)} > \frac{P(\text{Ref} \mid B)}{P(\text{Non-Ref} \mid B)}$$

Dataset

- Targets four types of operators that modulate whether discourse entity is introduced:

  - **Affirmative vs. negation**
    A: John **owns** a dog.
    B: John **doesn't own** a dog.

  - **Embedding under factive/non-factive predicates**
    A: I **know** that John owns a dog.
    B: I **doubt** that John owns a dog.

Dataset

- Targets four types of operators that modulate whether discourse entity is introduced:

  - **Modals**
    A: John **owns** a dog.
    B: John **wants to own** a dog.

  - **Embedding under implicative/negative implicative predicates**
    A: John **managed to** adopt a dog.
    B: John **failed to** adopt a dog.

16 hand-written items —> 64 pairs

Language models

- GPT-2 in various sizes:
    - **GPT-2**: 117M parameters
    - **GPT-2-medium**: 345M parameters
    - **GPT-2-large**: 762M parameters
    - **GPT-2-xl**: 1542M parameters
- **GPT-3 (davinci)**: 175B parameters?

trained on
~ 8 billion tokens

trained on
~ 500 billion tokens

Human experiment

Please read the following sentence (or part of a sentence)
and click on the continuation that makes more sense to you:

John owns a dog

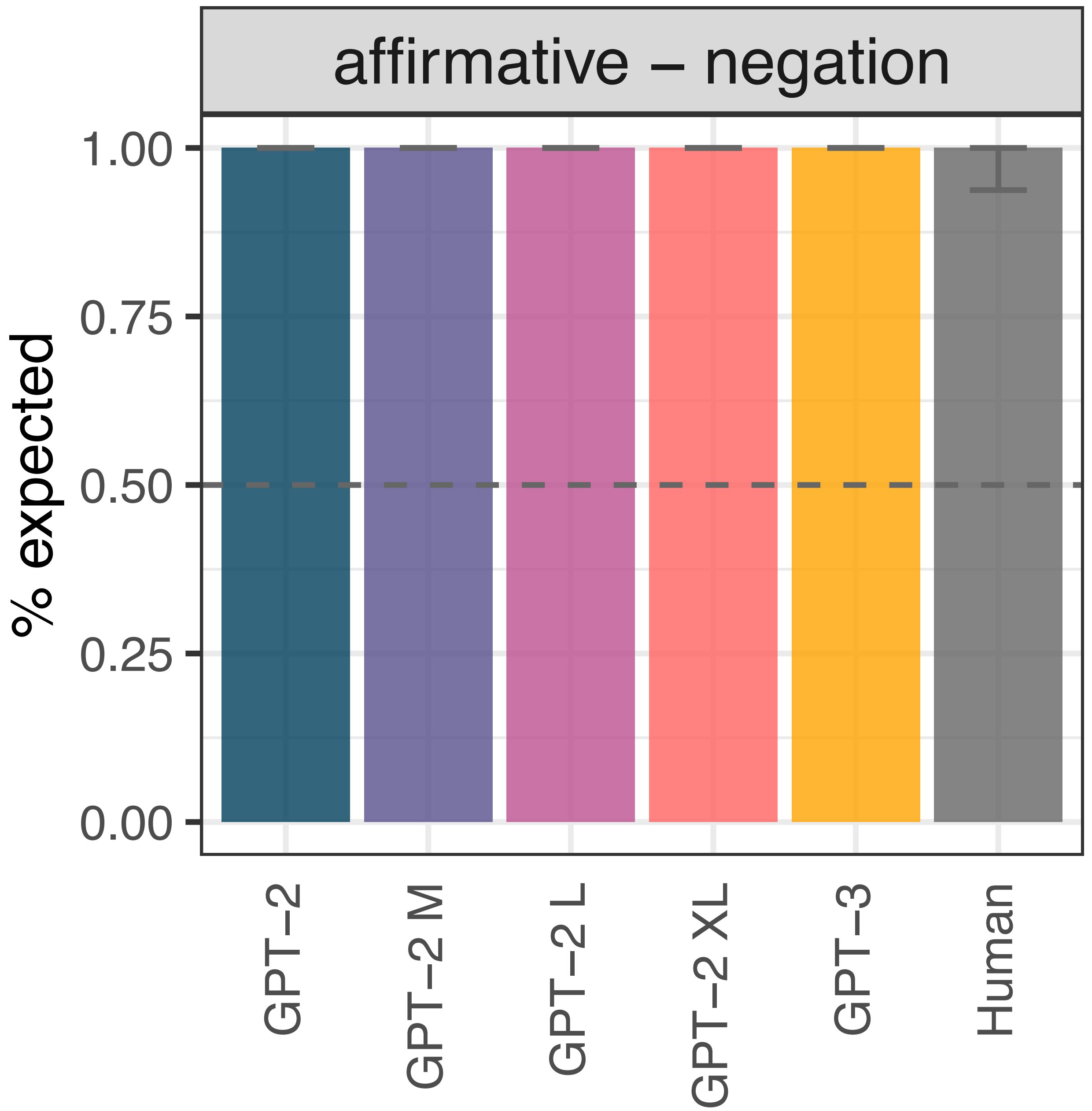
Continuations:

and it's not a big deal.

and it follows him everywhere he goes.

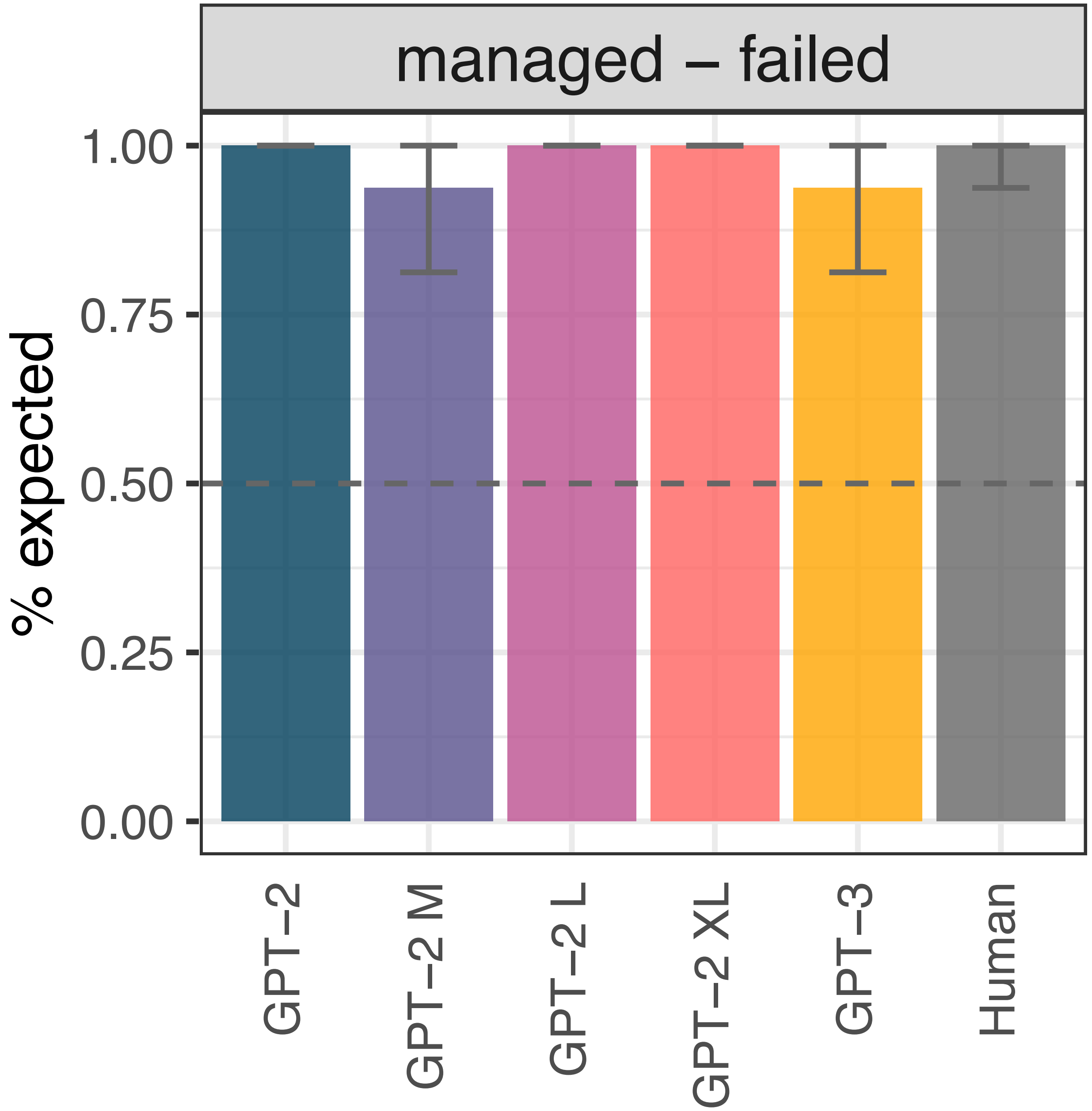
Results

$$\frac{P(\text{Ref} \mid A)}{P(\text{Non-Ref} \mid A)} > \frac{P(\text{Ref} \mid B)}{P(\text{Non-Ref} \mid B)}$$



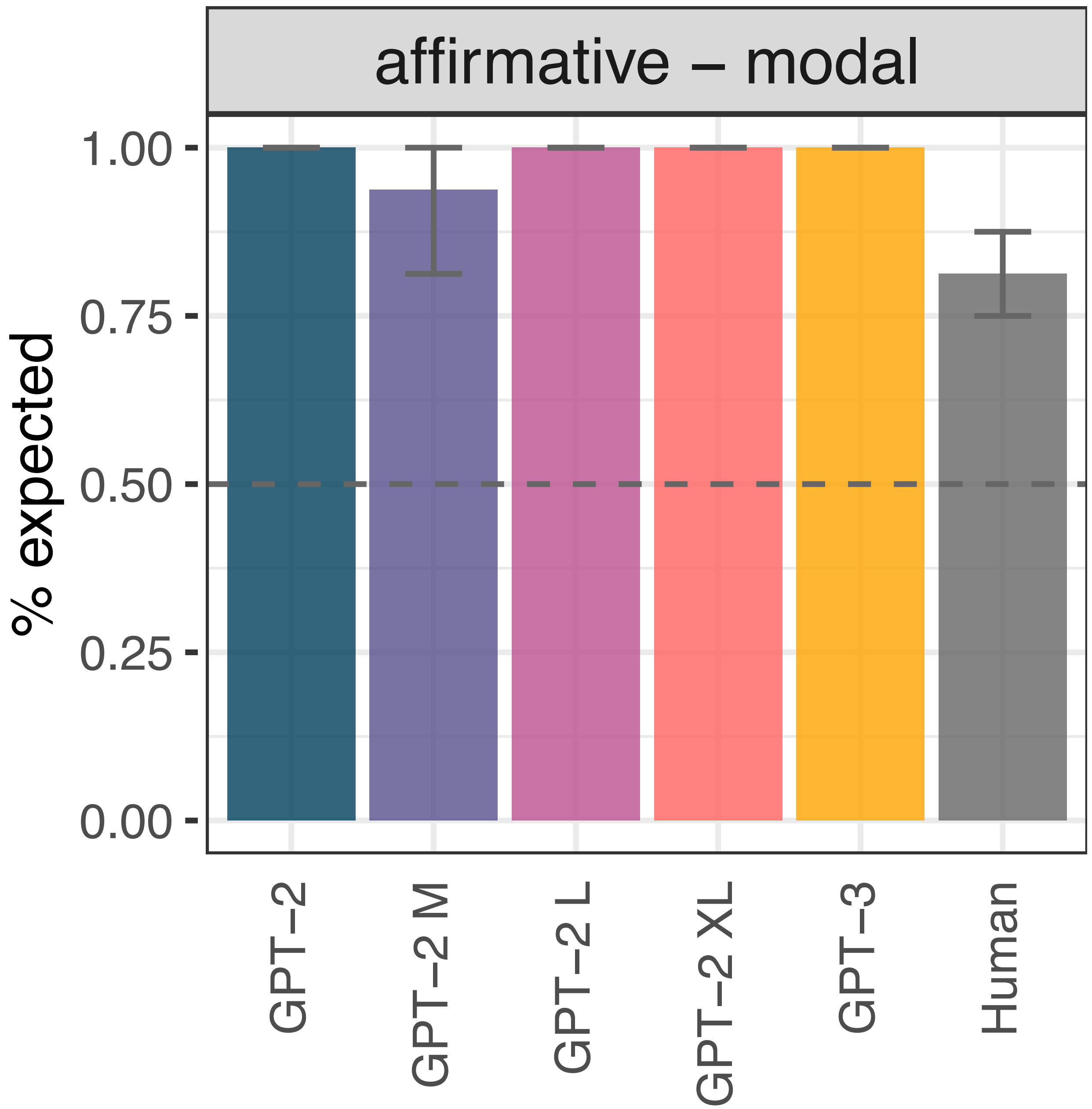
Results

$$\frac{P(\text{Ref} \mid A)}{P(\text{Non-Ref} \mid A)} > \frac{P(\text{Ref} \mid B)}{P(\text{Non-Ref} \mid B)}$$



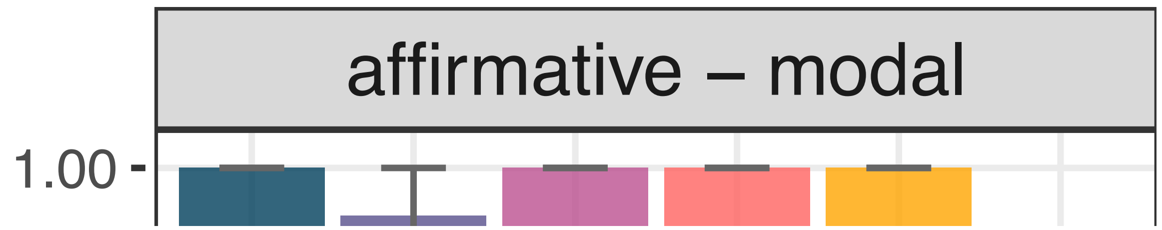
Results

$$\frac{P(\text{Ref} \mid A)}{P(\text{Non-Ref} \mid A)} > \frac{P(\text{Ref} \mid B)}{P(\text{Non-Ref} \mid B)}$$

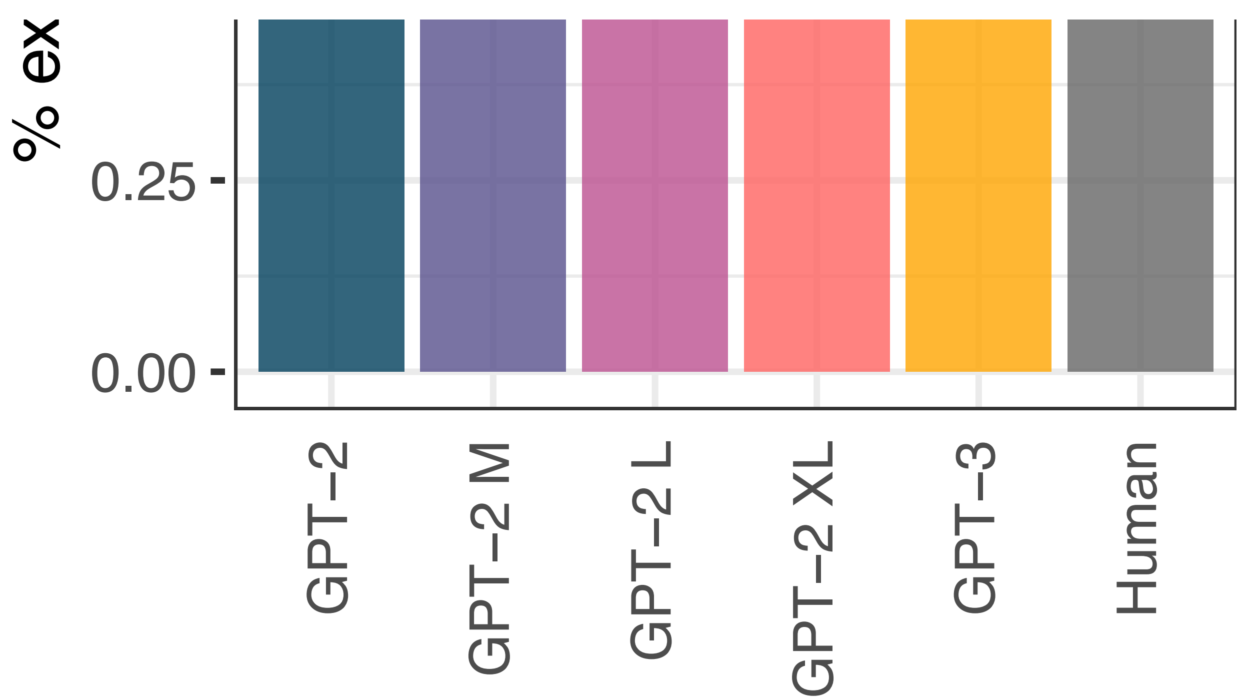


Results

$$\frac{P(\text{Ref} \mid A)}{P(\text{Non-Ref} \mid A)} > \frac{P(\text{Ref} \mid B)}{P(\text{Non-Ref} \mid B)}$$

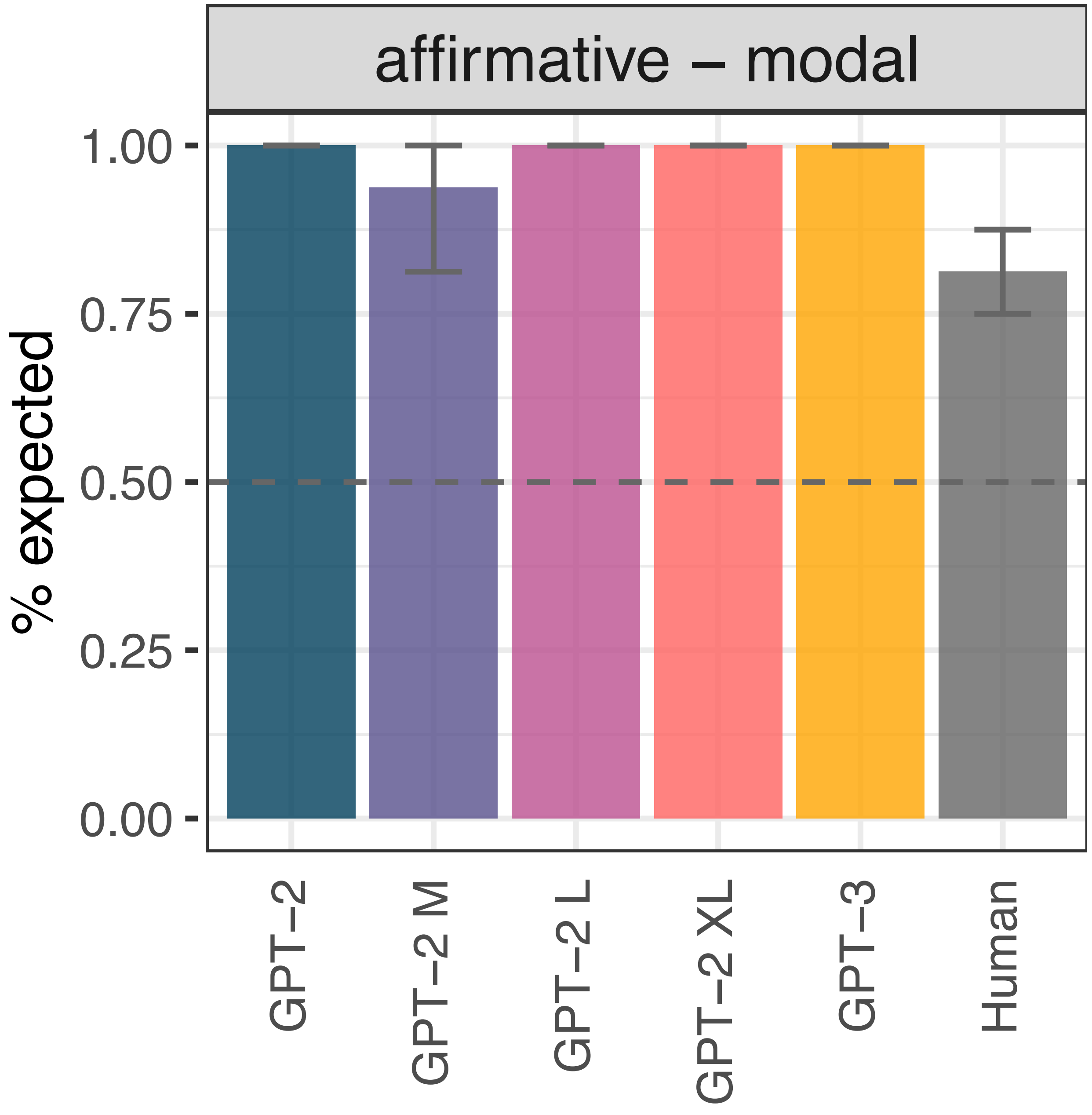


Michael wants to bake a cake … and it ~~was~~ **will be** the best thing at the picnic



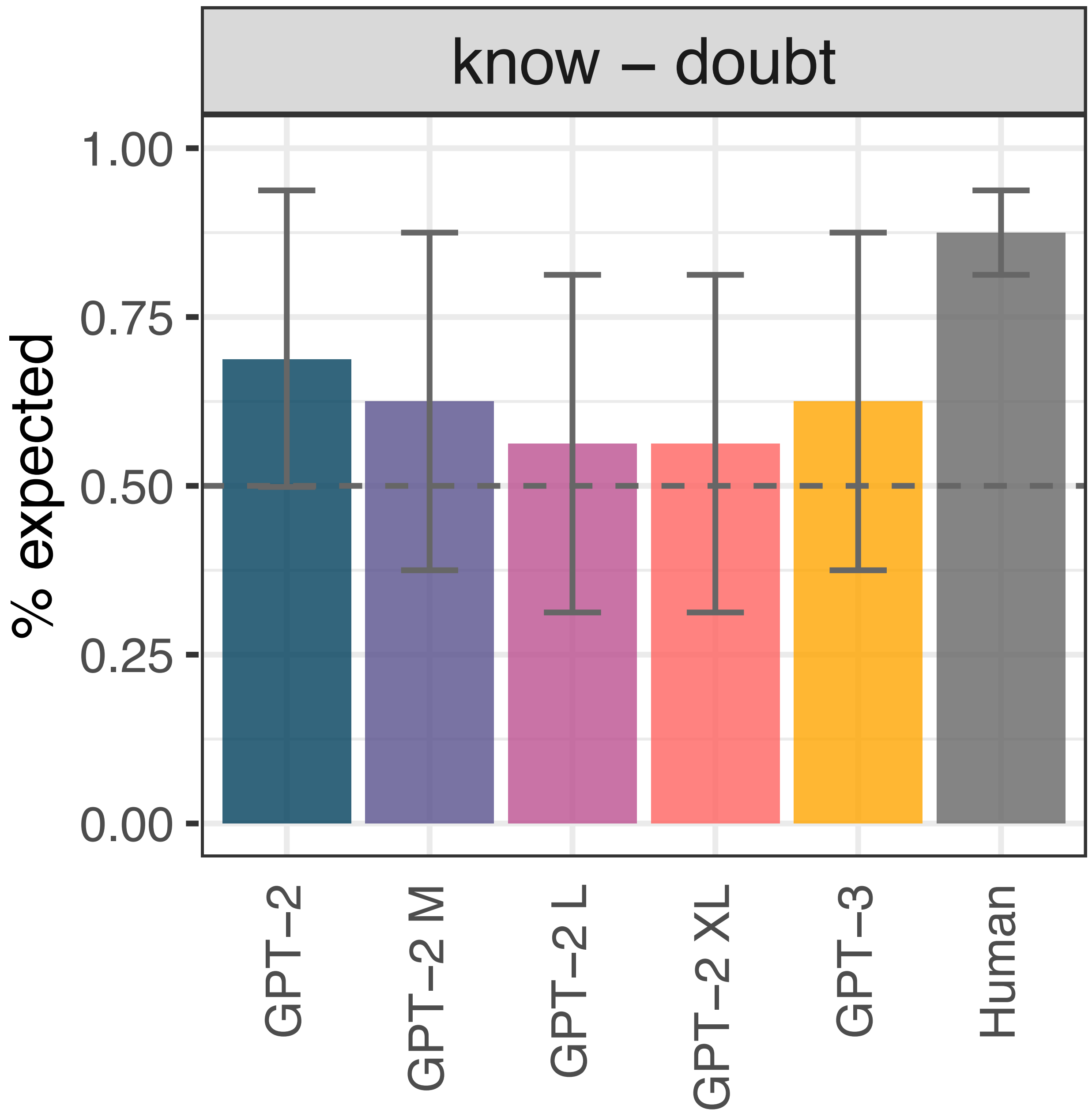
Results

$$\frac{P(\text{Ref} \mid A)}{P(\text{Non-Ref} \mid A)} > \frac{P(\text{Ref} \mid B)}{P(\text{Non-Ref} \mid B)}$$



Results

$$\frac{P(\text{Ref} \mid A)}{P(\text{Non-Ref} \mid A)} > \frac{P(\text{Ref} \mid B)}{P(\text{Non-Ref} \mid B)}$$



Interim conclusions

- Human preferences for continuations are largely in line with patterns predicted by most linguistic theories
- Except for the factive vs. non-factive (*know* and *doubt*) condition, all language models seem to be sensitive to the contrasts
- Is this a result of combining sentential operators and embedding predicates with indefinite noun phrases as humans do? Or could these be spurious correlations?

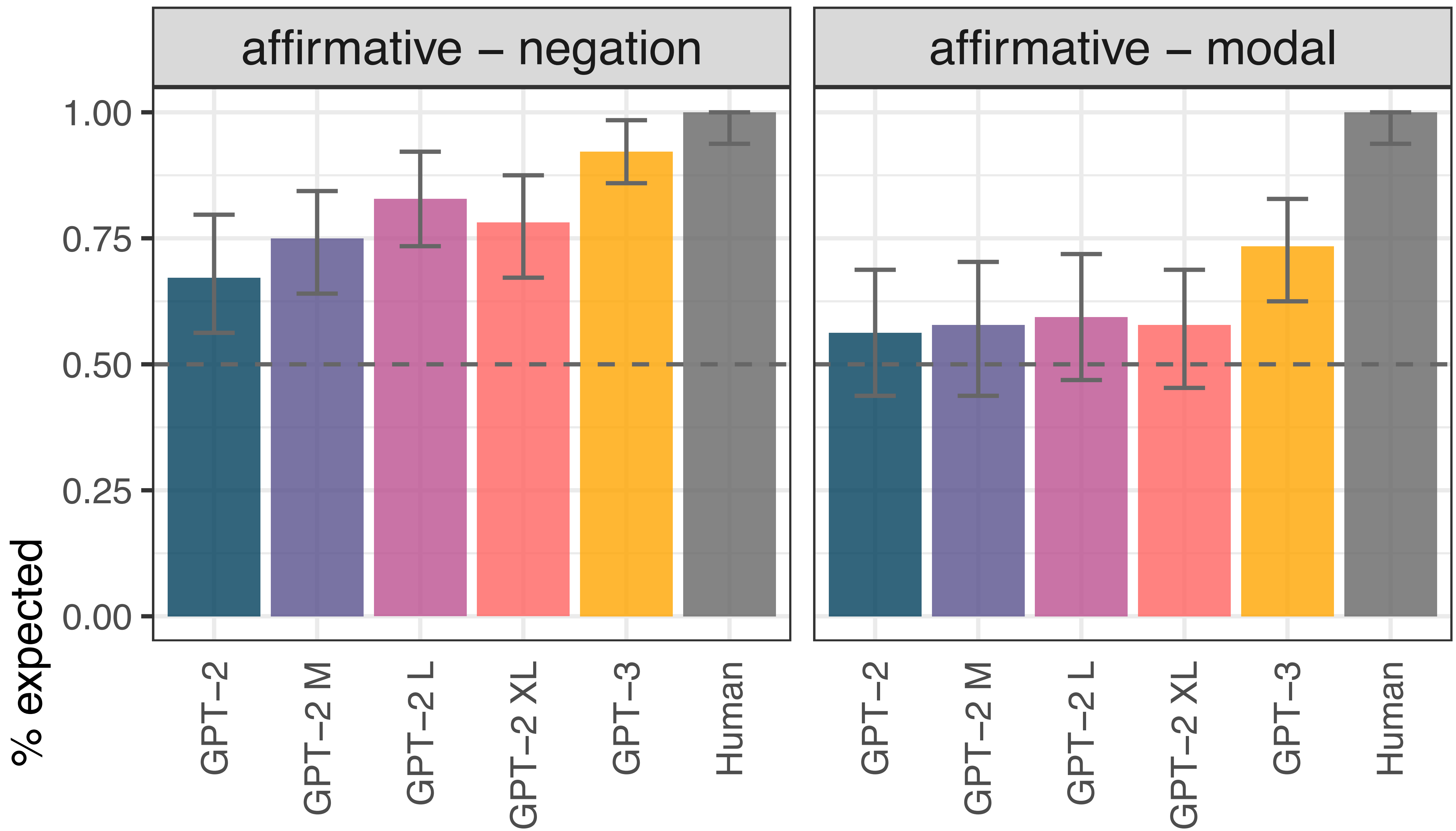
Multiple noun phrases

- Mary **found a shirt** at the store but she **didn't find a hat**
- Coreferential continuations:
  - P("The **shirt** was blue") > P("The **hat** was blue")
- Non-coreferential continuations:
  - P("The **hat** that she tried on didn't fit") >
    P("The **shirt** that she tried on didn't fit)

Results: Co-referential continuations

Mary **found a shirt** at the store but she **didn’t find a hat**

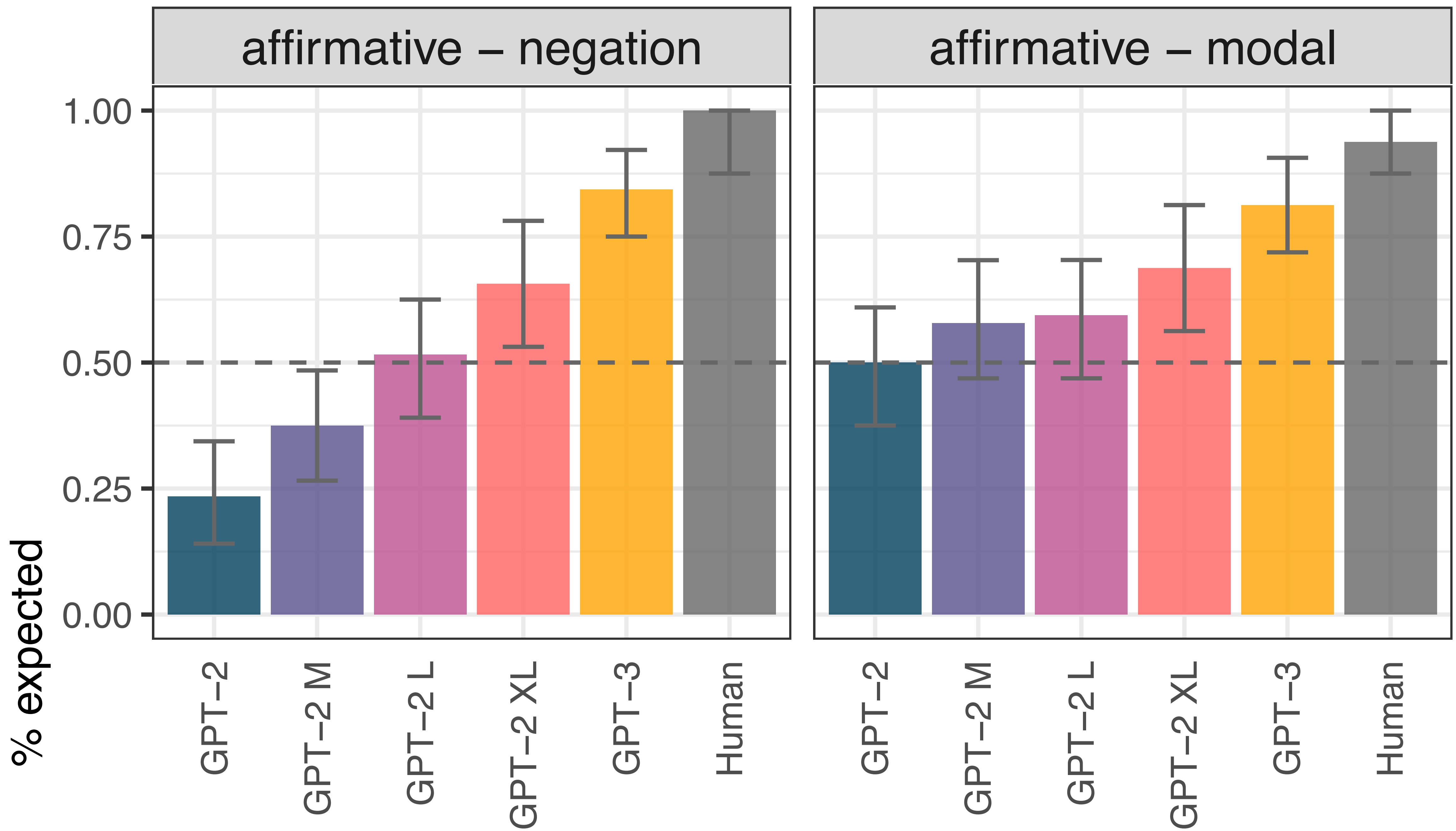
P(“The **shirt** was blue”) > P(“The **hat** was blue”)



Results: Non-coreferential continuations

Mary **found a shirt** at the store but she **didn't find a hat**

P("The **hat** that she tried on…") > P("The **shirt** that she …")



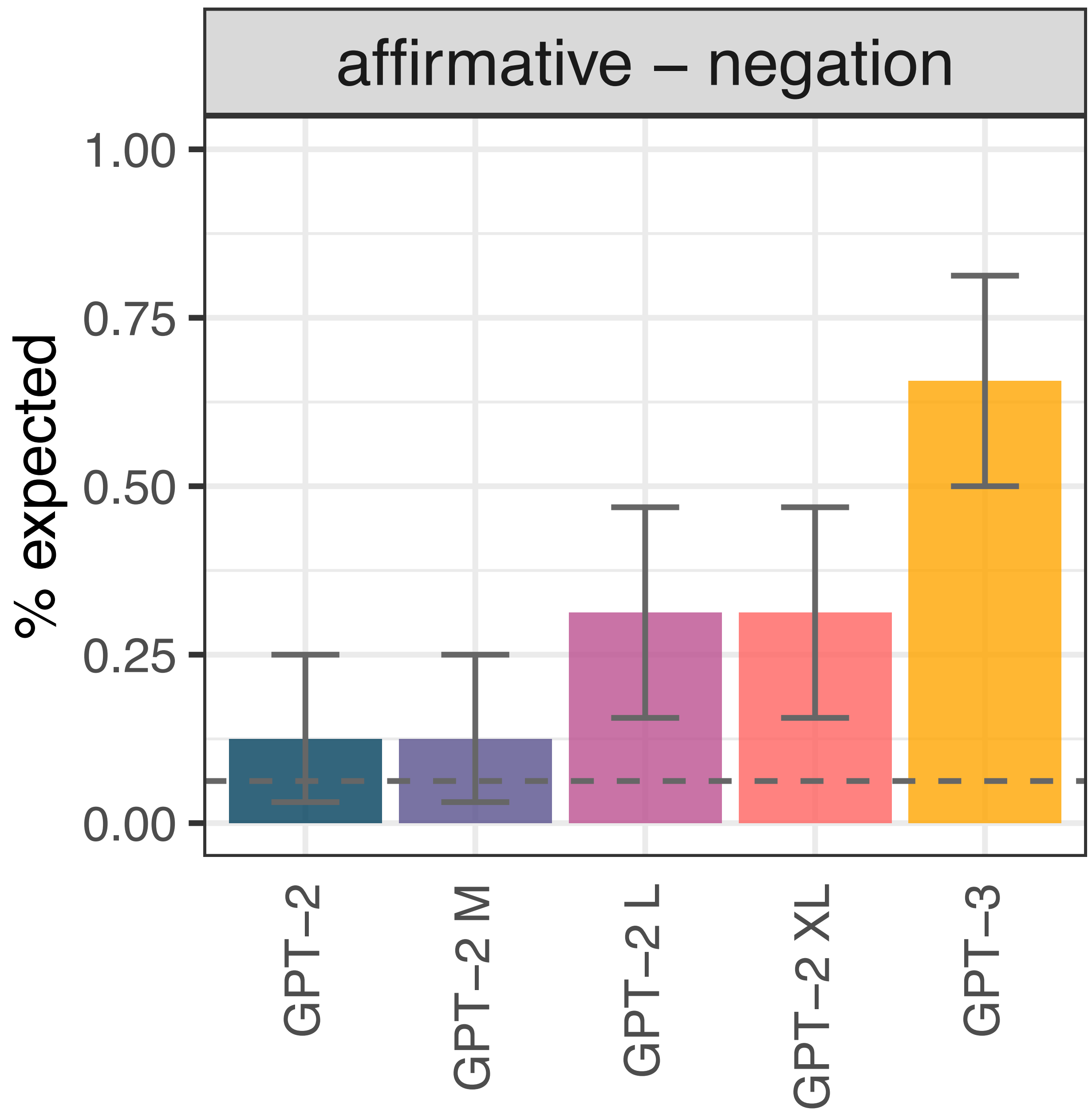
Evaluating systematicity

- **All orderings and combinations** of sentential operators and indefinite noun phrases

Mary found a shirt at the store but she didn't find a hat.
Mary found a hat at the store but she didn't find a shirt.
Mary didn't find a shirt at the store but she found a hat.
Mary didn't find a hat at the store but she found a shirt.

- Measure whether the model predictions are **as expected for all four combinations** for a specific item

Results: Systematicity



Conclusions

- Large-scale language models (especially GPT-3) are **to some extent** sensitive to interactions between sentential operators and indefinite noun phrases
- All models **lack systematicity** in their behavior, suggesting that their behavior deviates from human behavior
- Considering the size of the model and the training corpus of GPT-3, it seems unlikely that training even bigger models on even more data is going to lead to the expected behavior

General conclusions

- Large pre-trained LMs (especially more recent ones) exhibit to some extent pragmatic behavior
  - They can predict context-sensitive **scalar inferences** in many cases
  - They can predict **presuppositions** in many cases
  - They are often sensitive to whether sentential operators introduce **discourse entities**

- BUT: most behavior seems to be driven by **heuristics** and **lacks the systematicity** that we observe in humans thank you!

Collaborators:

Judith Degen

Tal Linzen

Yuxing Chen

Alex Warstadt

Alicia Parrish

Sam Bowman