Foundations: How to define and measure “understanding”

Sebastian Schuster

Seminar “What do language models really understand”?
April 27, 2023

Plan for today

- The octopus thought experiment and referentialism
- Responses:
  - Opportunities for grounding
  - Alternative views of what it means to understand
- Methods for evaluating understanding abilities
  - Benchmarks
  - Behavioral experiments
  - Probing
- Guidelines for presentations, reading papers, and commentaries

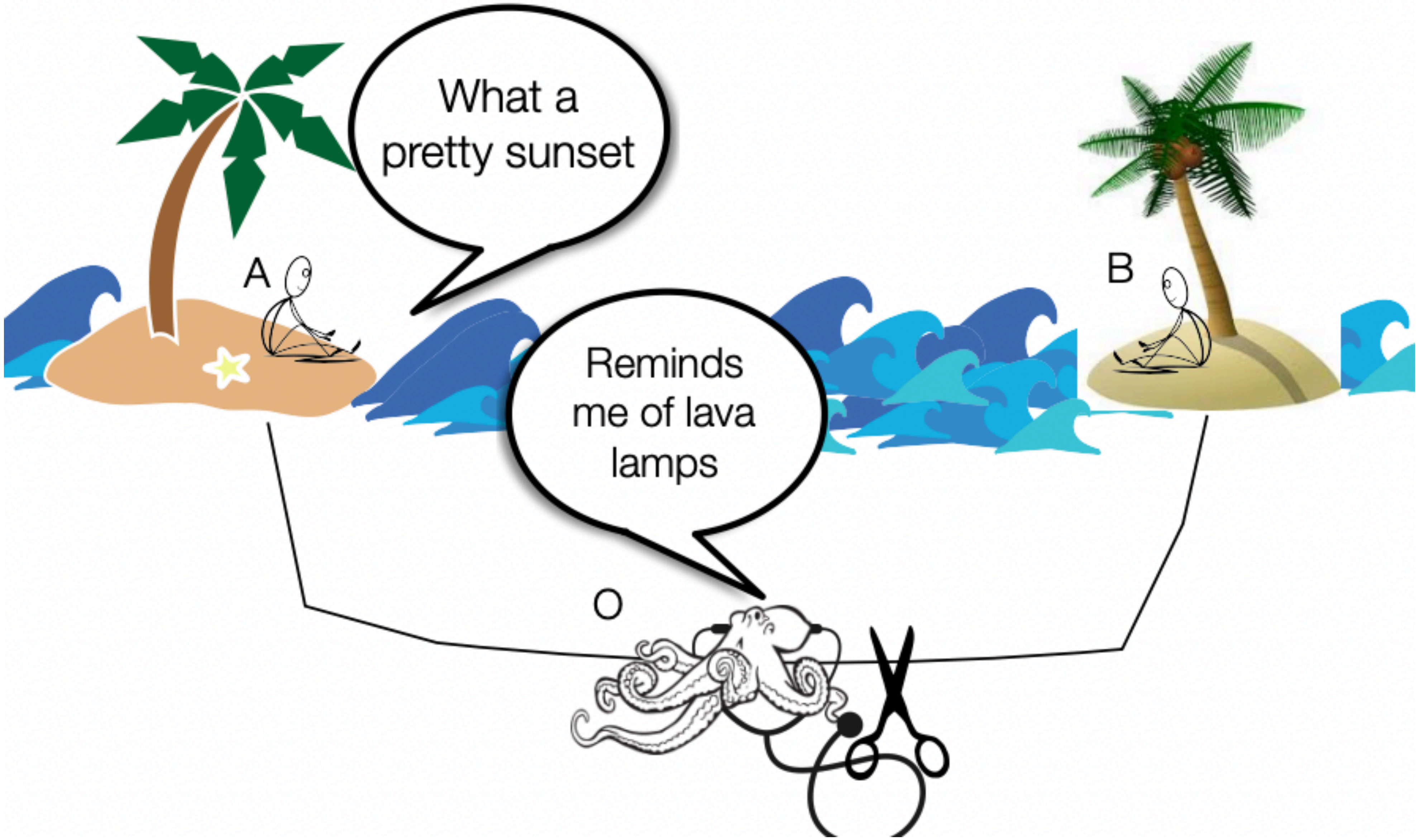
What is understanding?

- One view:
  - We don’t just use language for fun — we use language to achieve **communicative intents**
  - Formalization:

    Meaning: $M \subseteq E \times I$ (relation between natural language expressions *e* and communicative intents *i*)
  - Communicative intents are **something outside of language and grounded in the real world**
  - **Understanding:** given an expression *e*, in a context, recover the communicative intent *i*

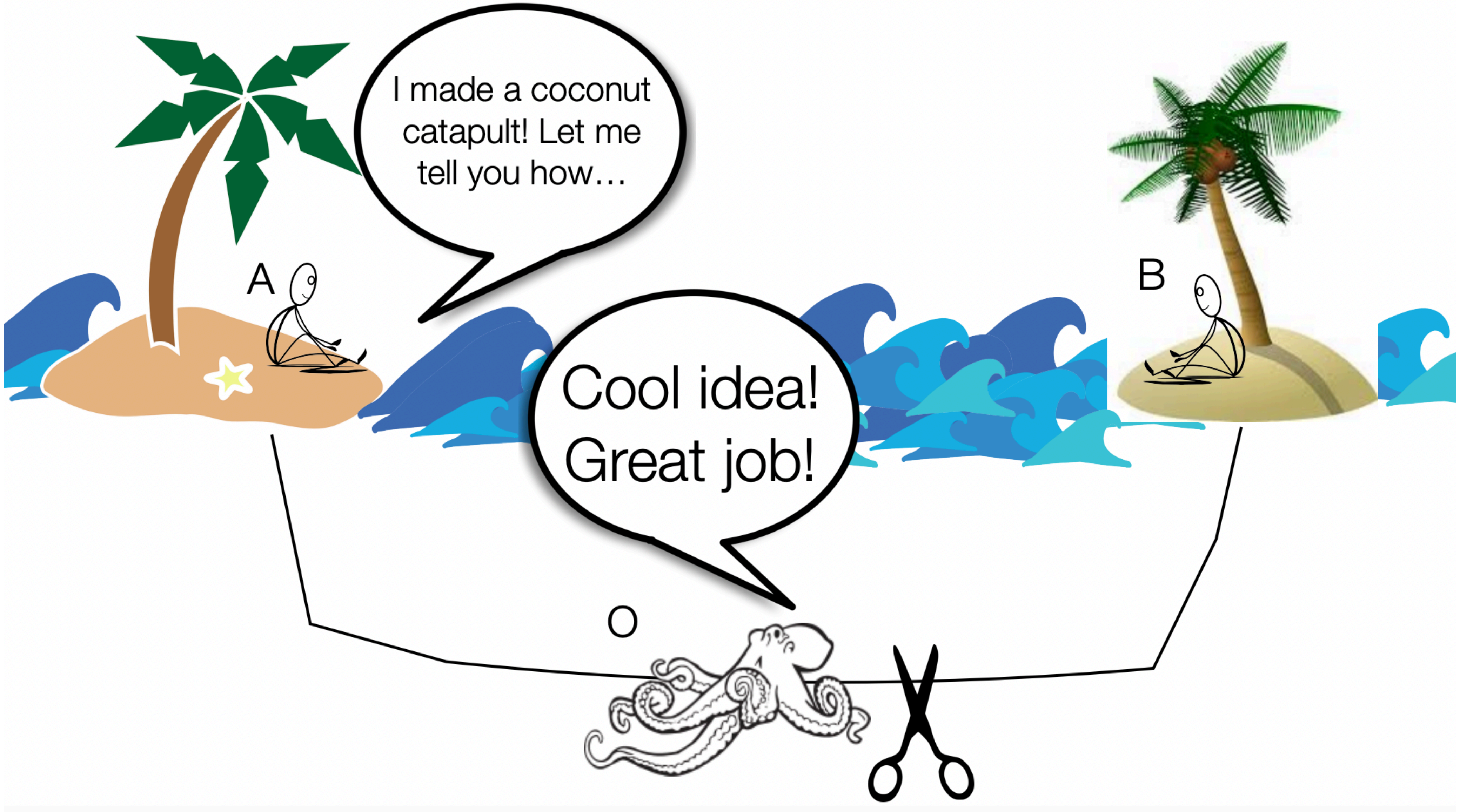
Bender and Koller (2020)

The octopus test



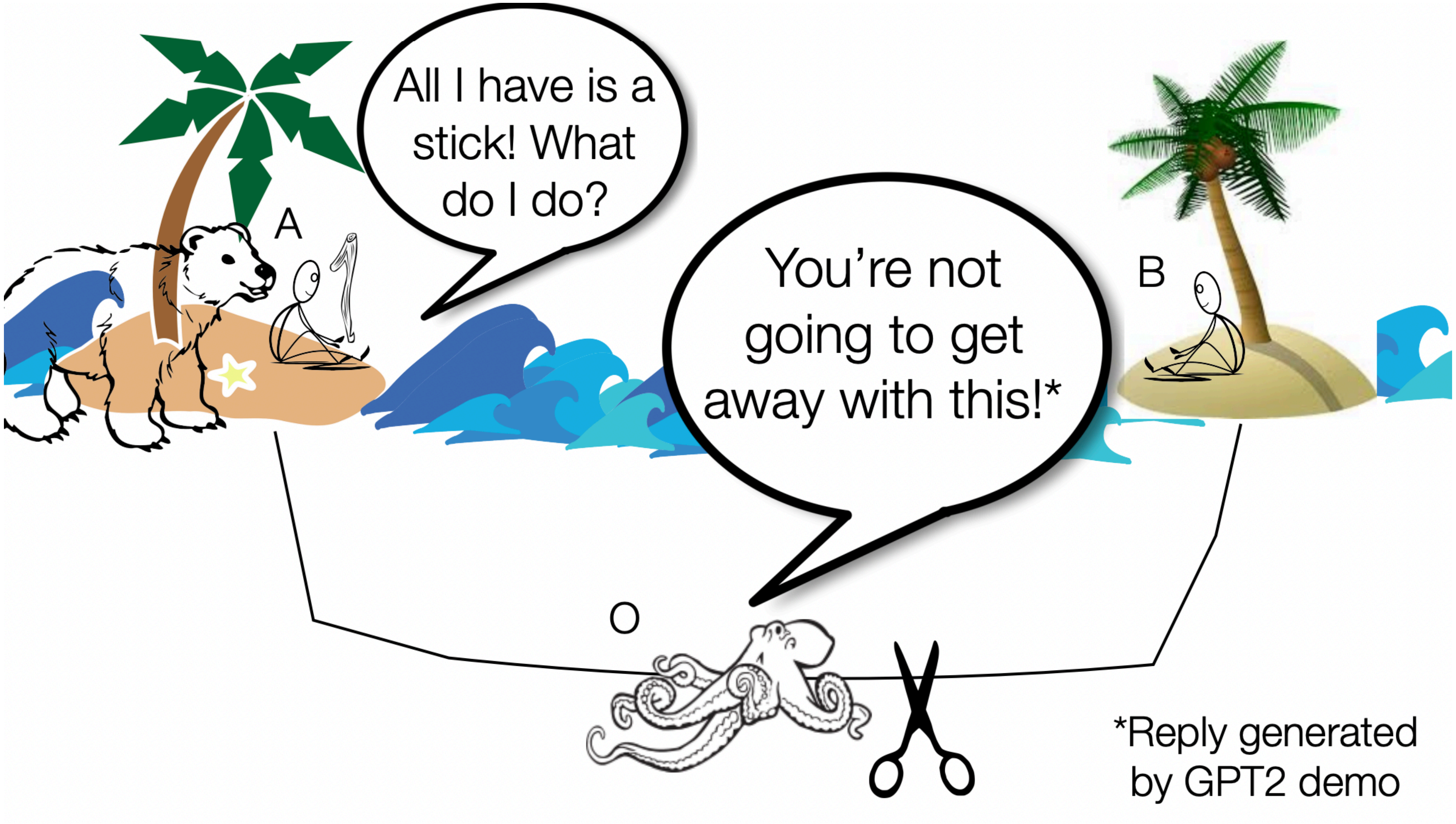
adapted from https://faculty.washington.edu/ebender/papers/Bender-Koller-2020-slides.pdf

The octopus test



adapted from https://faculty.washington.edu/ebender/papers/Bender-Koller-2020-slides.pdf

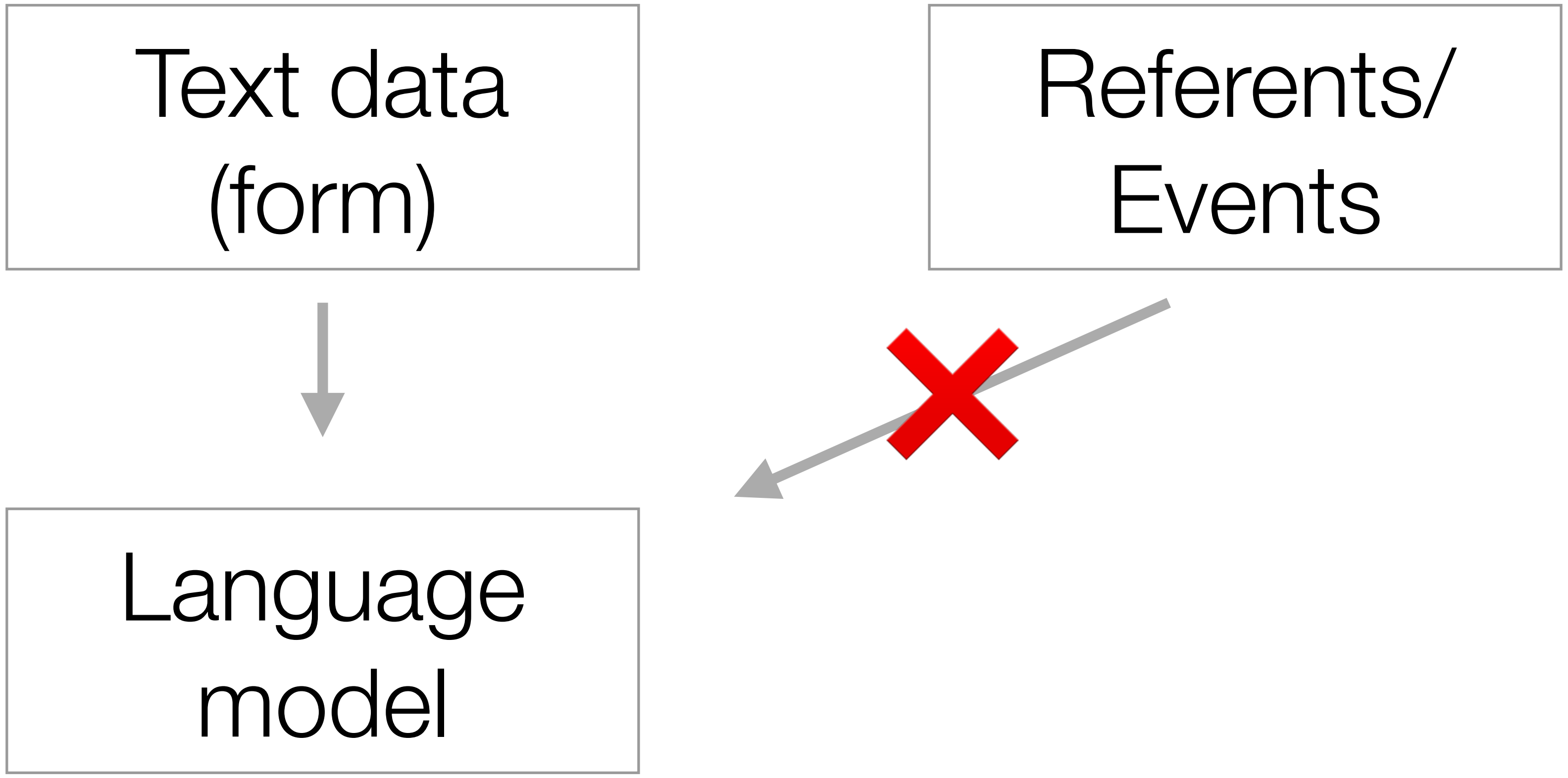
The octopus test


All I have is a stick! What do I do?
A
You're not going to get away with this!*
B
O

*Reply generated by GPT2 demo

adapted from https://faculty.washington.edu/ebender/papers/Bender-Koller-2020-slides.pdf

Can language models understand?



Text data (form)

Referents/ Events

Language model

Bender & Koller, 2020; Bender et al. 2021

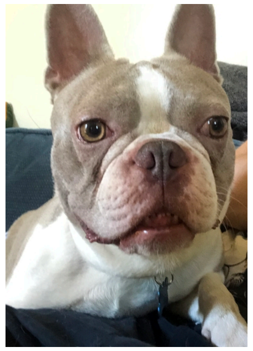
Responses

- Some grounding may happen when training only on form
    - e.g., unit tests in code
- Still relevant? Best LLMs are grounded in several ways (how?)
- Under specific assumptions about language use, pure LMs can learn whether one statement entails another statement (Merrill et al., 2022)
- There are alternative views of "understanding" than the one expressed by Bender & Koller.

Bender and Koller’s view: Referentialism

- **Referentialism**:
  - Words and phrases **map to entities and events** in the real world
  - An agent understands language if it is able to do this **mapping** and to **evaluate whether statements are true in the world**

Archie ⟷



Archie is a dog ⟷ `True`

One alternative view: Pragmatism

- **Pragmatism**
  - What matters is that the agent be disposed **to use language in the right way**
  - This may include appropriate inference and reasoning patterns, appropriate conversational moves, etc.
  - **Being able to use language in the right way** constitutes understanding

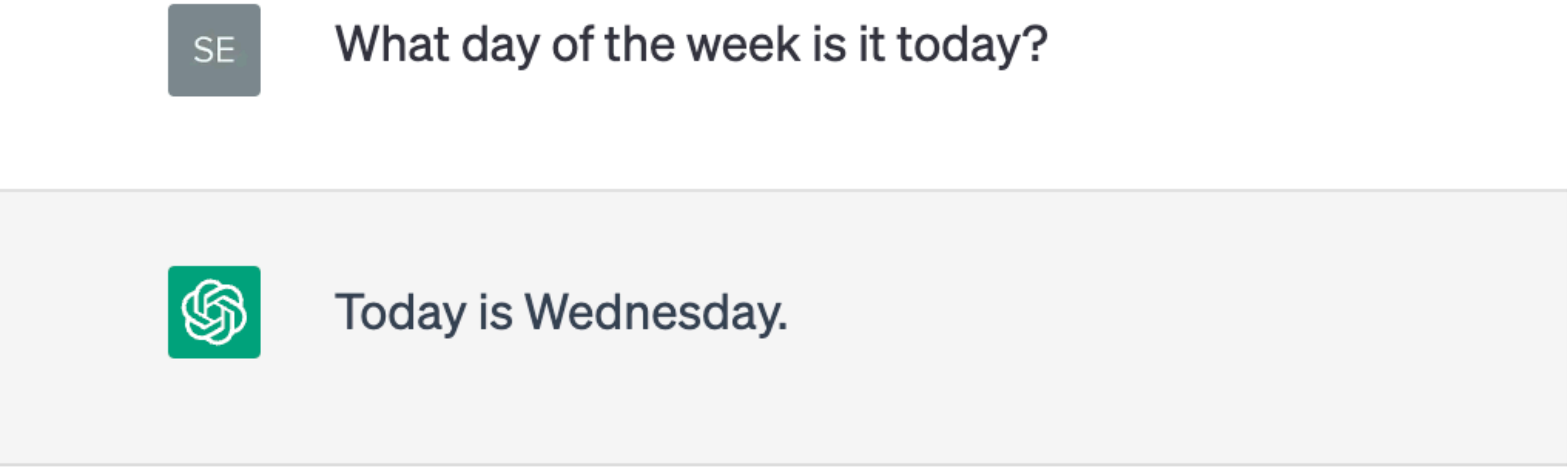
Archie is a dog → Archie is a mammal

Agent1: What day of the week is it today?

Agent2: It is Thursday.

Can language models understand?

- Under a pragmatist view: Maybe?



What day of the week is it today?

Today is Wednesday.

Methods for assessing understanding abilities

Methods for assessing understanding abilities

- Task benchmarks (e.g., Natural Language Inference benchmarks)

- Behavioral experiments (aka “Targeted evaluations”)

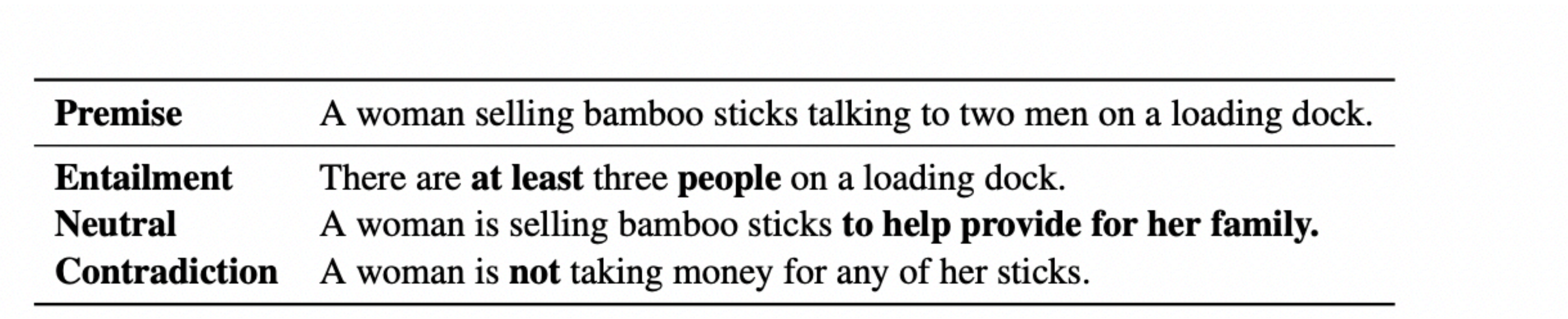
- Probing

Benchmarks

- A classic benchmark:
  - Crowdsourced examples
  - Randomly split into training/development/test examples
  - Model is trained on training split and evaluated on test split resulting in an overall accuracy score

Issues with benchmarking

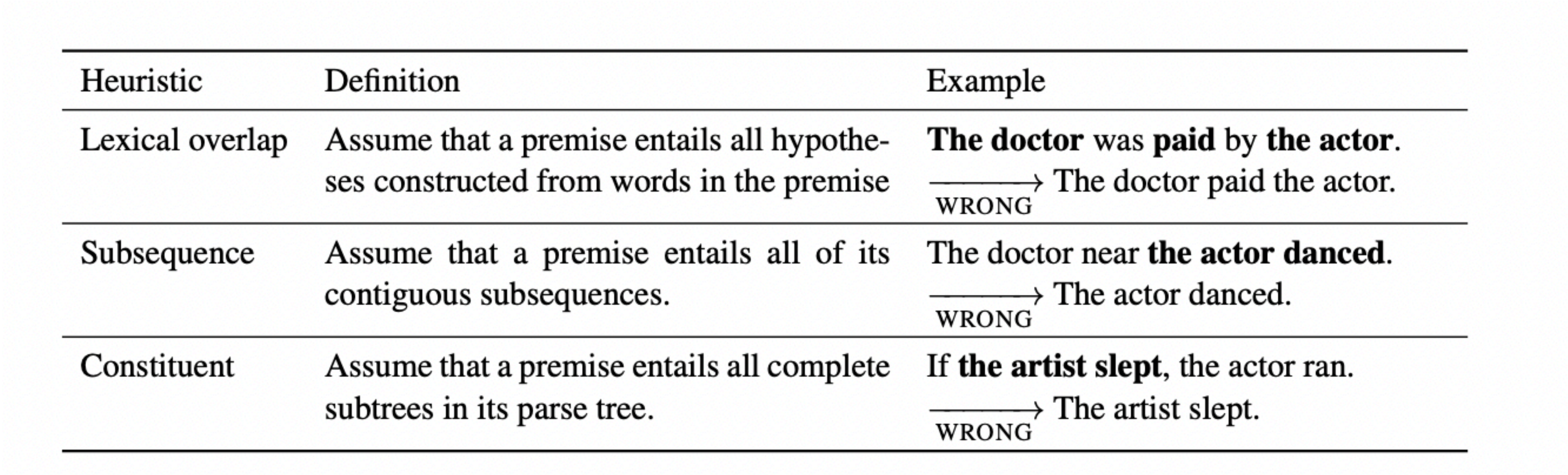
- Statistical artifacts in SNLI:



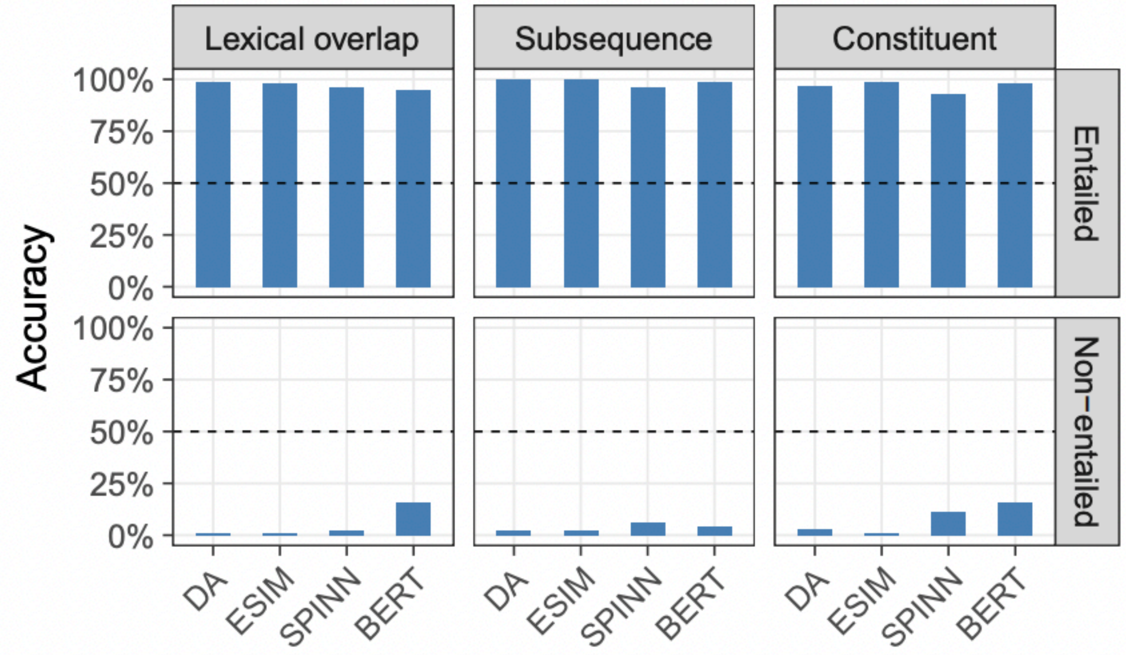
Premise	A woman selling bamboo sticks talking to two men on a loading dock.
Entailment	There are **at least** three **people** on a loading dock.
Neutral	A woman is selling bamboo sticks **to help provide for her family.**
Contradiction	A woman is **not** taking money for any of her sticks.

Gururangan (2018)

Issues with benchmarking



Heuristic	Definition	Example
Lexical overlap	Assume that a premise entails all hypotheses constructed from words in the premise	**The doctor** was **paid** by **the actor**. $\xrightarrow[\text{WRONG}]{}$ The doctor paid the actor.
Subsequence	Assume that a premise entails all of its contiguous subsequences.	The doctor near **the actor danced**. $\xrightarrow[\text{WRONG}]{}$ The actor danced.
Constituent	Assume that a premise entails all complete subtrees in its parse tree.	If **the artist slept**, the actor ran. $\xrightarrow[\text{WRONG}]{}$ The artist slept.



(b)

McCoy et al. (2019)

Issues with benchmarking

- Tasks are usually quite general
  - Question answering
  - Natural language inference
  - …
- **Difficult to identify systematic shortcomings**

Behavioral experiments / targeted evaluation suites

- Inspired by psycholinguistics experiments
- Small test sets that target a specific behavior, e.g., negation
- Models are usually not trained on similar examples
    - Evaluates out-of-distribution examples

e.g., Linzen (2020)

Example: Evaluating whether models learned dependencies necessary for reflexives

The bankers who the pilot embarrassed hurt ____

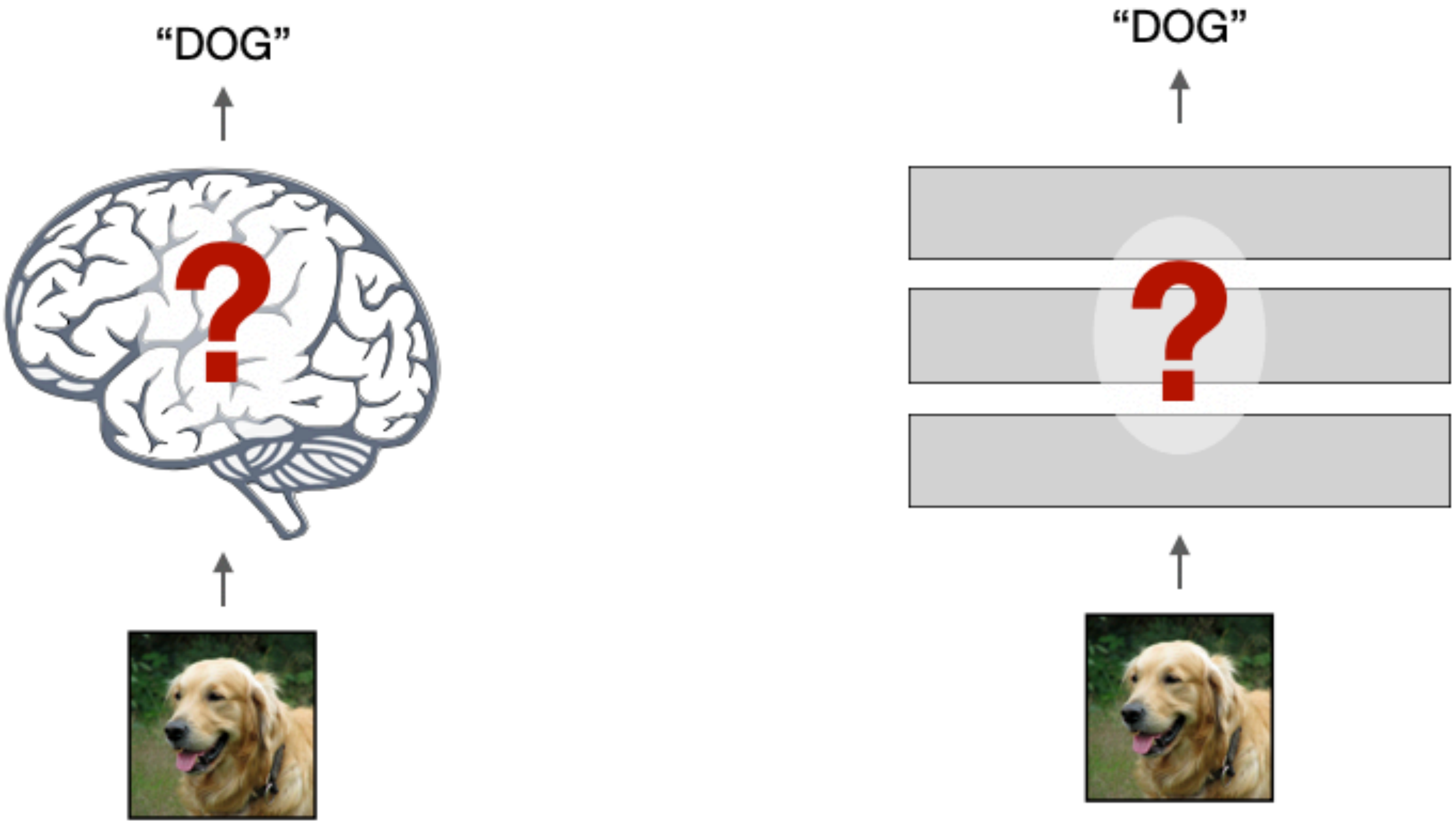
P(themselves | *Context*) > P (herself | *Context*)?

The bankers thought the pilot embarrassed ____

P(herself | *Context*) > P (themselves | *Context*)?

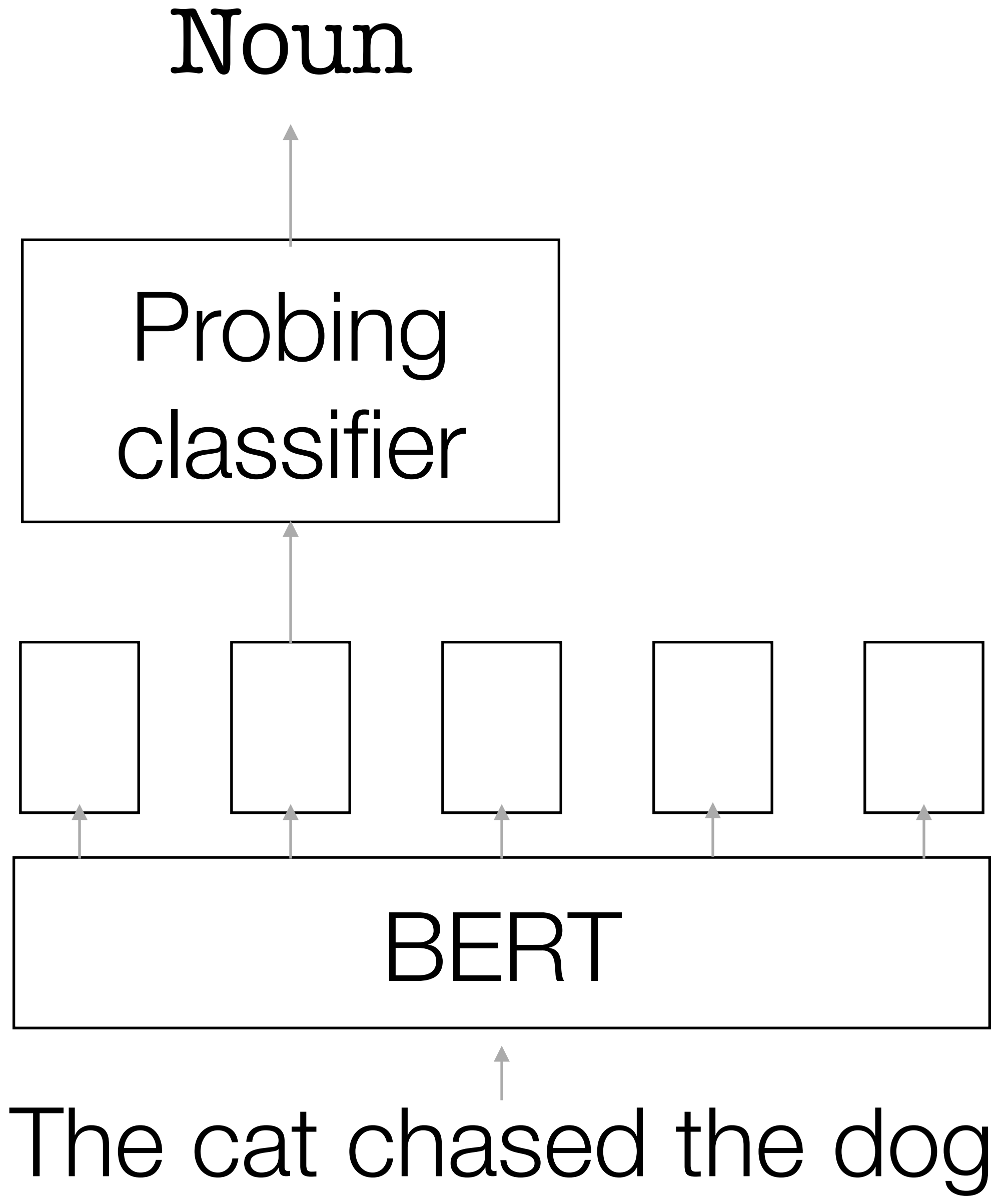
Hu et al. (2020)

Probing



Adapted from https://neuranna.mit.edu/news/probing-brains-vs-neural-networks

Example: determining whether representations encode something about part-of-speech tags



Takeaways

- Two prominent views of what it means to understand:
  - Referentialism and Pragmatism
- Methods for evaluating abilities of language models
  - Benchmarks
  - Targeted evaluations
  - Probing

Guidelines for readings and presentations

Questions to keep in mind while doing the readings

- What are the properties of the model(s) being used?
    - Autoregressive model? Masked LM? Something else?
    - How was the model trained? Additional training objectives on top of LM task?
- How was the understanding ability evaluated? Did the evaluation task potentially provide additional supervision?
- What kind of data was being used? Naturalistic? Hand-crafted? Can we rule out statistical artifacts in the data? Could the model have learned shortcuts?
- Do the authors talk about “understanding”? If so, what kind of definition of “understanding” do they (seem to) assume?
- Does the paper report results from models of different size? Does size seem to matter for the evaluated ability?

Guidelines for presentations

- Length: 15-25min + 15-25min of discussion
- Contents:
  - Summary of the **main questions, methods and results**
  - (optional) Background on model and data
  - Discussion of the strengths and weaknesses of the paper
- Slides and handouts are optional

Guidelines for weekly comments

- Superficial questions/comments:
    - Questions/comments that could have been written by reading just the abstract or a paragraph of the paper
    - “I didn’t understand X…” (fine to mention that as well but not as the only comment!)
- Examples of insightful questions/comments:
    - Connect multiple points made in the paper
    - Relate findings of a paper to other papers we’ve read
    - Relate to some of the higher-level questions we are asking in this course