UNIVERSITÄT
DES
SAARLANDES

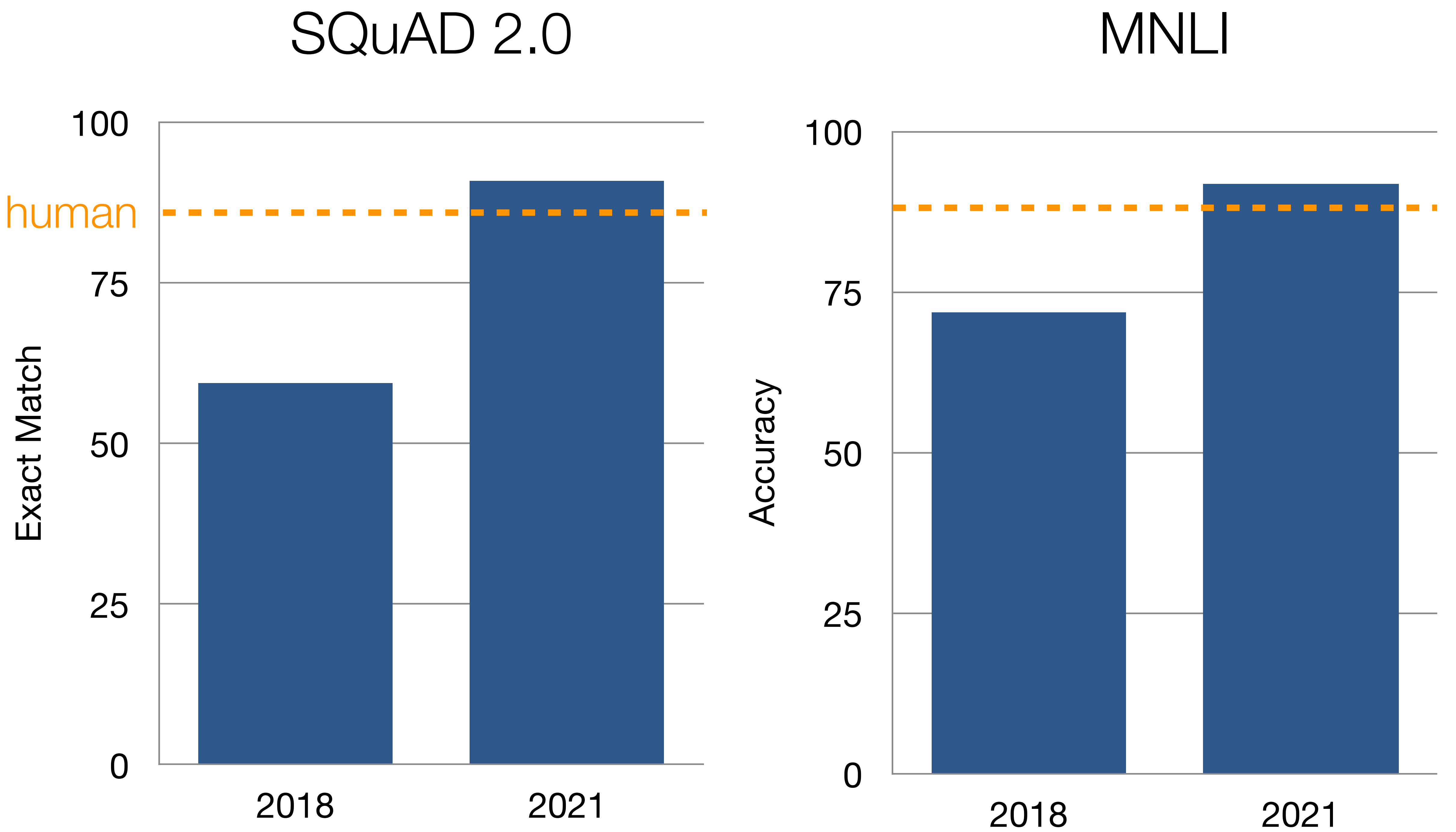
Entity tracking in pre-trained language models

November 4, 2022

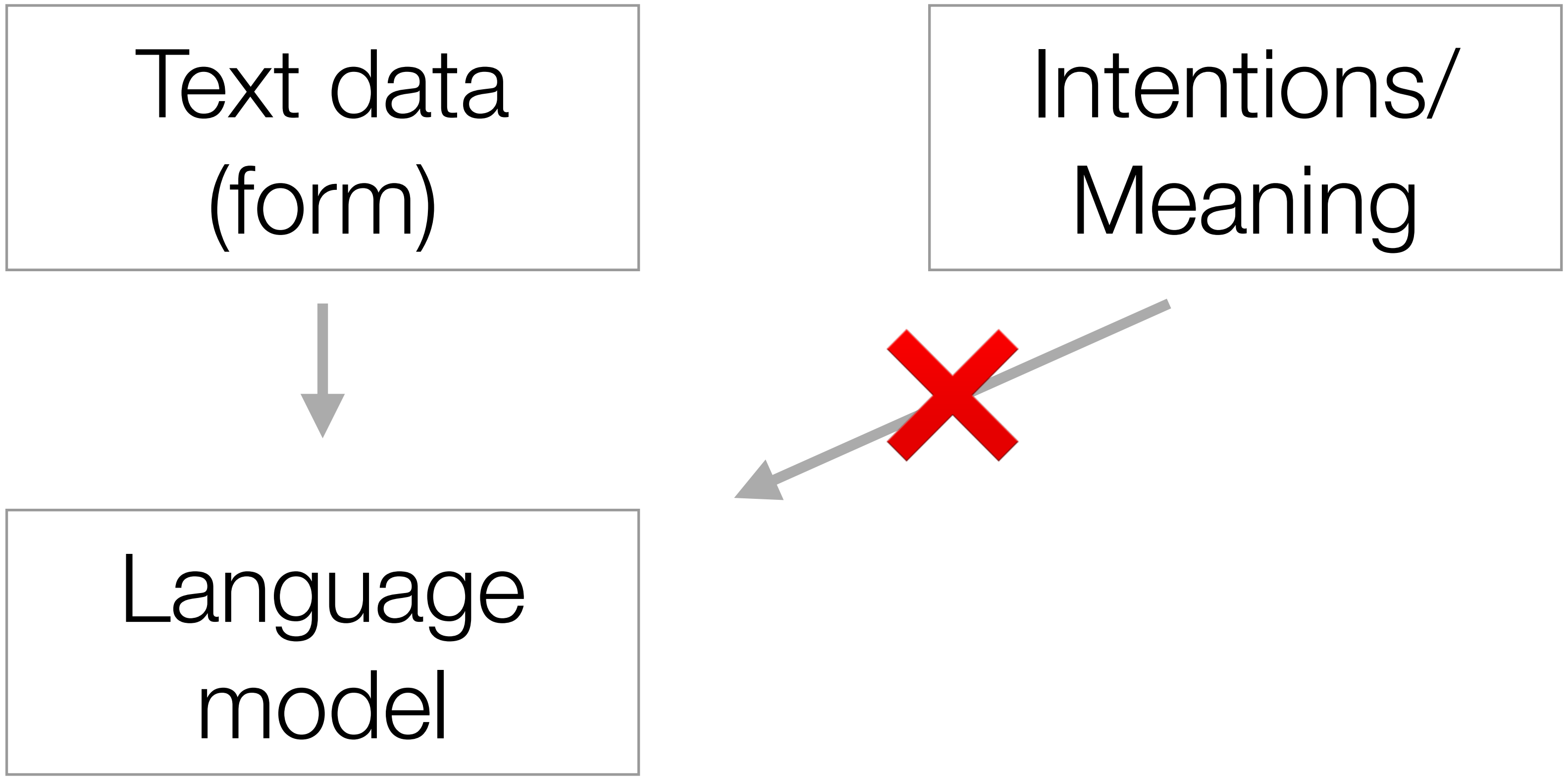
Sebastian Schuster
Computer Science and Computational Linguistics
Saarland University

**Contains some preliminary results,
please do not cite without permission!**

Progress in NLU



But do pre-trained language models understand language?



Bender & Koller, 2020; Bender et al. 2021

The pragmatist view

- “[…] what matters is that the agent be disposed to **use language in the right way.** This might include dispositions toward inference or reasoning patterns, appropriate conversational moves, and so on. Crucially, **the relevant verbal abilities constitute understanding.**” (Potts et al. 2021)

- Language models **may** learn these relevant verbal abilities from a lot of text

- For assessment, we need behavioral experiments!

Behavioral experiments at other linguistic levels

- Subject-verb agreement task:
  - Context: The keys to the cabinet ____
  - Examine the probabilities of the language model:
    Is P(are | context) > P(is | context) ?

Can we use similar experiments to assess semantic and discourse abilities of language models?

e.g., Linzen et al. 2016; Marvin and Linzen (2018)

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

“Consider a device designed to read a text in some natural language, interpret it, and store the content in some manner, say, for the purpose of being able to answer questions about it. To accomplish this task, the machine […] has to be able to build a file **that consists of records of-all the individuals**, that is events, objects, etc., **mentioned in the text**, and, for each individual, record whatever is said about it.”

Karttunen, 1976

Processing of discourse entities

1
John
owns(2)
gave(3)-to(2)

2
🐶
is-owned-by(1)
named-Spot
was-given(2)-by(1)

3
bandana
has-pumpkins
was-given-to(2)-by(1)

John owns a dog.
His name is Spot.
Because it is Halloween,
John got him a bandana with pumpkins.

Heim (1982)

Processing of discourse entities

1
John
owns(2)
gave(3)-to(2)

2
🐶
is-owned-by(1)
named-Spot
was-given(2)-by(1)

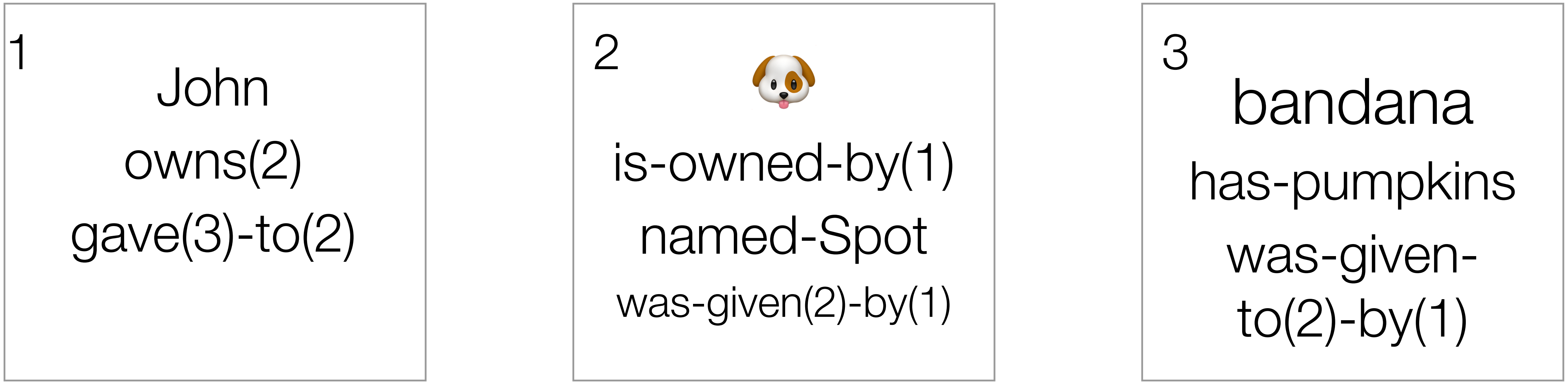
3
bandana
has-pumpkins
was-given-to(2)-by(1)

John owns a dog.
His name is Spot.
Because it is Halloween,
John got him a bandana with pumpkins.

Identifying when noun phrases introduce new entities.

Heim (1982)

Processing of discourse entities



1
John
owns(2)
gave(3)-to(2)

2
🐶
is-owned-by(1)
named-Spot
was-given(2)-by(1)

3
bandana
has-pumpkins
was-given-to(2)-by(1)

John owns a dog.
His name is Spot.
Because it is Halloween,
John got him a bandana with pumpkins.

Coreference resolution.

Heim (1982)

Processing of discourse entities

1

John

owns(2)

gave(3)-to(2)

2

🐶

is-owned-by(1)

named-Spot

was-given(2)-by(1)

3

bandana

has-pumpkins

was-given-to(2)-by(1)

John owns a dog.
His name is Spot.
Because it is Halloween,
John got him a bandana with pumpkins.

Updating information about entities as discourse unfolds.

Heim (1982)

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

1. Can LMs identify when noun phrases introduce discourse entities?
2. Can LMs resolve co-reference?
3. Do LMs update information about entities as the discourse unfolds?

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

1. Can LMs identify when noun phrases introduce discourse entities?

Are language models **sensitive to contextual factors** that modulate whether an **indefinite noun phrase introduces a discourse entity**?

Schuster and Linzen (2022),
see also Loáiciga et al. (2022)

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

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

Expected language model behavior

	ent-referential It has a red collar	**evt-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

Expected language model behavior

	ent-referential It has a red collar	**evt-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

$$\frac{P(\text{ent-ref} \mid A)}{P(\text{evt-ref} \mid A)} > \frac{P(\text{ent-ref} \mid B)}{P(\text{evt-ref} \mid B)}$$

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.
  - ***know* vs. *doubt***
    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:

  - **affirmative vs. *want***
    A: John **owns** a dog.
    B: John **wants to own** a dog.

  - ***managed to* vs. *failed to***
    A: John **managed to** adopt a dog.
    B: John **failed to** adopt a dog.

16 hand-written items —> 64 pairs

Language models

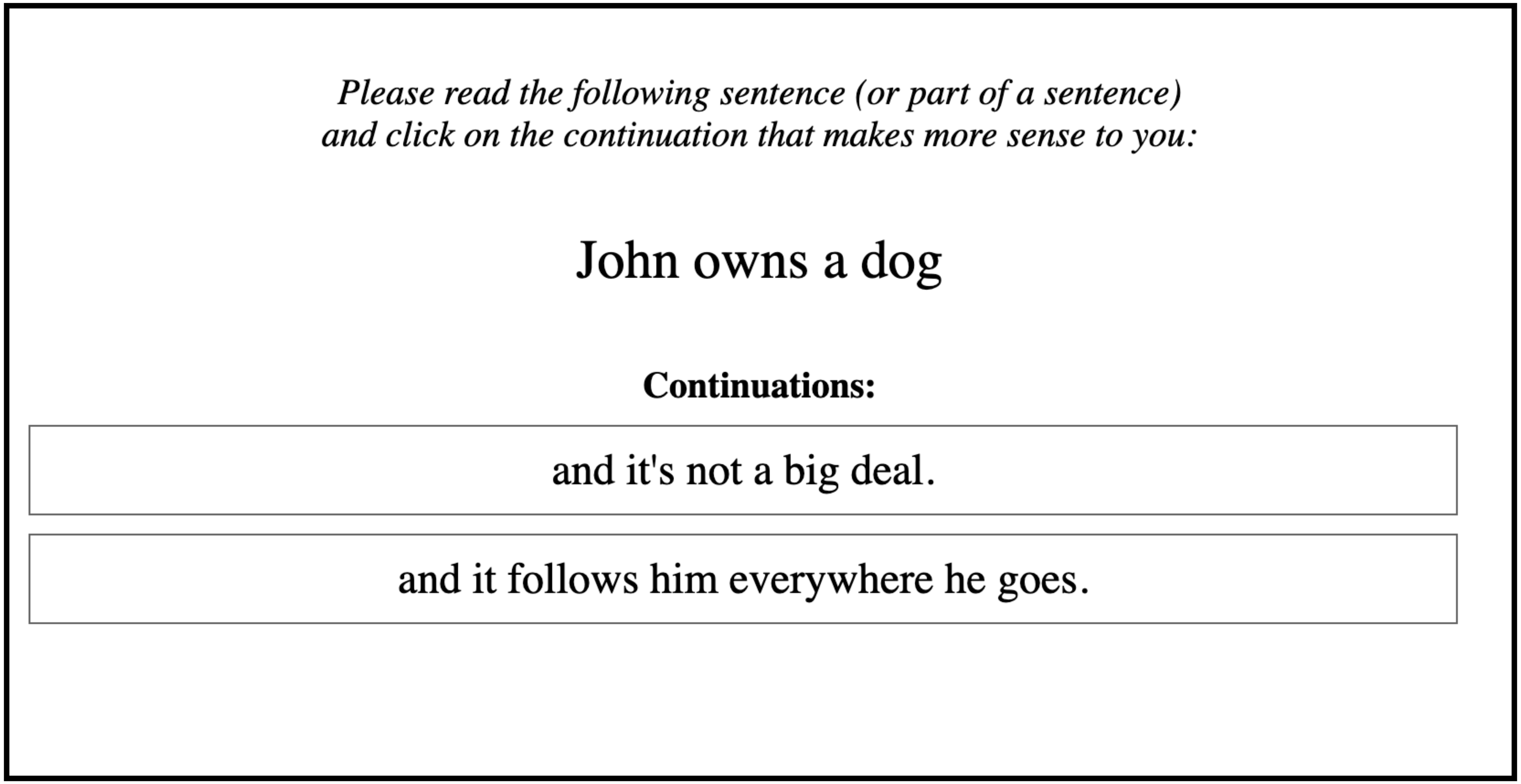
- GPT-2 in various sizes:
  - **GPT-2**: 117M parameters
  - **GPT-2-medium**: 345M parameters
  - **GPT-2-large**: 762M parameters
  - **GPT-2-xl**: 1542M parameters

trained on
~ 8 billion tokens

- **GPT-3 (davinci-001)**: 175B parameters?

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

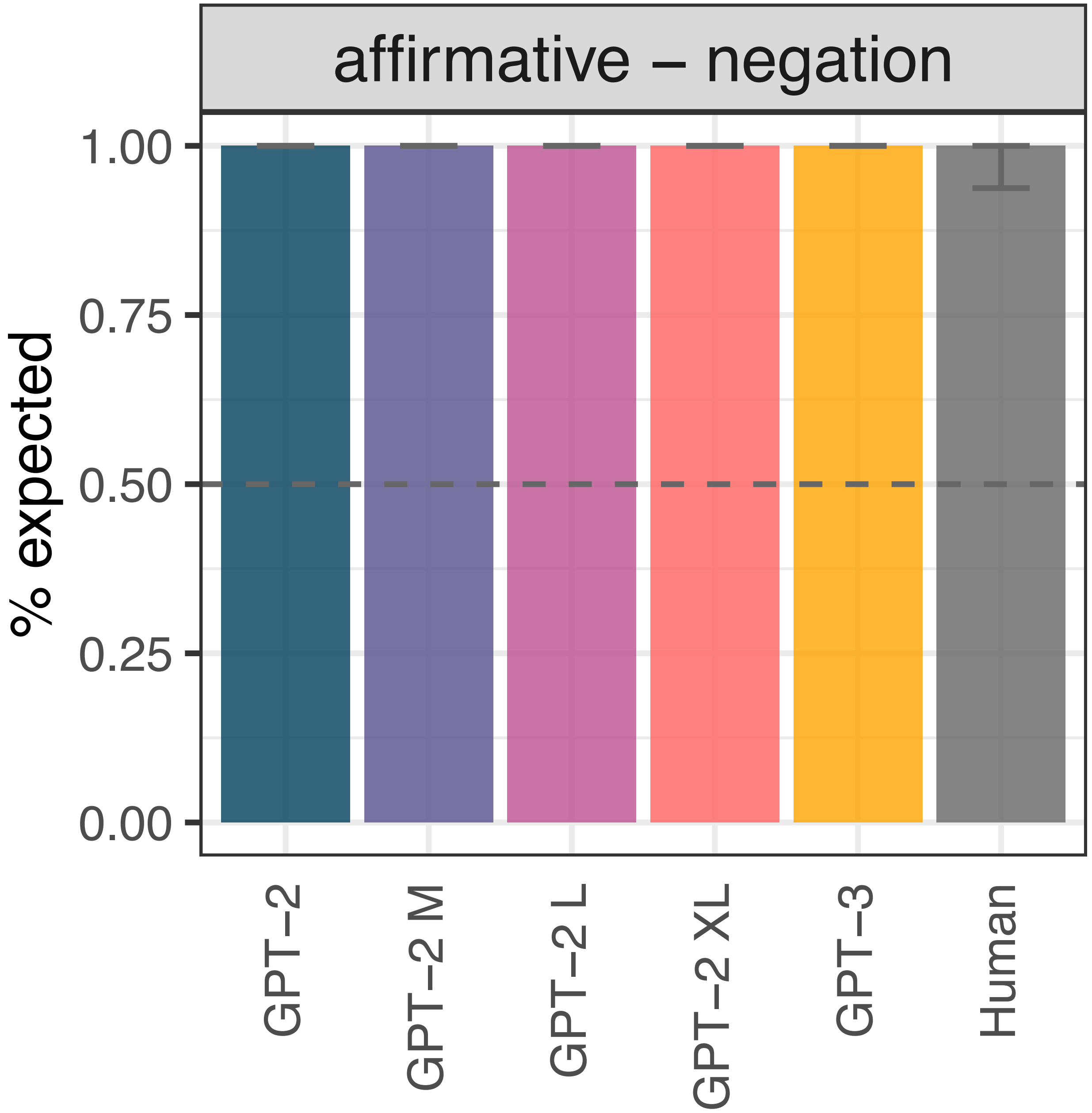
Continuations:

and it's not a big deal.

and it follows him everywhere he goes.

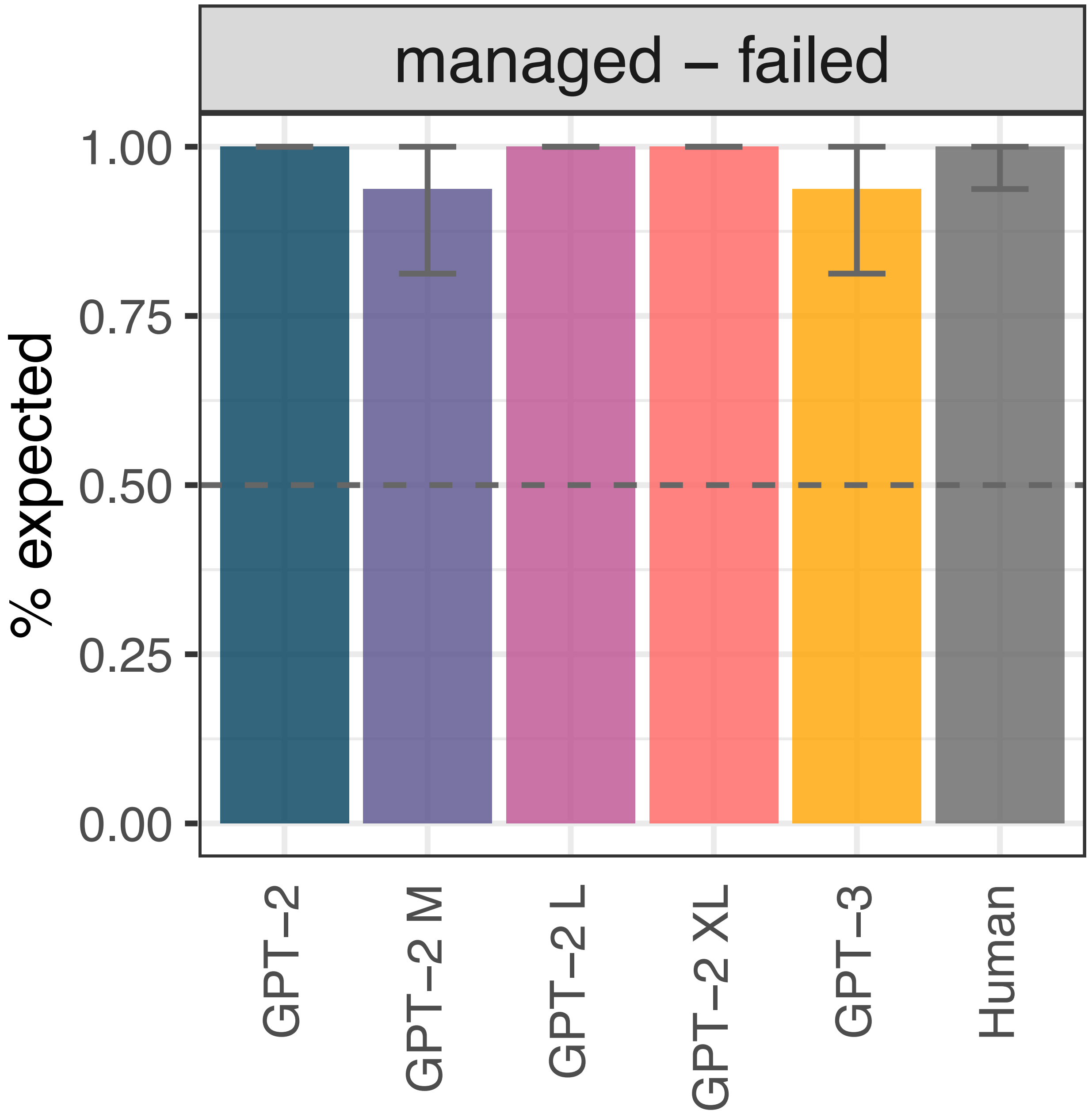
Results

$$\frac{P(\text{ent-ref} \mid A)}{P(\text{evt-ref} \mid A)} > \frac{P(\text{ent-ref} \mid B)}{P(\text{evt-ref} \mid B)}$$



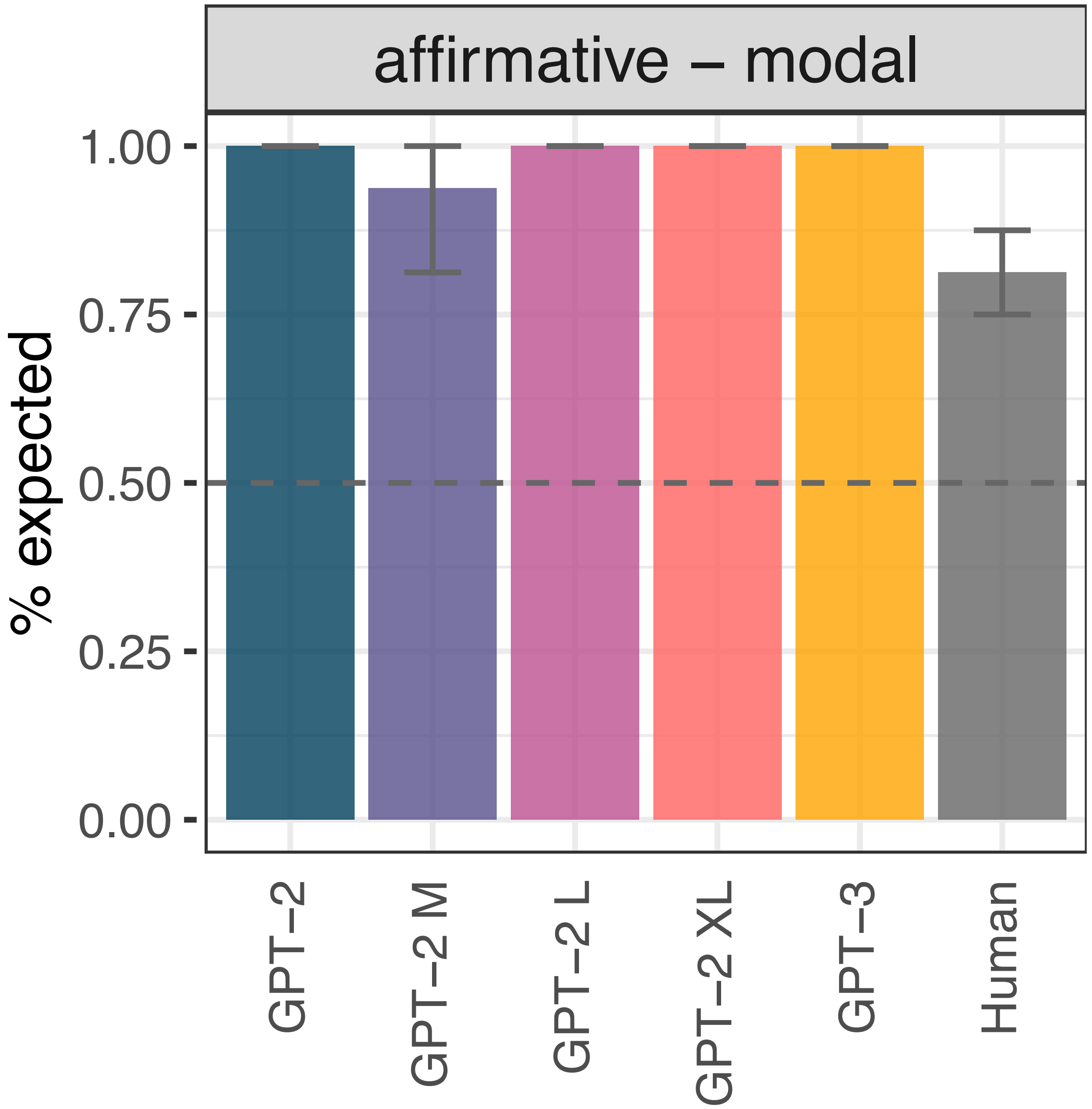
Results

$$\frac{P(\text{ent-ref} \mid A)}{P(\text{evt-ref} \mid A)} > \frac{P(\text{ent-ref} \mid B)}{P(\text{evt-ref} \mid B)}$$



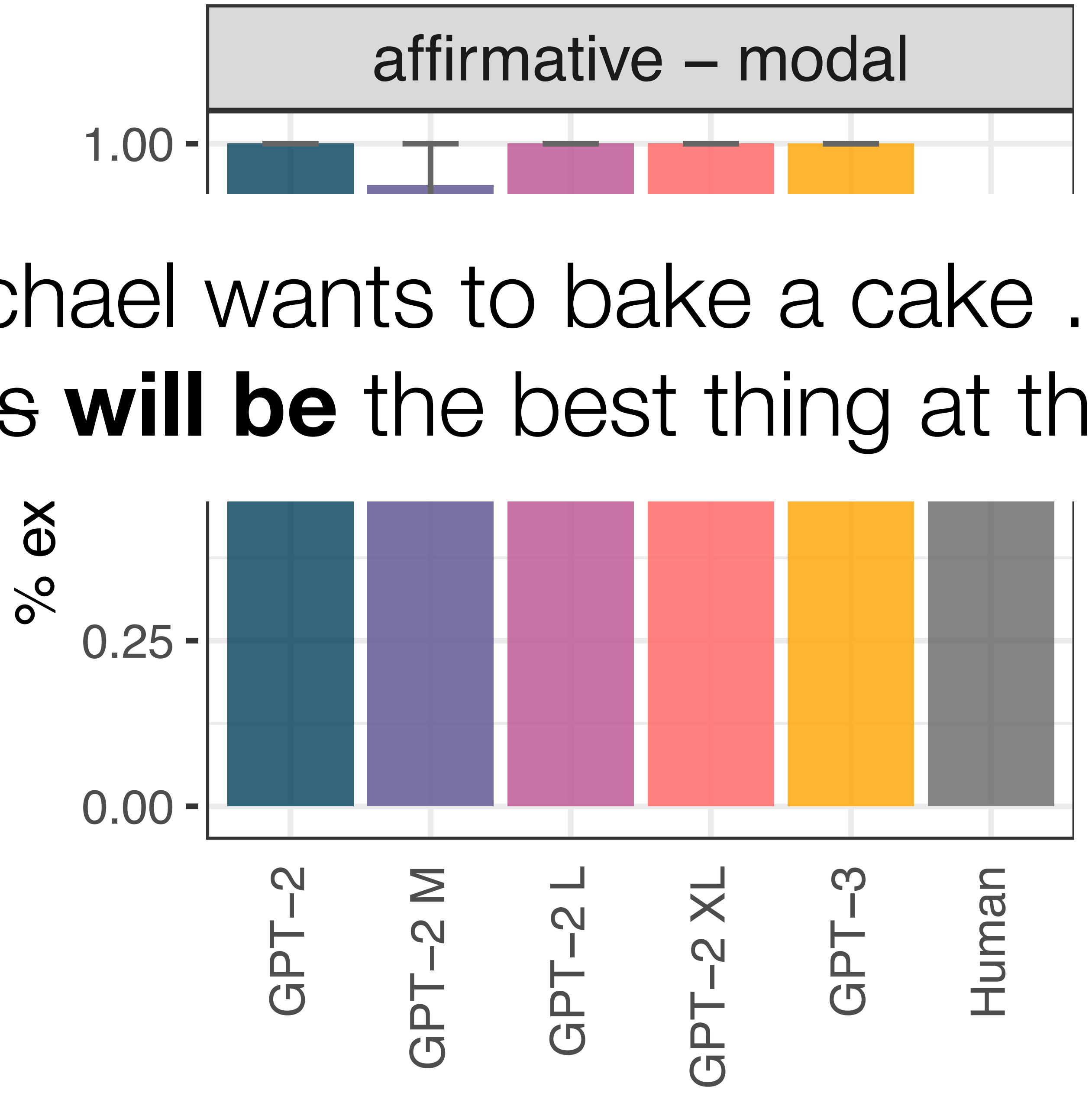
Results

$$\frac{P(\text{ent-ref} \mid A)}{P(\text{evt-ref} \mid A)} > \frac{P(\text{ent-ref} \mid B)}{P(\text{evt-ref} \mid B)}$$



Results

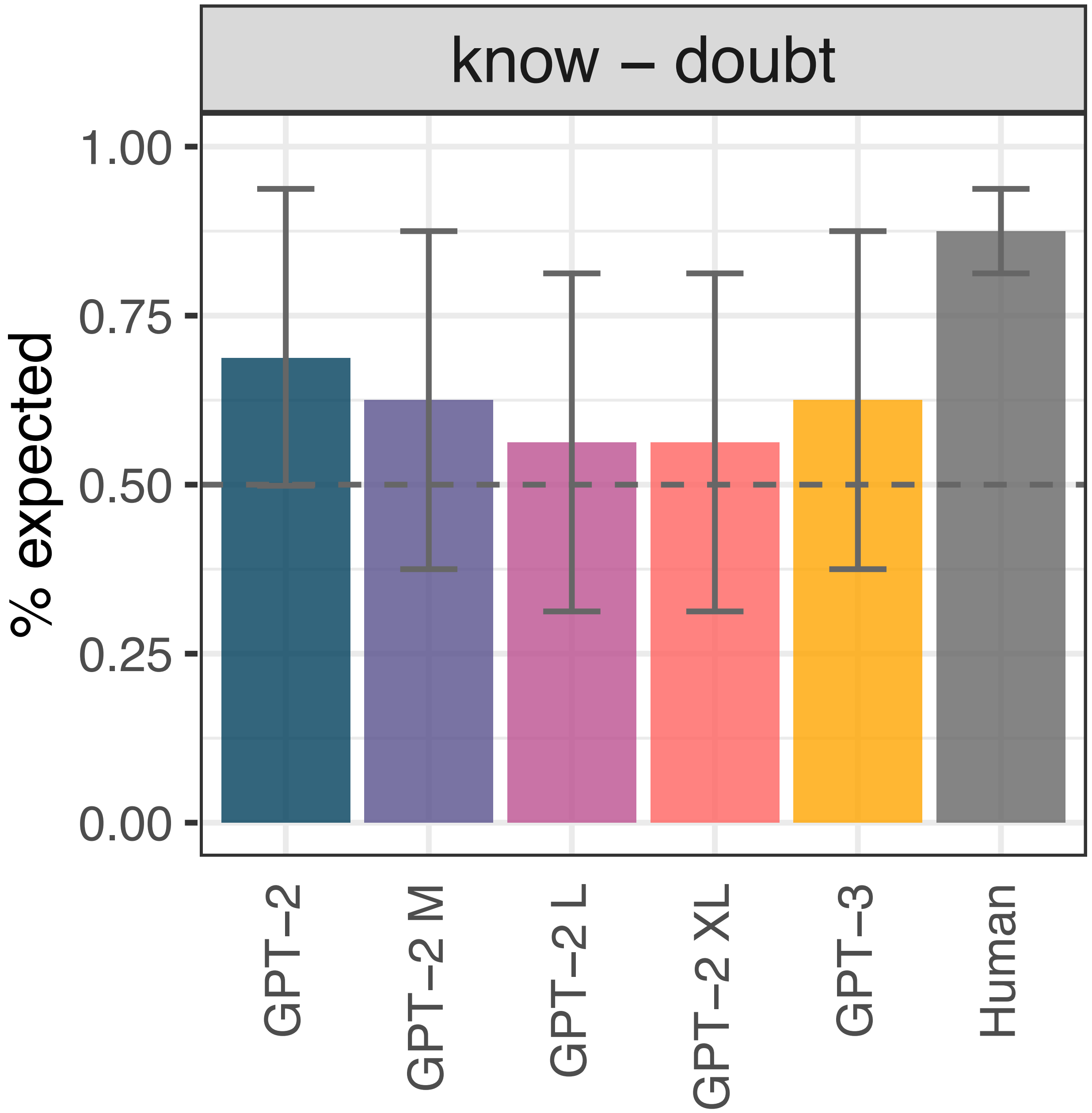
$$\frac{P(\text{ent-ref} \mid A)}{P(\text{evt-ref} \mid A)} > \frac{P(\text{ent-ref} \mid B)}{P(\text{evt-ref} \mid B)}$$



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

Results

$$\frac{P(\text{ent-ref} \mid A)}{P(\text{evt-ref} \mid A)} > \frac{P(\text{ent-ref} \mid B)}{P(\text{evt-ref} \mid B)}$$



Interim conclusions

- Human preferences for continuations are largely in line with patterns predicted by most linguistic theories
- Except for the *know* vs *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”)

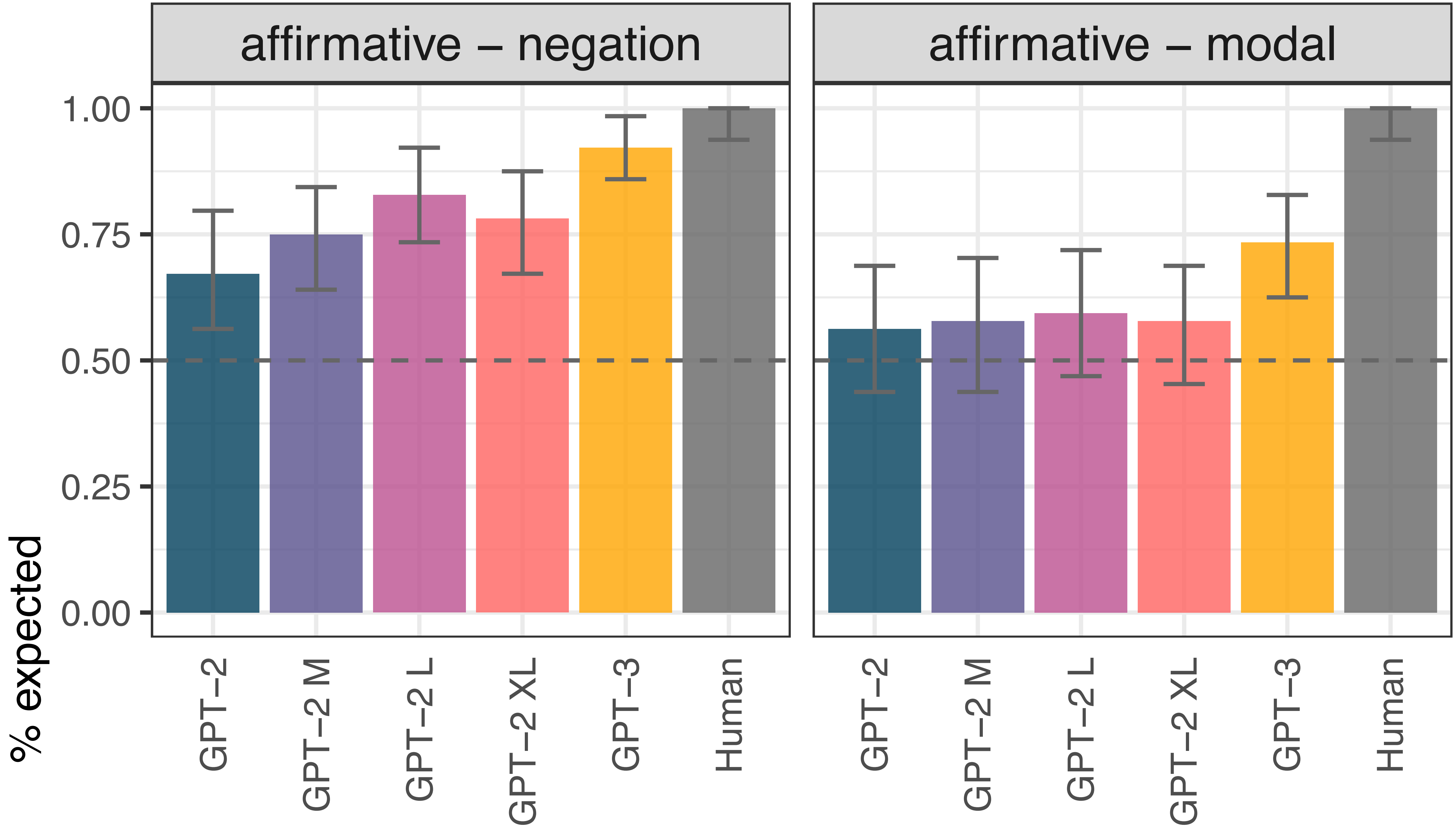
- Non-coreferential continuations:

  - 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**

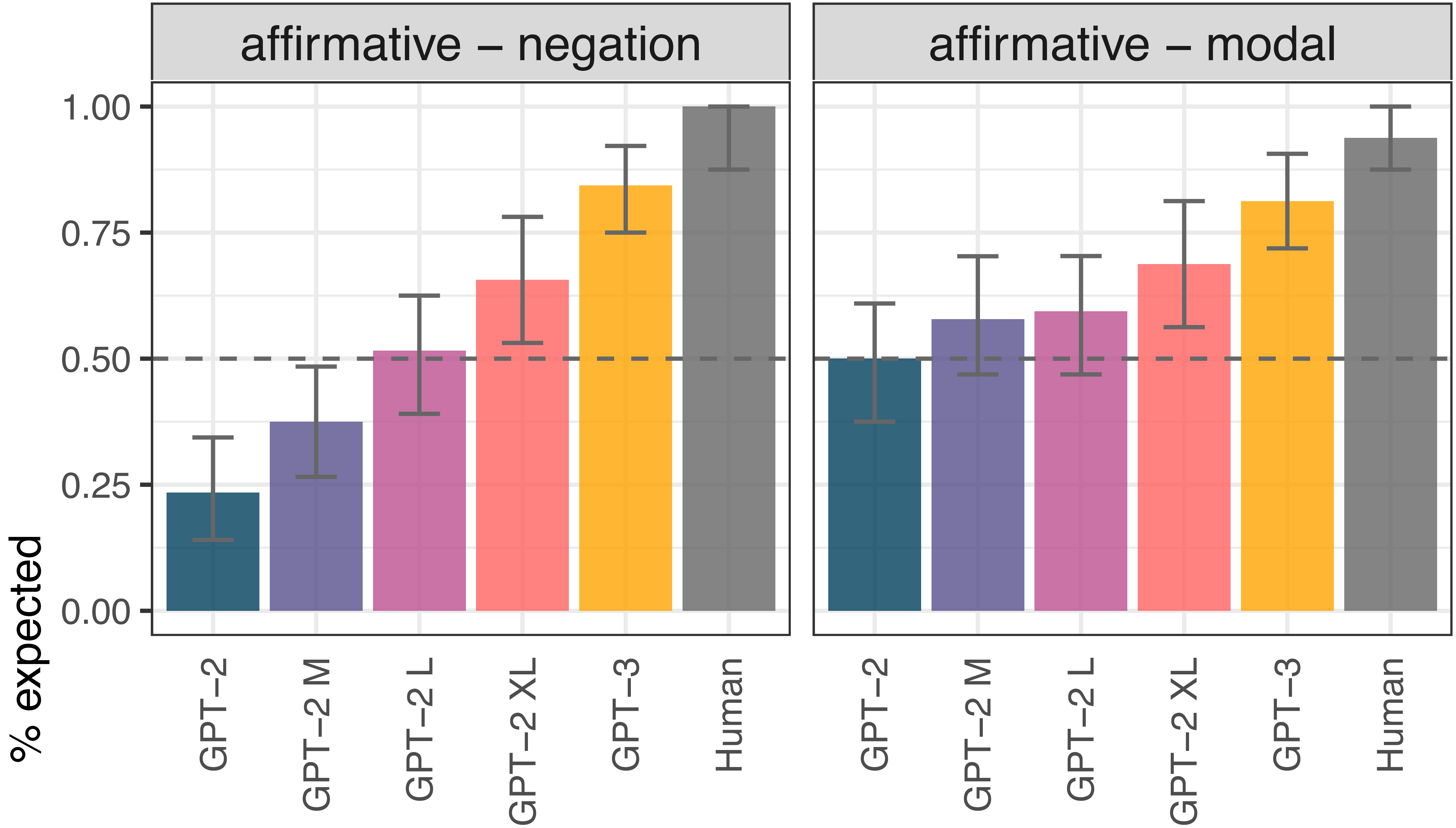
P("The **shirt** was blue") > P("The **hat** was blue")



Results: Non-coreferential continuations

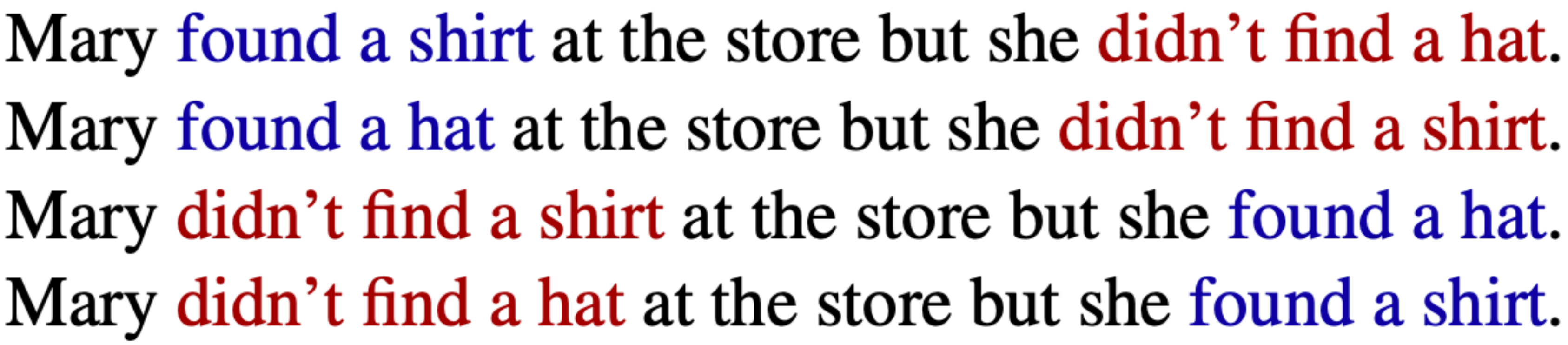
Mary **found a shirt** at the store but she **didn't find a hat**

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



Evaluating systematicity

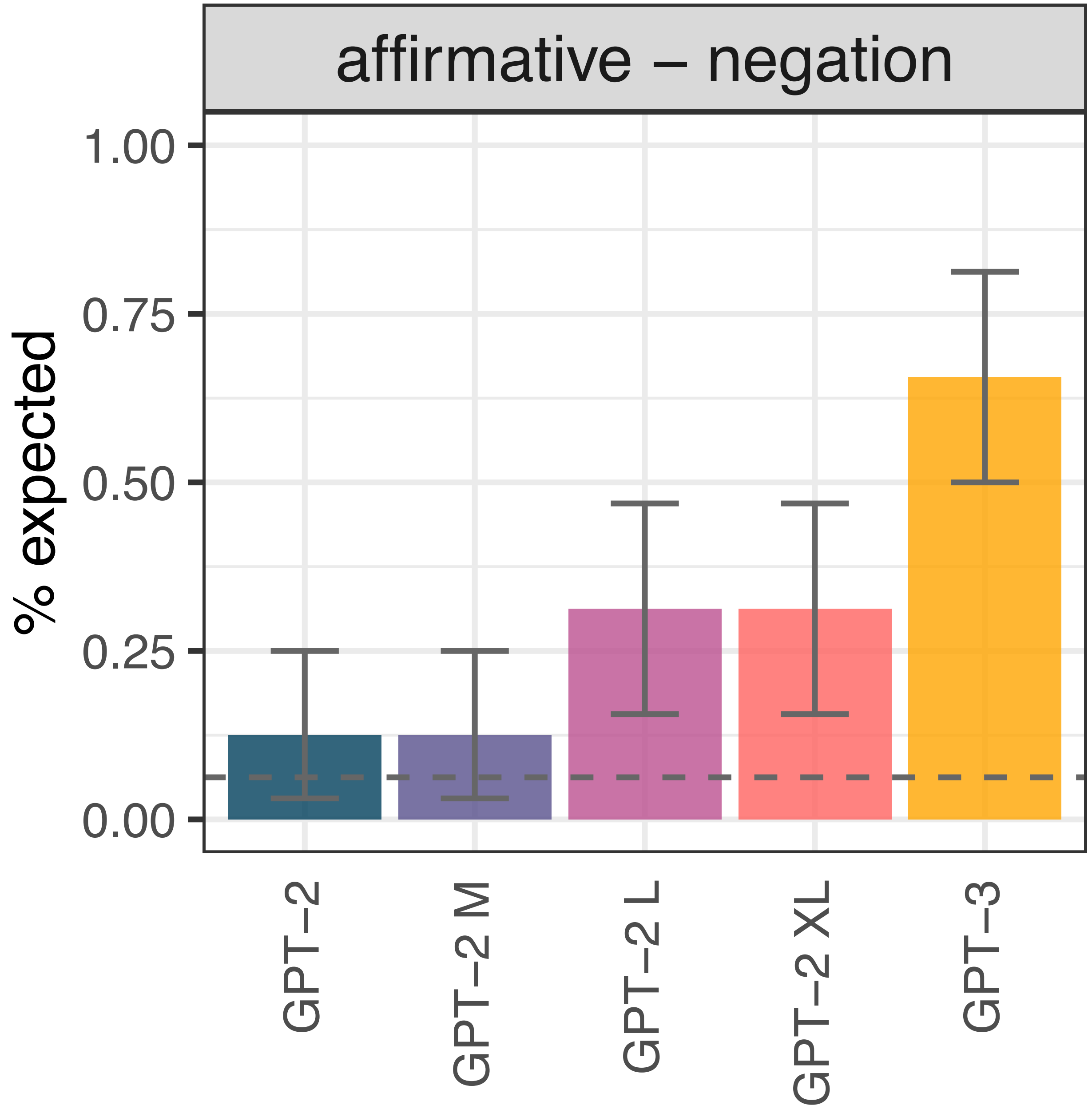
- **All orderings and combinations** of sentential operators and indefinite noun phrases



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



Likely continuations

- The previous experiments used **specific continuations**
- It may be that both the expected and the unexpected continuation are very unlikely and thus would almost **never be generated by an LM** in practice
- How often do LMs refer back to noun phrases that do not introduce discourse entities in likely continuations?

Manual analysis of continuations

- Sample continuations from GPT-2 XL and GPT-3 using prompts such as

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

- A linguistics graduate student annotated each continuation for whether it contained a referring expression that referred back to the NP that introduced or did not introduce a discourse entity

Manual analysis of continuations

John owns a dog but he doesn't own a cat. The …

… dog has a red collar. … cat has a red collar.

Model	Discourse entity introducing NP	Non-discourse entity introducing NP
GPT-2 XL	43.8	22.3
GPT-3	52.3	21.1

1 in 5 likely continuations contained a referring expression that referred back to an NP that did not introduce a discourse entity!

Manual analysis of continuations

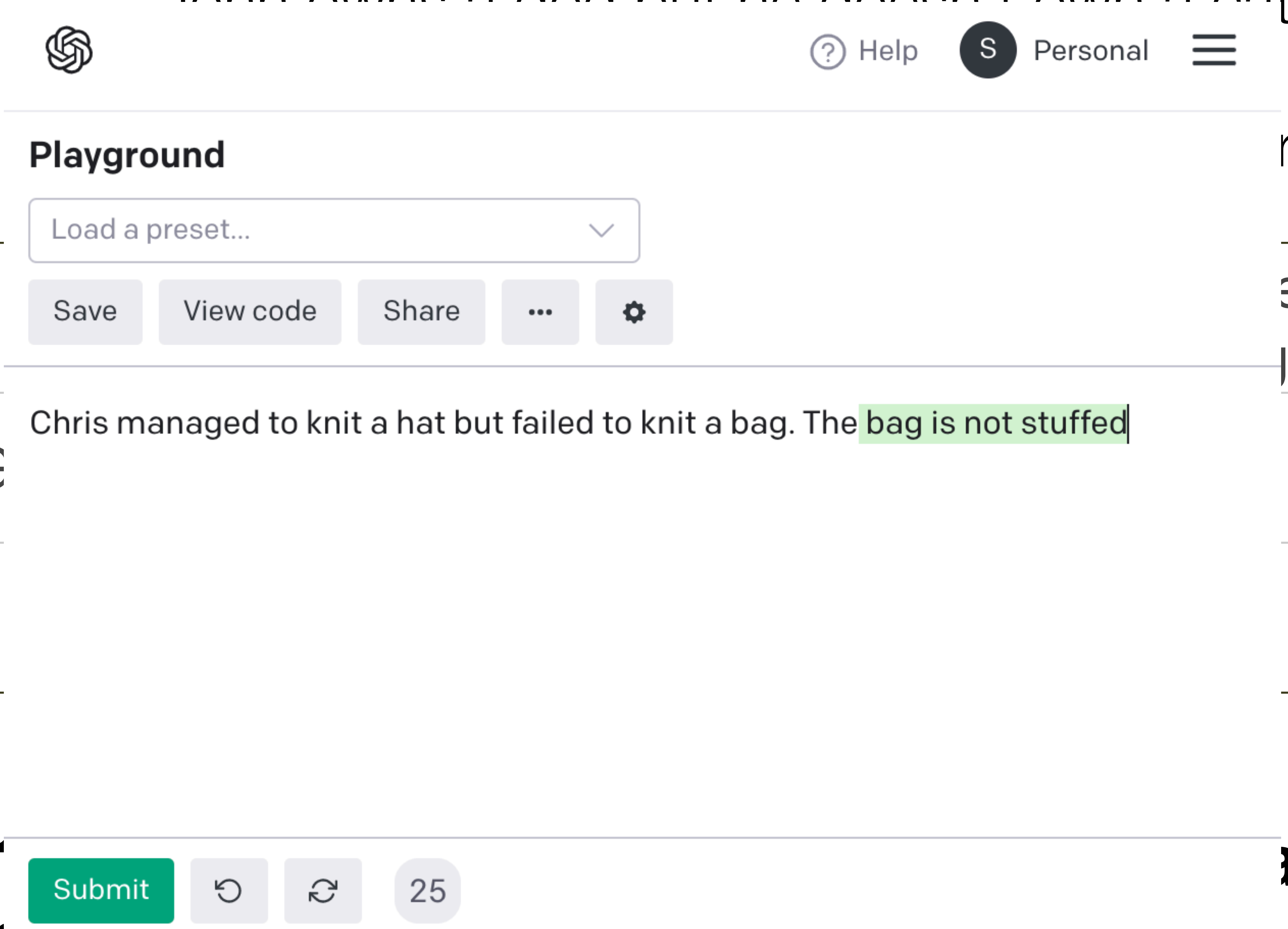
John owns a dog but he doesn't own a cat. The …

red collar.

e entity

NP

G



Playground

Chris managed to knit a hat but failed to knit a bag. The bag is not stuffed

referring expression that referred back to
an NP that did not introduce a discourse entity!

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

- There are effects of model size, so potentially scaling up could work. However, this may require order of magnitudes of more data.

Methodological conclusions

- Zooming into specific phenomena and comparing to human behavior can reveal systematic limitations of current models

- At first glance, models often seem to behave as expected, so it's important to evaluate from multiple angles

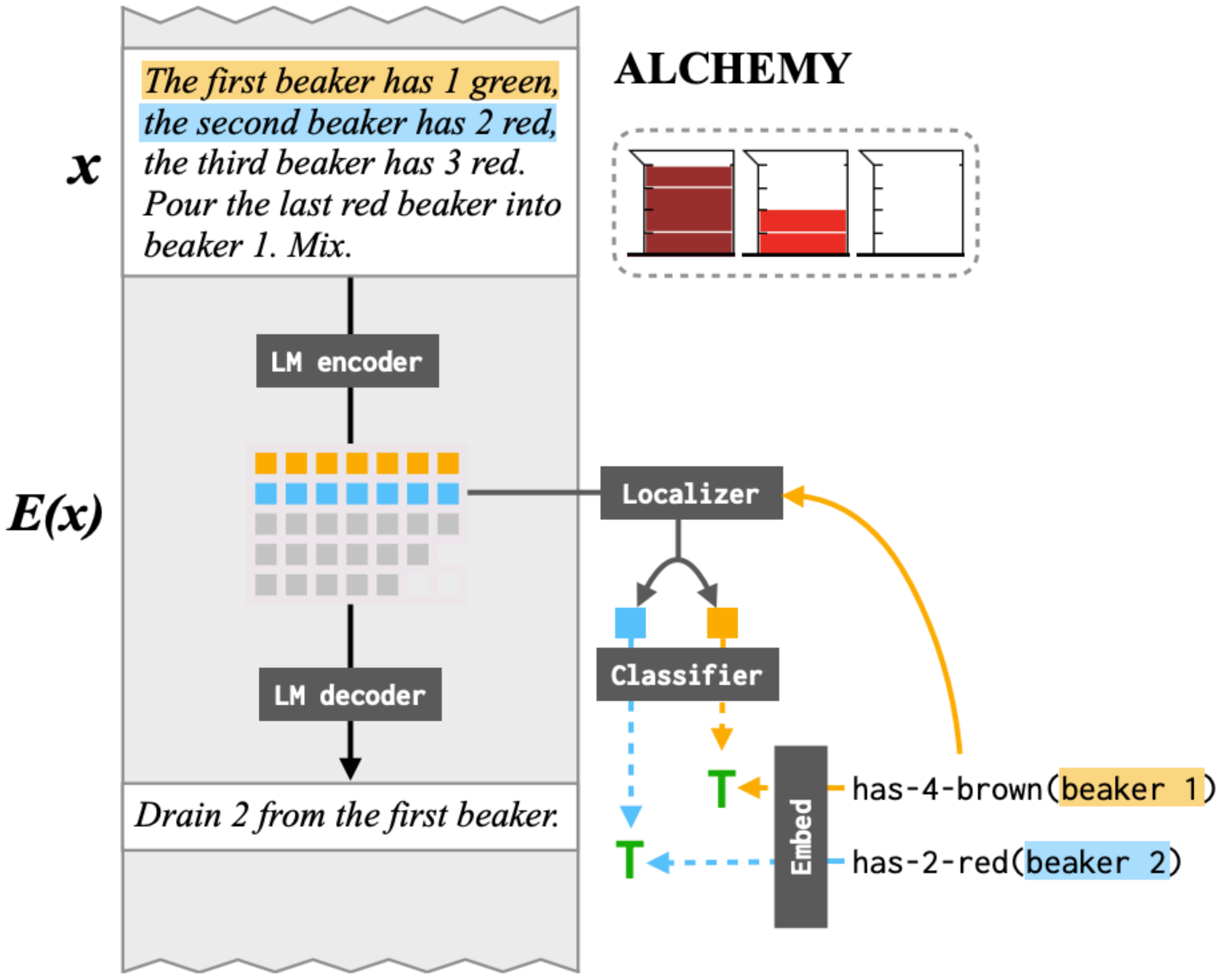
- Carefully constructed behavioral experiments complement benchmarks to track progress

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

1. Can LMs identify when noun phrases introduce discourse entities?
2. Can LMs resolve co-reference?
3. **Do LMs update information about entities as the discourse unfolds?**

joint work w/ Najoung Kim

Evidence from Probing


ALCHEMY
x
The first beaker has 1 green, the second beaker has 2 red, the third beaker has 3 red. Pour the last red beaker into beaker 1. Mix.
LM encoder
E(x)
Localizer
Classifier
LM decoder
Drain 2 from the first beaker.
T
T
Embed
has-4-brown(beaker 1)
has-2-red(beaker 2)

Overall accuracy: 76.5%

Accuracy for non-trivial cases: 3.1%

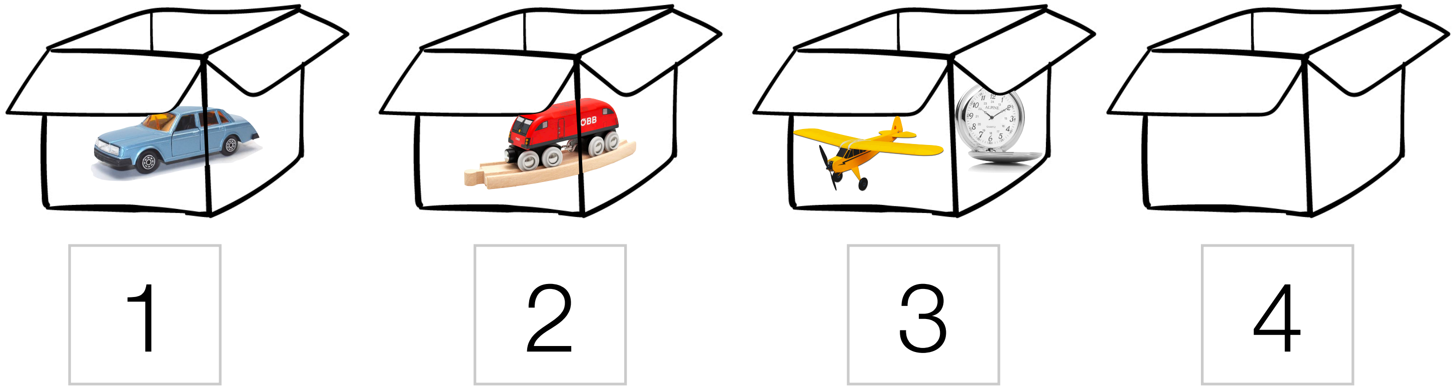
Li et al. (ACL 2021)

Do LMs update information about entities as the discourse unfolds?

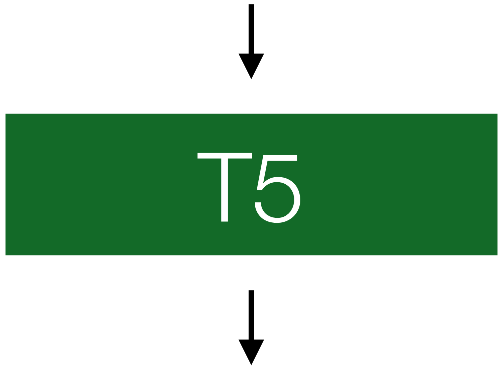
- Probing experiments make it difficult to disentangle how much the probe learns and how much is captured by the representation —> use behavioral experiments

- Crowdsourced dataset may contain a lot of biases that allow the model to seemingly correctly do the task while it is just learning patterns of the data —> generate highly controlled dataset

Setup

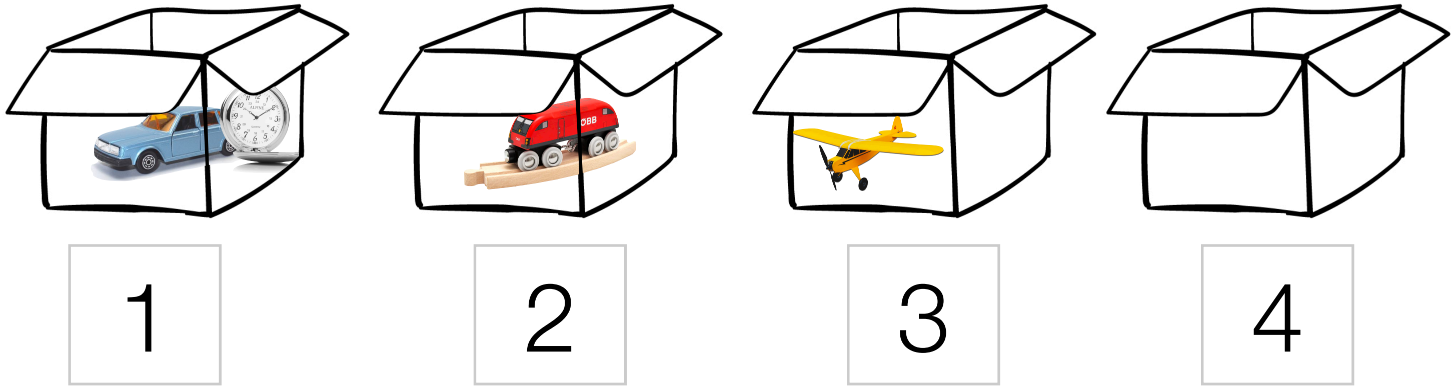


```
Box 1 contains the car,
Box 2 contains the train,
Box 3 contains the plane and the watch,
Box 4 is empty.
Box 1 [MASK1] .
```


T5

```
[MASK1] contains the car
```

Setup



```
Box 1 contains the car,
Box 2 contains the train,
Box 3 contains the plane and the watch,
Box 4 is empty.
Move the watch from Box 3 to Box 1.
Box 1 [MASK1] .
```

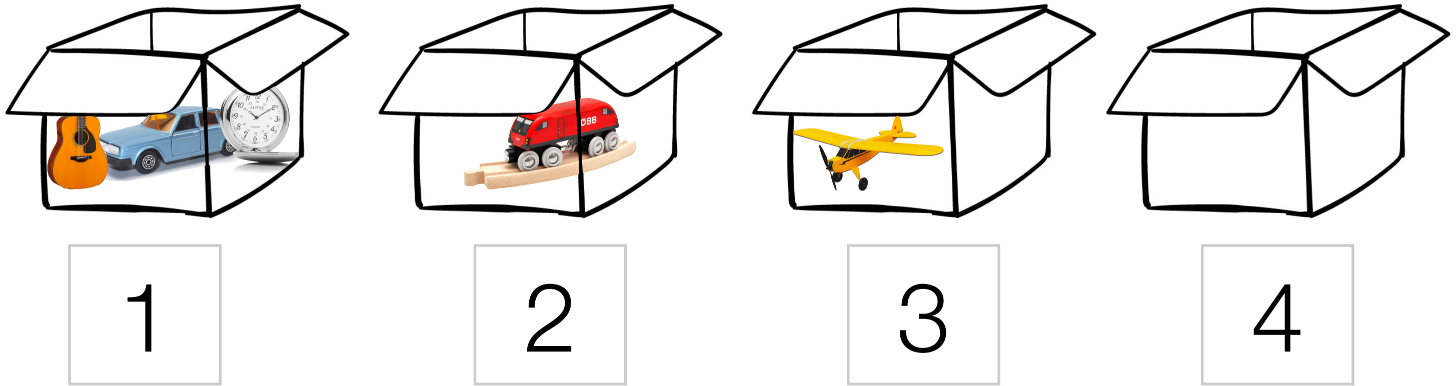
↓

T5

↓

[MASK1] contains the car **and the watch**

Setup



```
Box 1 contains the car,
Box 2 contains the train,
Box 3 contains the plane and the watch,
Box 4 is empty.
Move the watch from Box 3 to Box 1.
Add the guitar to Box 1.
Box 1 [MASK1] .
```

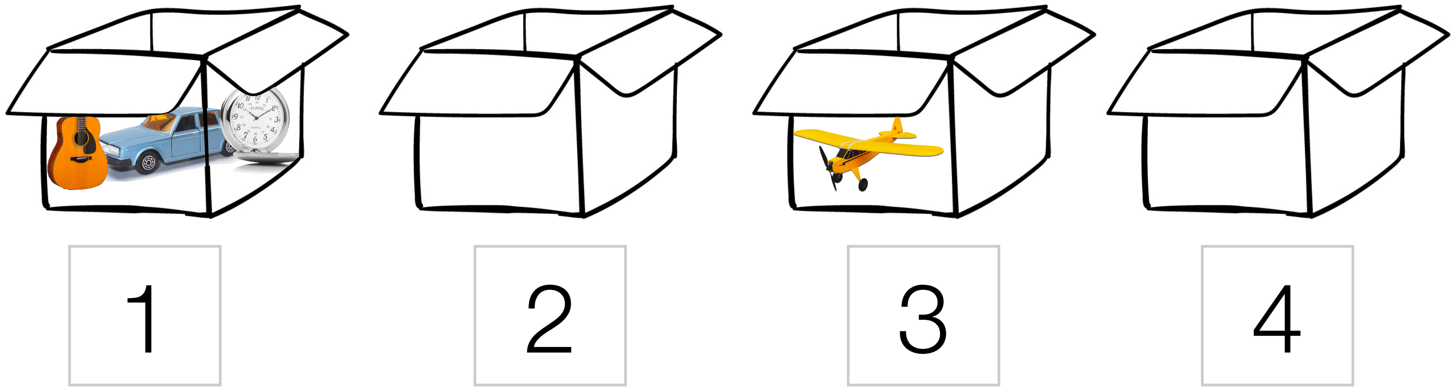
↓

T5

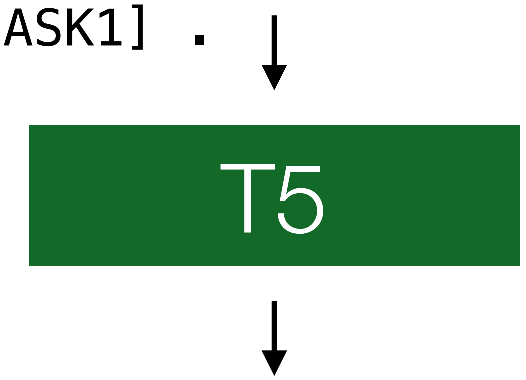
↓

`[MASK1] contains the car and the watch` **`and the guitar`**

Setup

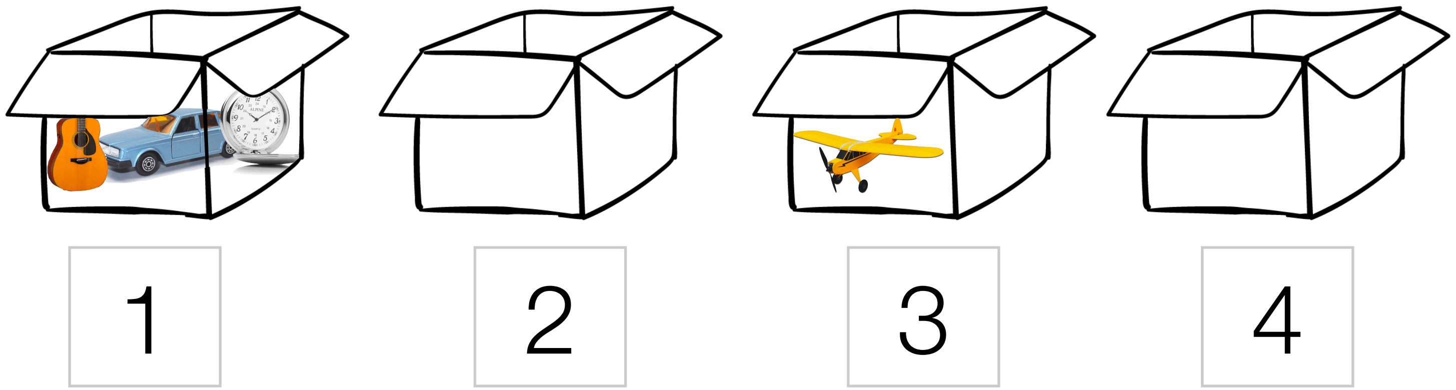


```
Box 1 contains the car,
Box 2 contains the train,
Box 3 contains the plane and the watch,
Box 4 is empty.
Move the watch from Box 3 to Box 1.
Add the guitar to Box 1.
Remove the train from Box 2.
Box 1 [MASK1] .
```

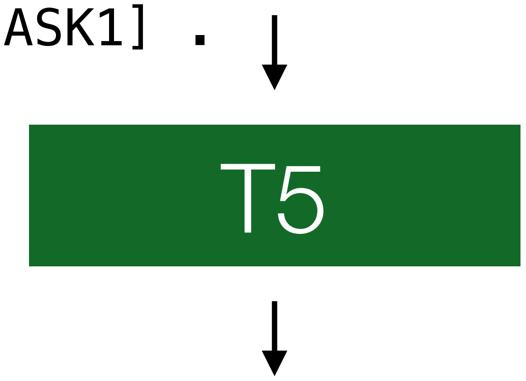

T5

```
[MASK1] contains the car and the watch and the guitar
```

Setup



Box 1 contains the car,
Box 2 contains the train,
Box 3 contains the plane and the watch,
Box 4 is empty.
Move the watch from Box 3 to Box 1.
Add the guitar to Box 1.
Remove the train from Box 2.
Box **2** [MASK1] .



[MASK1] is empty

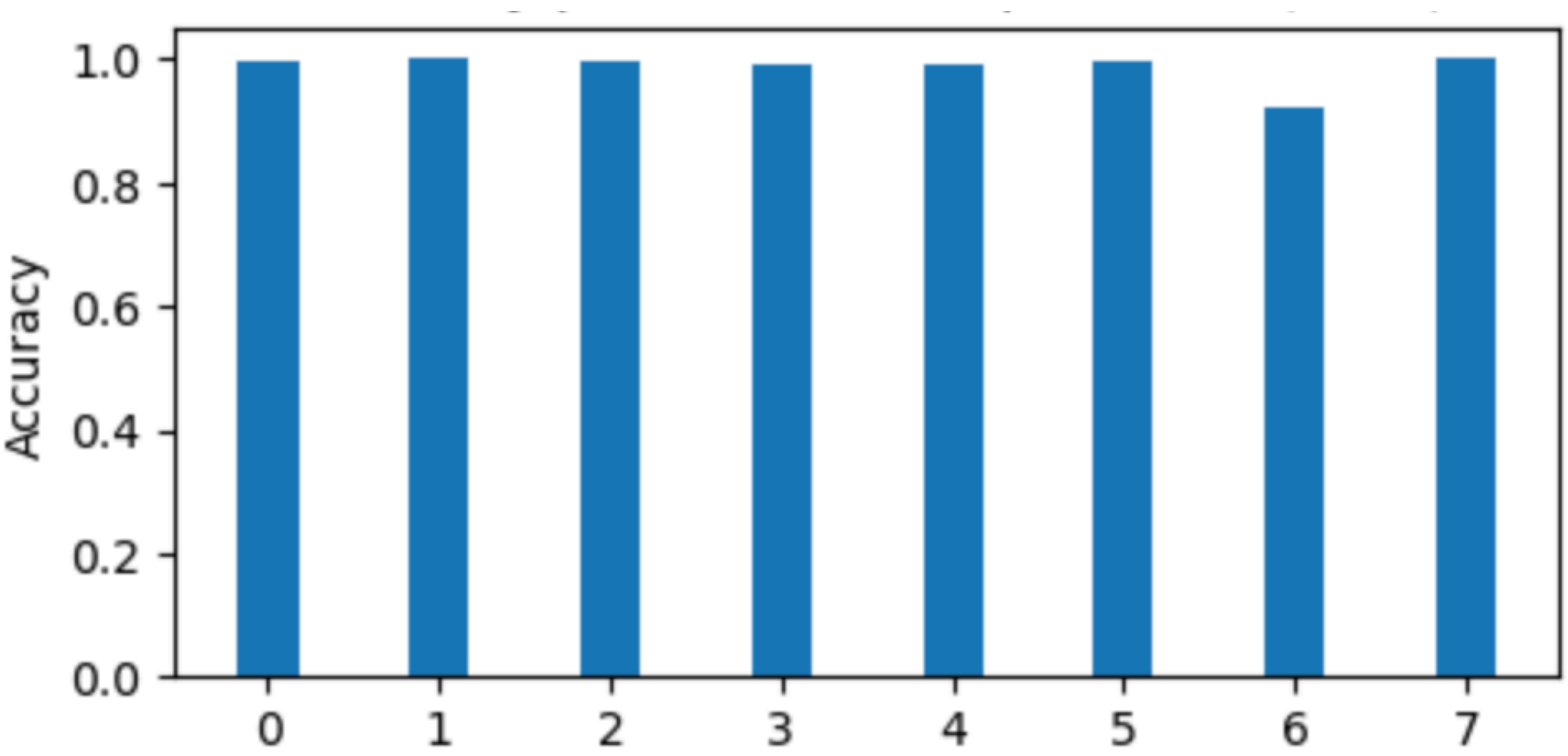
Dataset generation

- 7 boxes
- up to 9 objects per box
- Randomly sample initial states and operations (move, add, remove)
- Initial states are unique and “signature” does not overlap between train and test

Can T5 learn this task?

```
Box 1 contains the car,
Box 2 …
Move the watch from Box 3 to Box 1.
Add the guitar to Box 1.
Remove the train from Box 2.
Box 1 [MASK1] .
```

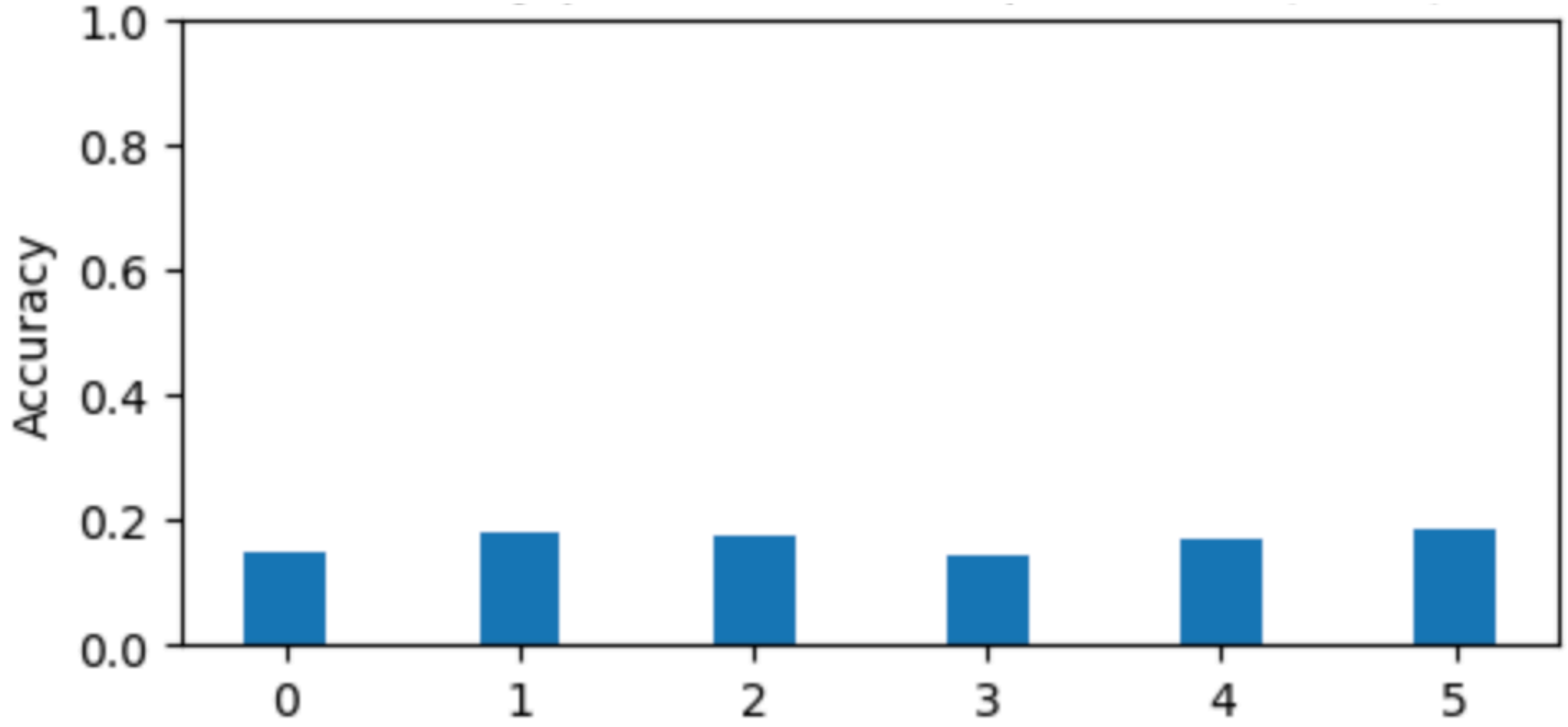
After 1 epoch of fine-tuning



Number of operations acted on box

Is fine-tuning doing all the work?

Randomly initialized model,
10 epochs of fine-tuning



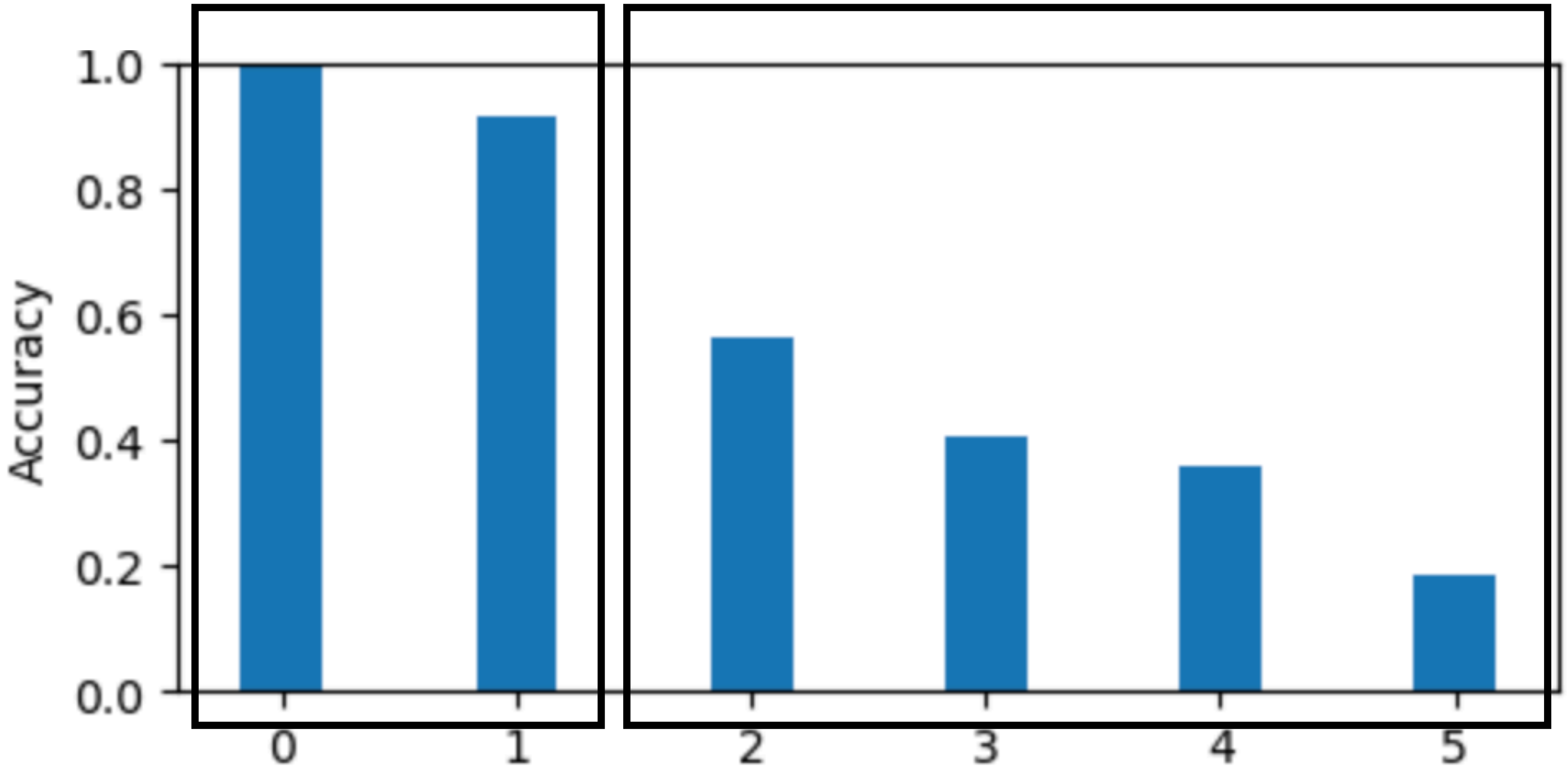
Number of operations acted on box

Is the model properly generalizing?

After 1 epoch of fine-tuning

In training data

generalization



Number of operations acted on box

Preliminary conclusions

- T5 can **learn** to track updates to entities as discourse unfolds (at least in this toy domain…)

- Randomly initialized models lack this ability — **pre-training does seem to contribute to this ability**

- **Generalization performance drops rapidly** and it remains unclear to what extent a LM is able to track updates without fine-tuning

- Stay tuned for zero-shot results!

Takeaways

1. Can LMs identify when noun phrases introduce discourse entities?

   Above chance but not as systematically as humans

2. Do LMs update information about entities as the discourse unfolds?

   They can learn to do it within a toy domain
   BUT: T5 doesn’t seem to generalize reliably and unclear whether they also do it spontaneously

Do language models understand language?

- Models trained only on next-word prediction don’t seem to be able to systematically track discourse entities
- Instruction fine-tuned models seem promising to improve results on a lot of tasks (e.g., Chung et al. 2022, Ruis et al. 2022)
- Challenge datasets will help us to track progress!

thank you!

Collaborators and RAs:

Najoung Kim

Tal Linzen

Alicia Chatten

CCC
Computing Community Consortium
Catalyst

NSF