

Large Language Models and the Lawyer's Search for Meaning

Emily M. Bender
University of Washington

Critical Legal AI Literacies
Yale Law School
September 9, 2025

New York lawyers sanctioned for using fake ChatGPT cases in legal brief

By Sara Merken

June 26, 2023 1:28 AM PDT · Updated June 26, 2023

LAW

Michael Cohen says he unwittingly sent AI-generated fake legal cases to his attorney

DECEMBER 30, 2023 · 8:20 AM ET

Judge weighs sanctioning lawyers over 'fictitious' case citations

By Sara Merken

July 25, 2024 2:33 PM PDT · Updated July 25, 2024

Civil Litigation

Aug. 20, 2025

In Onlyfans case, Hagens Berman filed AI-generated briefs with fake case citations

Partner Robert Carey blamed the errors on an outside attorney dealing with a family crisis who used AI tools without proper disclosure or review.

For more:

<https://www.damiencharlotin.com/hallucinations/>

ChatGPT wrote California
Khanna's new

Judges Given the OK to Use ChatGPT in Legal Rulings

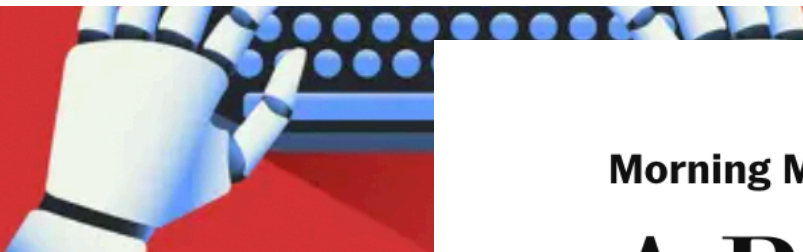
The UK now permits judges to use the “jolly useful” AI chatbot in court.

By Thomas Germain

Published December 12, 2023 | Comments (0) | [Share](#)

Judges and Masters Have ChatGPT Fever

JUAN DAVID GUTIÉRREZ / APR 19, 2023



Lawmakers are using ChatGPT to draft legislation
Chronicle photo illustration from Getty Images

By **Shira Stein**, Washington Correspondent
July 20, 2023

Morning Mix

A Brazilian city passed a law about water meters. ChatGPT wrote it.

Councilman Ramiro Rosário, who sponsored the bill, revealed only after it had passed that it was written by AI

December 4, 2023

More than **1 year ago**

Outline

- Brief overview & history of language models
- Form vs. meaning: Why language models don't “understand”
- Automation and legal practice
- The “god trick” in law and “AI”

What's a language model?

- Better term: “corpus model” (Veres 2022)
- Given a collection of text (corpus) representing a language, how likely is a given string to appear?
- Earliest were n-gram language models (Shannon 1948)
 - Unigram: relative frequency of single words
 - Bigram: relative frequency of words given one previous word
 - Trigram: relative frequency of words given two previous words

What are language models good for?

- Ranking spelling correction candidates
- Ranking acoustic model outputs in automatic transcription
- Ranking translation model outputs in machine translation
- Simplified text entry (T9)



What's a neural language model?

- So-called “neural nets” are not artificial brains/minds
- Collections of “perceptrons”: Mathematical model based on a simplified version of 1940s understanding of neurons

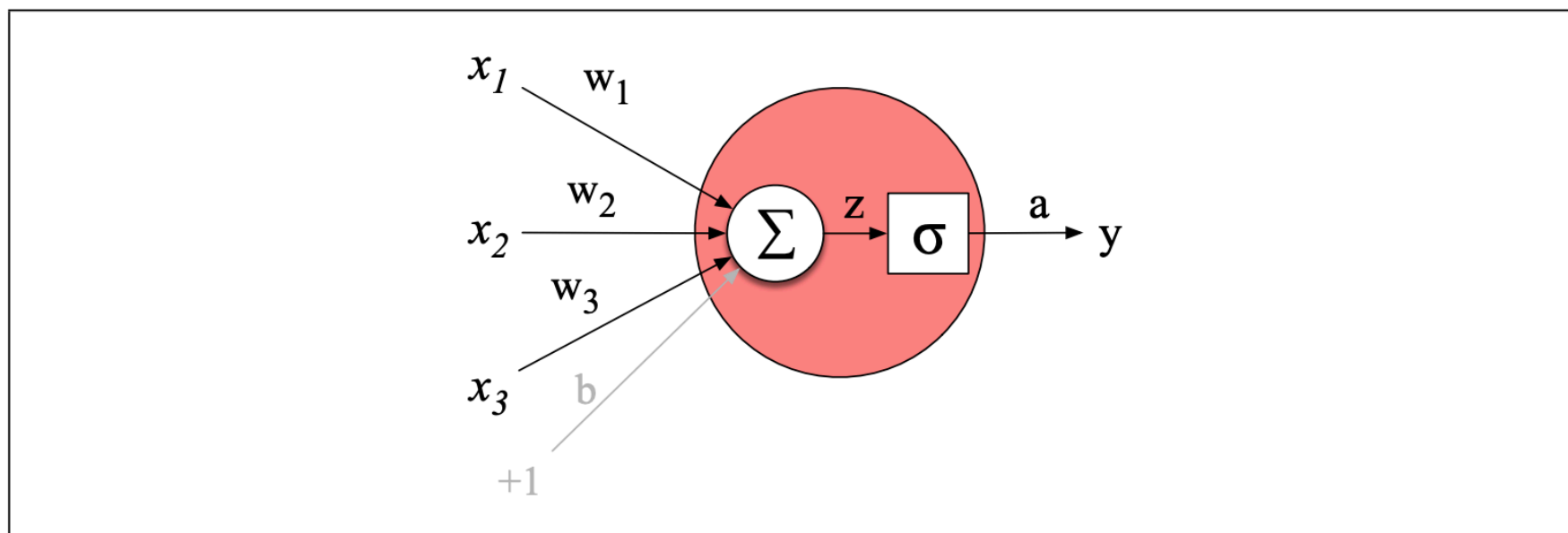


Figure 7.2 A neural unit, taking 3 inputs x_1 , x_2 , and x_3 (and a bias b that we represent as a weight for an input clamped at $+1$) and producing an output y . We include some convenient intermediate variables: the output of the summation, z , and the output of the sigmoid, a . In this case the output of the unit y is the same as a , but in deeper networks we'll reserve y to mean the final output of the entire network, leaving a as the activation of an individual node.

(Jurafsky & Martin 2023, Ch 7)

What's a neural language model?

- “Neural net” whose input is a sequence of words and output is a probability distribution over the vocabulary — how likely is each word to come next?
- Trained with “back propagation”: compare actual next word to predictions and, when different, adjust weights throughout the network (slightly) (Bengio et al 2003)
- Represent words as “embeddings” (dense vectors reflecting word co-occurrence) rather than character strings, for better generalization across words (Mikolov et al 2013)
- Performance improvement through architecture innovations like Long Short-Term Memory (Hochreiter and Schmidhuber, 1997) and Transformer (Vaswani et al 2017) models and training paradigms (BERT; Devlin et al 2017)

What are neural language models good for?

- Much smoother automatic transcription and machine translation output
- Query expansion in search
- Grammar checker
- Autocorrect
- Word “embeddings” => dramatic improvements to almost every kind of language technology

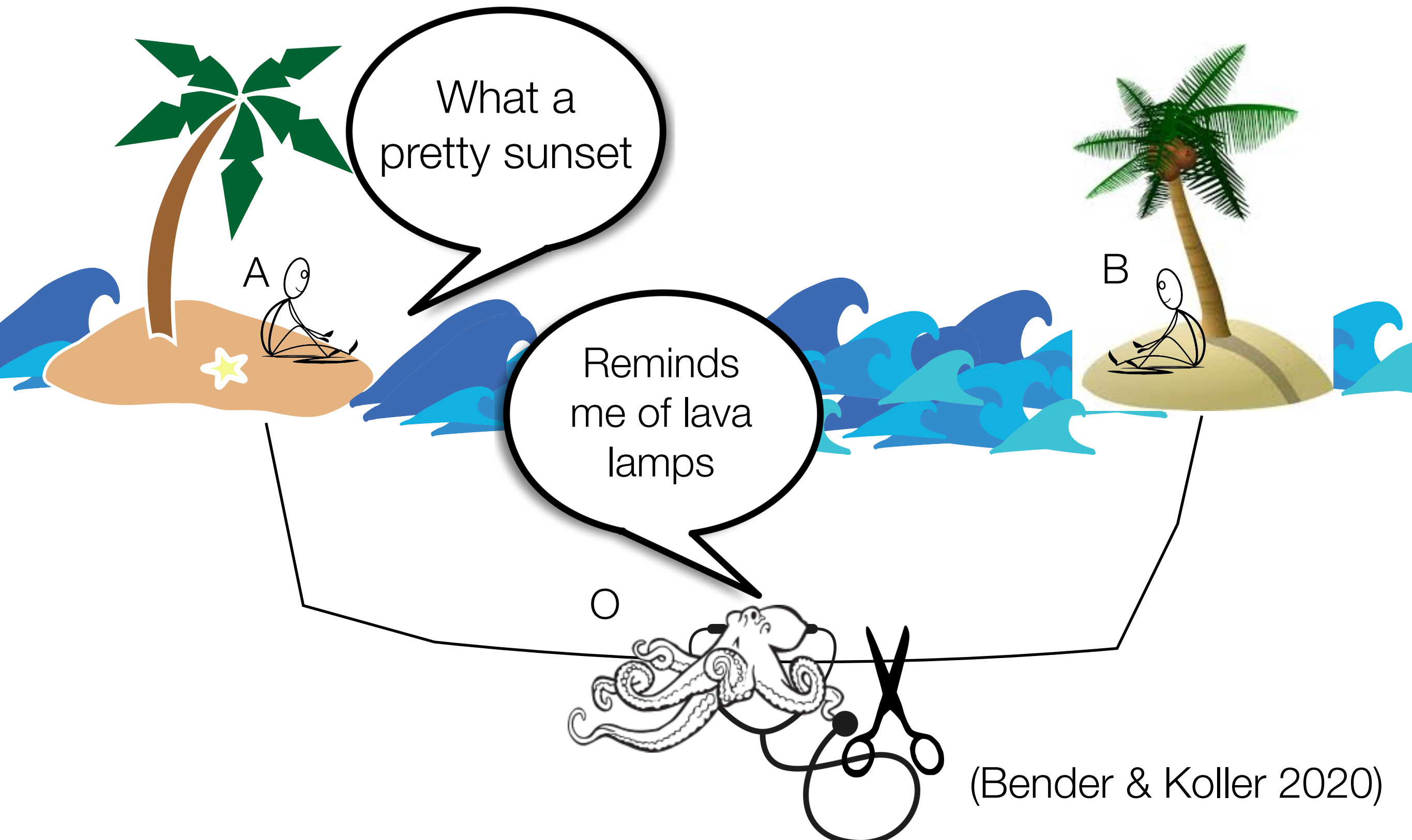
What is “generative AI”?

- Turning systems meant for classification/ranking inside-out
- Instead of “Which string is more plausible?” we get “What word comes next?”
- Cover term for other kinds of synthetic media machines (audio, image, video) as well
- Not “AI”, and definitely not “AGI”

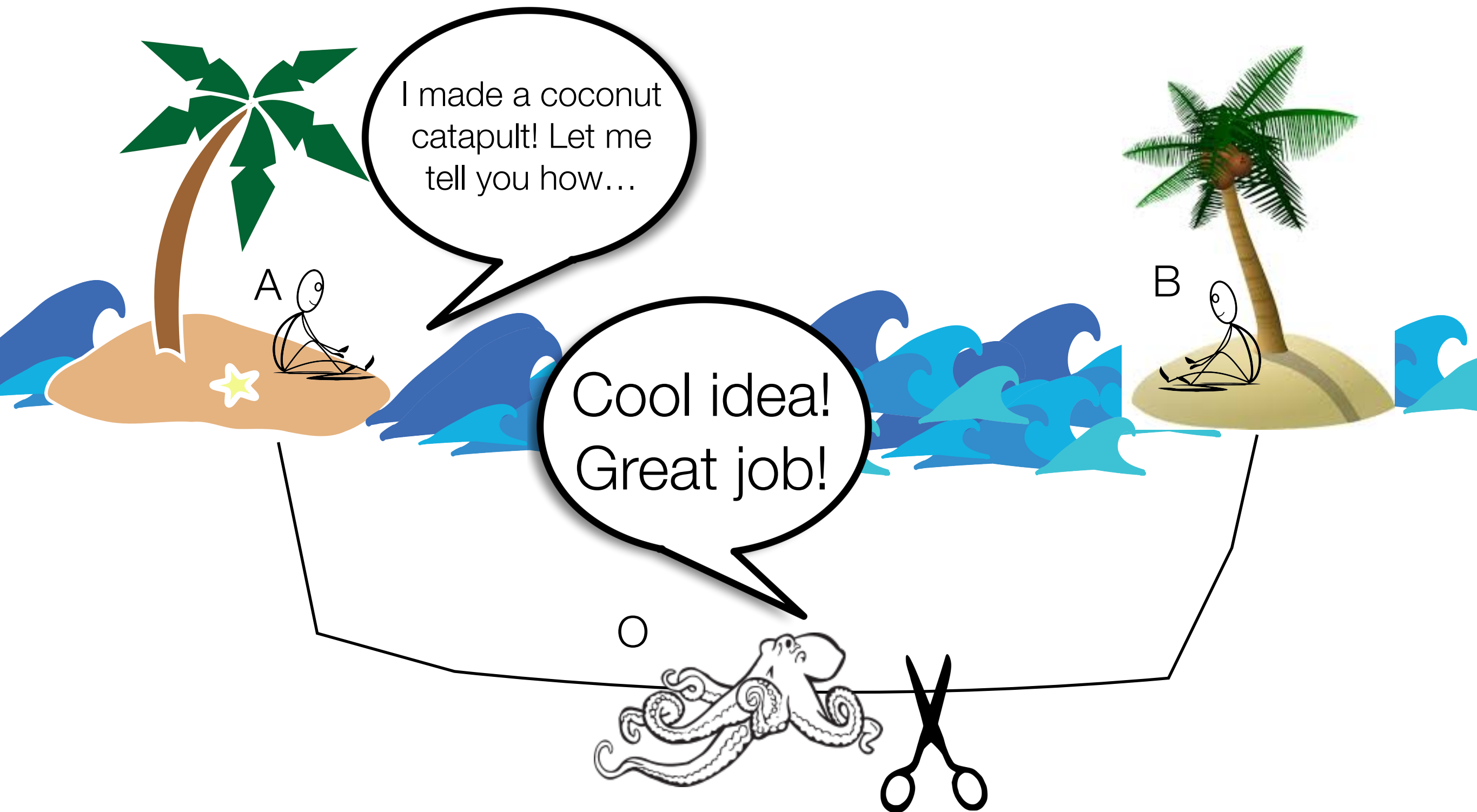
Chatbots/large language models are a parlor trick

- A language model is a model of the distribution of word spellings in some collection of text
- Today's large language models are very effective tools for mimicking how people use language: what words we say in what order
- They provide the illusion of reasoning/"intelligence" because of how we understand language: by imagining a mind behind the text

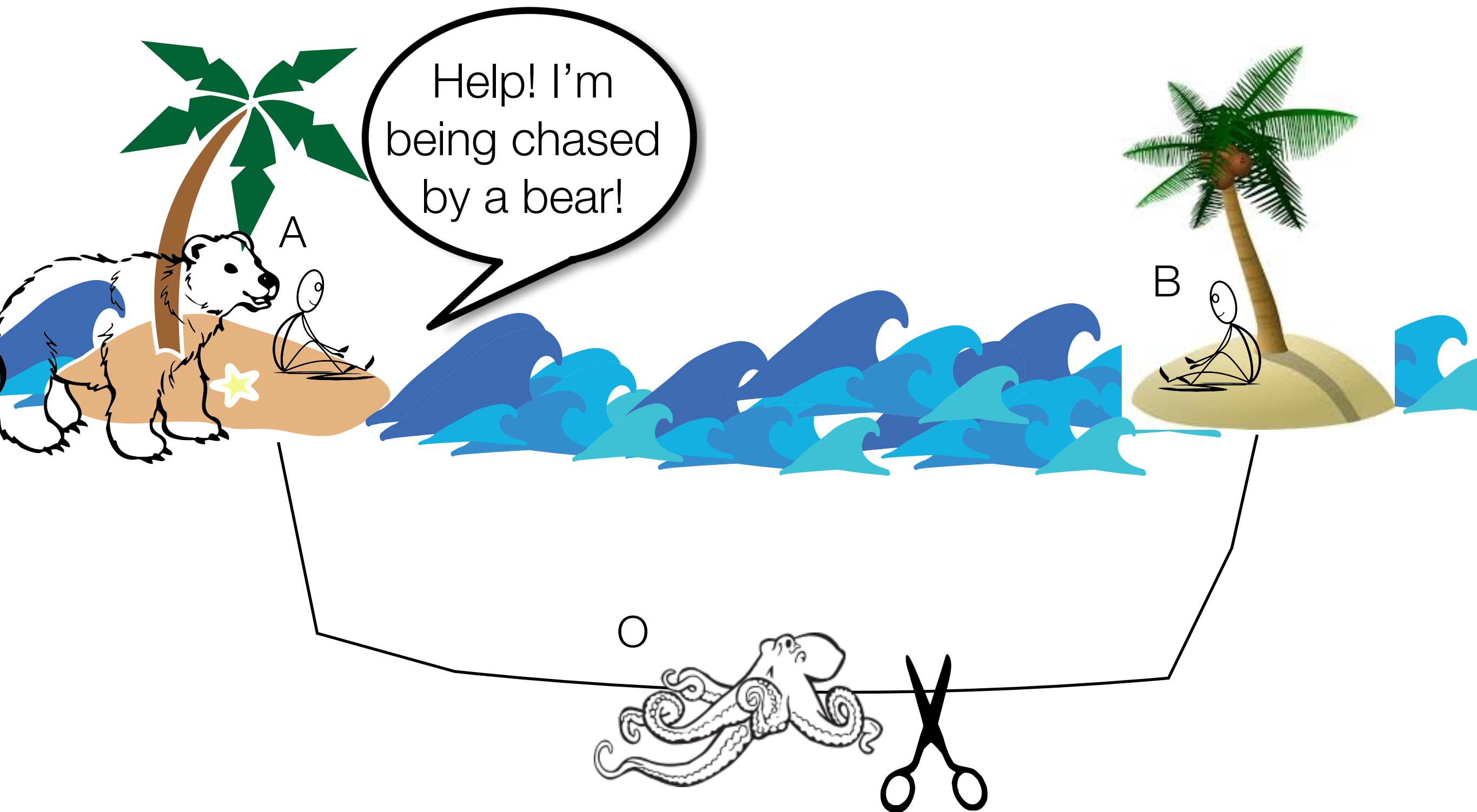
Thought experiment: Meaning from form alone



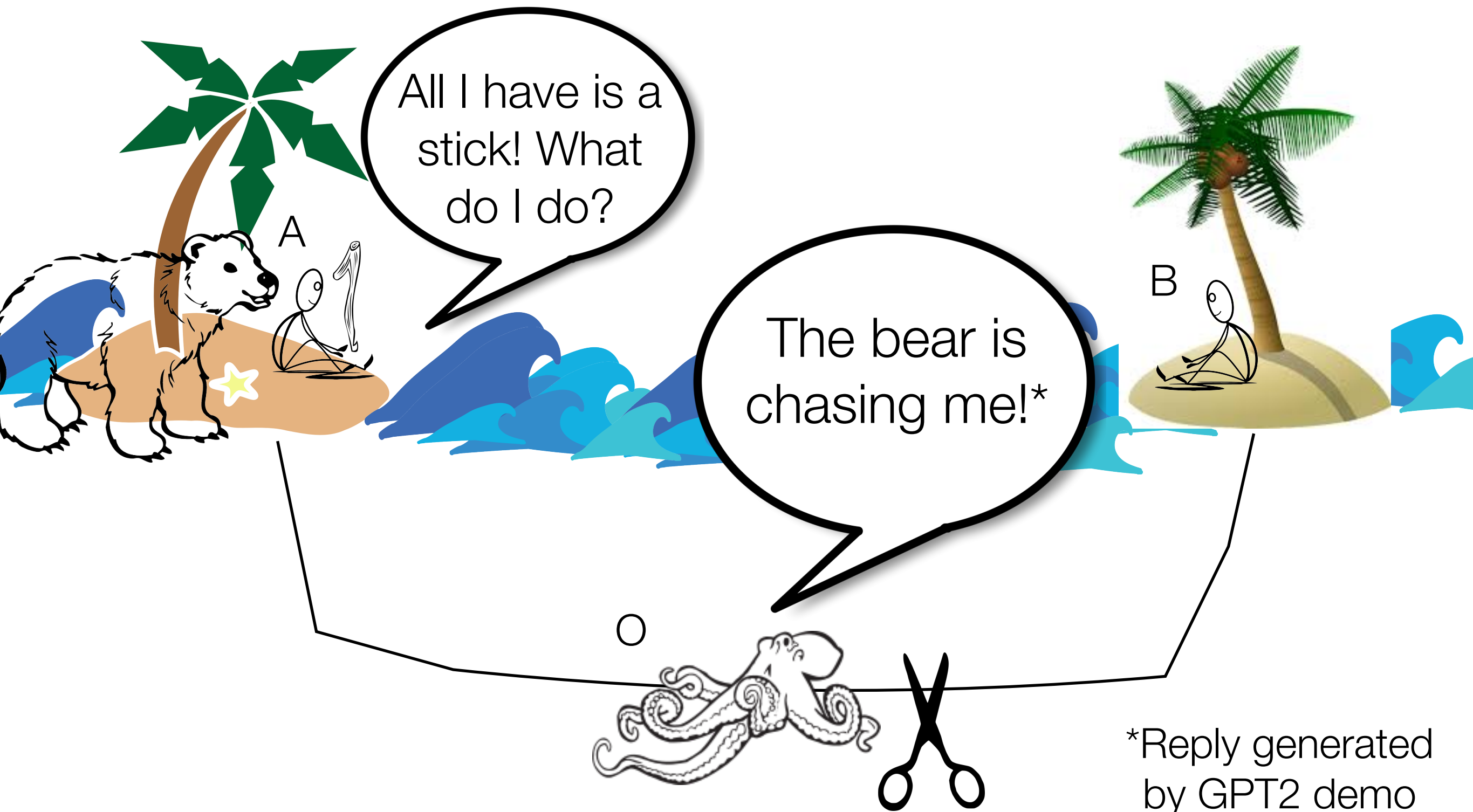
Thought experiment: Meaning from form alone



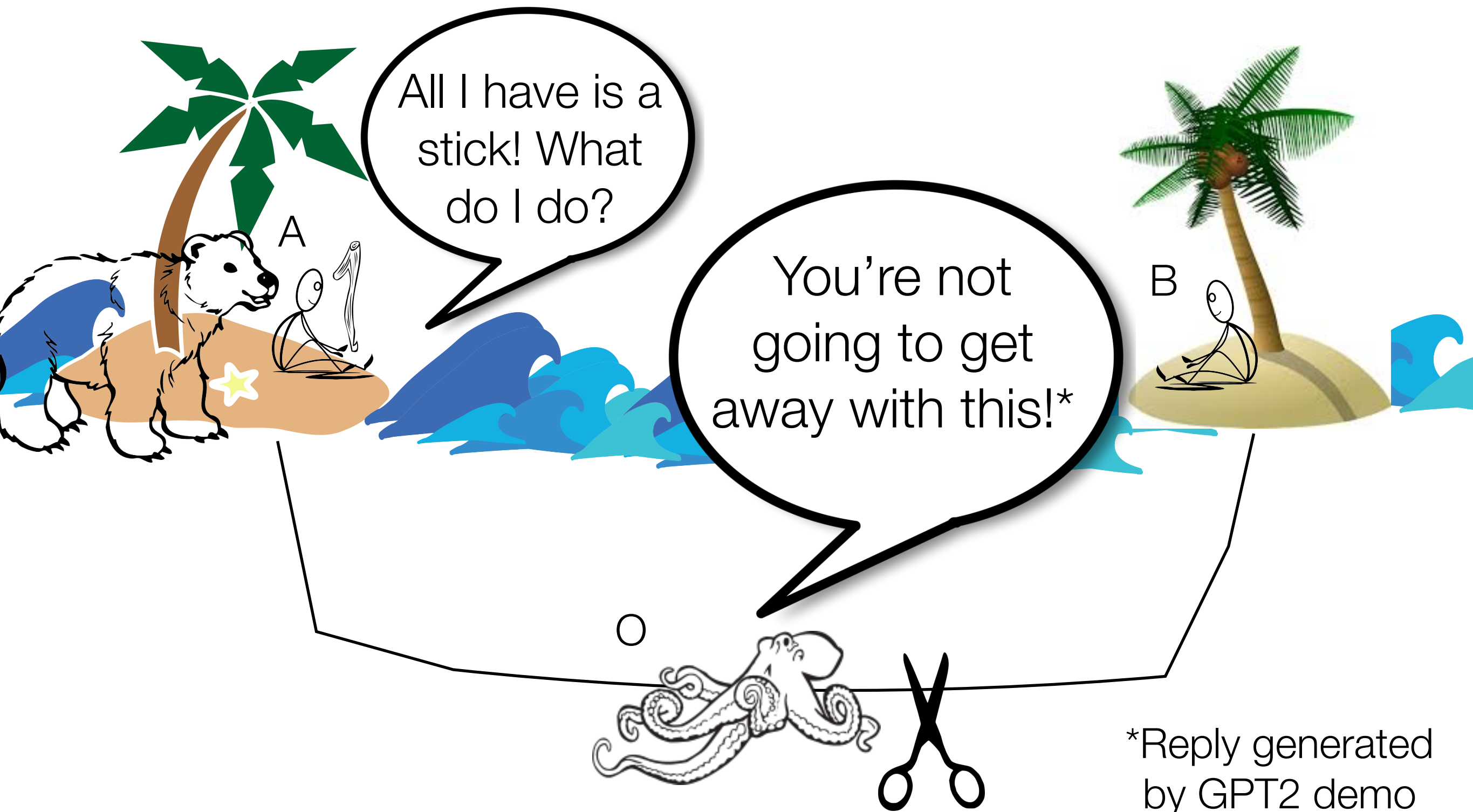
Thought experiment: Meaning from form alone



Thought experiment: Meaning from form alone



Thought experiment: Meaning from form alone



2023 update: National Library of Thailand

bit.ly/Bender-NLT

- You're in the National Library of Thailand
- Unlimited time, unlimited delicious Thai food, no people to interact with
- All documents with images or non-Thai text removed
- Can you learn Thai?
- How?

(Photo credit:
Pat Roengpitya)



But wouldn't it be so wonderfully useful if?

- ChatGPT isn't a general-purpose "answer machine"
- But what if one could be built?
- I would argue that still would have no place in the practice of law
- What are you doing when you write a legal brief/do an analysis/search for precedents?
- What would it mean to off-load that work to an automated process?

Conditions for appropriate automation

- Specificity of what is being automated — tools are tools *for* something
- Evaluation in context of the automated system
- Transparency of the affordances of the system
- Accountability for system functionality
- Accountability for appropriate system use

The “god trick” (Haraway 1988) in the law and “AI”

- “seeing everything from nowhere” (p.581)
- In “AI”: The fantasy of all-knowing “answer machines”, or the idea that any question articulated has a definitive answer, that can be provided due simply to the processing of enough text
- Possible analog in law: The idea that legal systems are functions, with inputs and outputs that can be applied most ideally when applied mechanically, rather than structures we build to better treat each other fairly

Weizenbaum 1976

Computers can make judicial decisions, computers can make psychiatric judgments. They can flip coins in much more sophisticated ways than can the most patient human being. The point is that **they ought not be given such tasks.** [...] What emerges as the most elementary insight is that, since we do not now have any ways of making computers wise, **we ought not now to give computers tasks that demand wisdom.**

Outline

- Brief overview & history of language models
- Form vs. meaning: Why language models don't "understand"
- Automation and legal practice
- The "god trick" in law and "AI"



References

- Bender, E. M. and Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.
- Bengio, Y., Ducharme, R., Vincent, P., and Janvin, C. (2003). A neural probabilistic language model. *J. Mach. Learn. Res.*, 3(null):11371155.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Haraway, D. (1988). Situated knowledges: The science question in feminism and the privilege of partial perspective. *Feminist Studies*, 14(3):575–599.
- Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9:1735–1780.
- Jurafsky, D. and Martin, J. H. (2023). *Speech and Language Processing (3rd ed. draft)*. Available from <https://web.stanford.edu/~jurafsky/slp3/>.
- Mikolov, T., Chen, K., Corrado, G. S., and Dean, J. (2013). Efficient estimation of word representations in vector space. In *International Conference on Learning Representations*.
- Shannon, C. E. (1948). A mathematical theory of information. *Bell System Technical Journal*, 27:379–423, 623–656.
- Vaswani, A., Shazeer, N. M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In *NeurIPS*.
- Veres, C. (2022). Large language models are not models of natural language: They are corpus models. *IEEE Access*, 10:61970–61979.
- Weizenbaum, J. (1976). *Computer Power and Human Reason: From Judgment to Calculation*. Freeman, San Francisco CA.