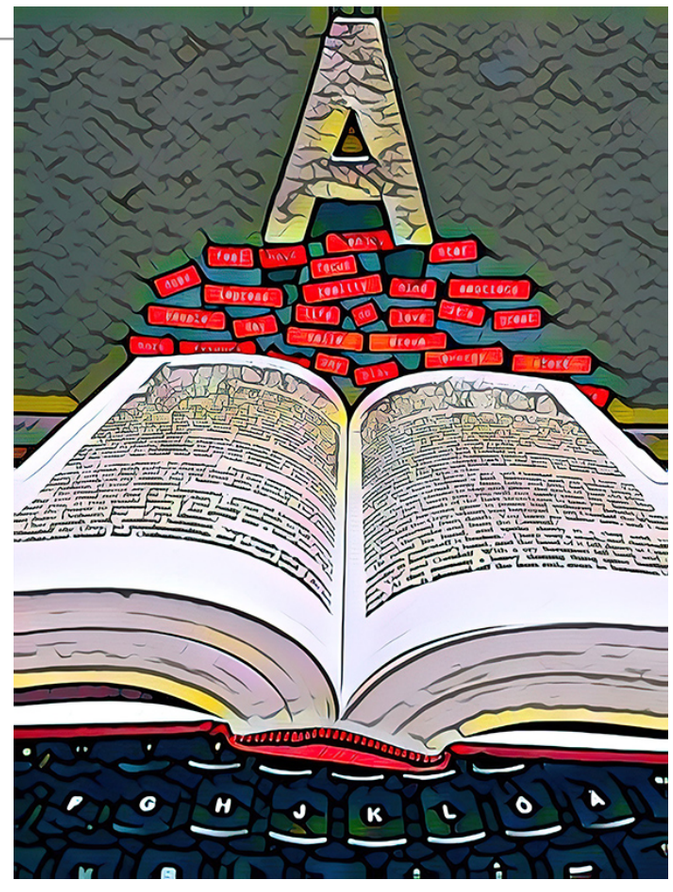
https://bit.ly/EMB-SM-23

ChatGPT in a Medical Setting: When, If Ever, Is Synthetic Text Safe, Appropriate and Desirable?

Emily M. Bender
University of Washington

Stanford Medicine Grand Rounds
Department of Medicine
Stanford University
15 November 2023



Teresa Berndtsson / Better Images of AI / Letter Word Text Taxonomy / CC-BY 4.0

Overview

- Large language models seem like nearly-there solutions to many problems, including in a medical context

- in fact, they only mimic language use, without understanding

- in addition, they absorb and amplify bias

- … while being misleadingly fluent.

- Despite strong sales pressure, there are almost no appropriate use cases for this technology

https://bit.ly/EMB-SM-23

Outline

- Brief overview & history of language models
- Form vs. meaning: Why language models don’t “understand”
- On the dangers of stochastic parrots 🦜
- Criteria for appropriate use cases in medicine
- Sample use cases held up to those criteria
- Take-aways

https://bit.ly/EMB-SM-23

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 give 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

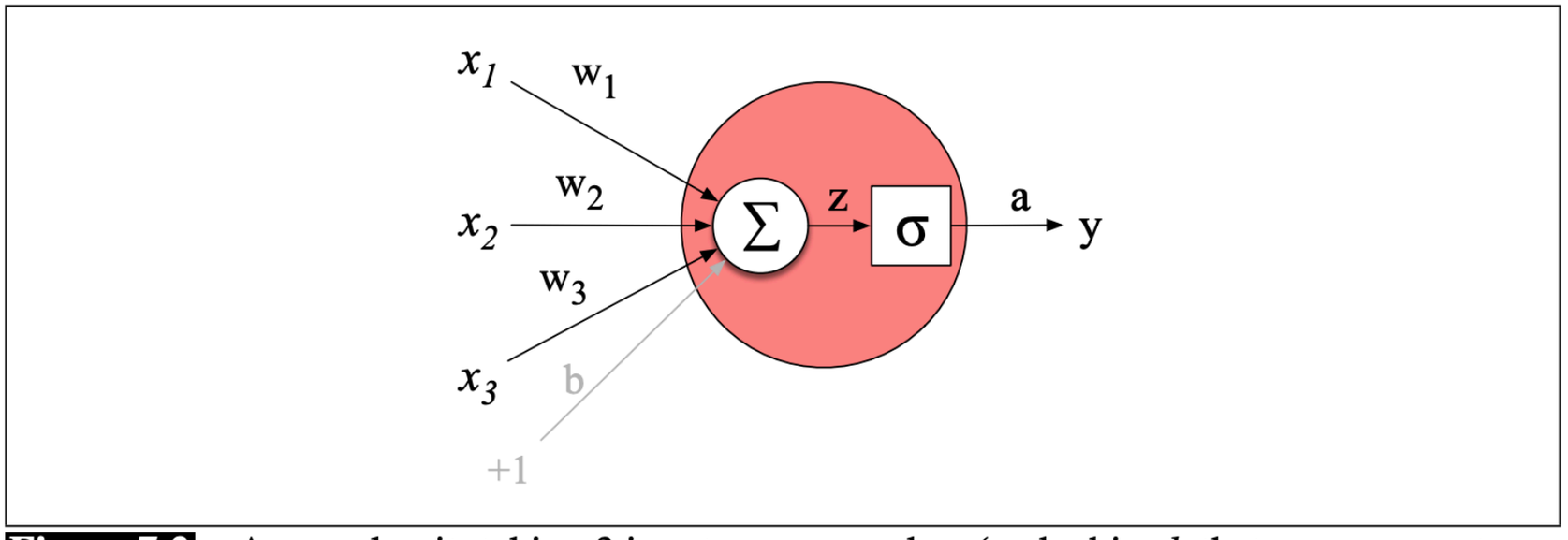


Figure 7.2 A neural unit, taking 3 inputs x_1, x_2, and x_3 (and a bias b that we represent as a weight for an input clamped at +1) and producing an output y. We include some convenient intermediate variables: the output of the summation, z, and the output of the sigmoid, a. In this case the output of 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**?

- **Represent words as "embeddings"** (dense vectors reflecting word co-occurrence) rather than character strings, for better generalization across words (Mikolov et al 2013)

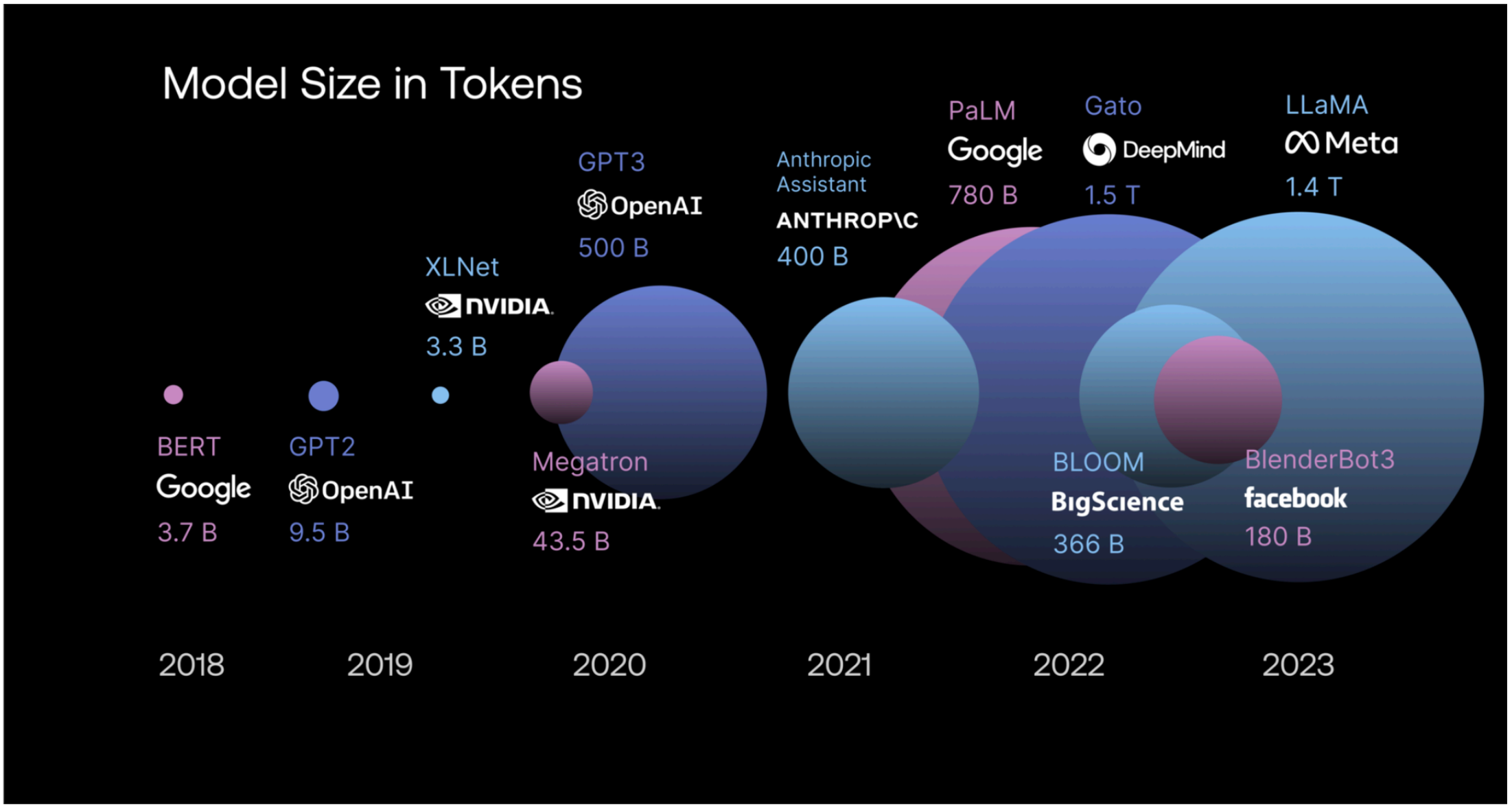
- Trained with "back propagation": compare actual next word to predictions and, when different, adjust weights throughout the network (slightly) (Bengio et al 2003)

- 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’s a large language model?



https://scale.com/guides/large-language-models

What are large language models good for?

- Automatic transcription, machine translation
- “End-to-end” approaches to many, many language technology tasks:
    - Summarization
    - Sentiment analysis
    - Taking multiple-choice tests
    - …

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”

What is “generative AI” good for?

When, if ever, is synthetic text safe, appropriate, and desirable?

Outline

- Brief overview & history of language models
- Form vs. meaning: Why language models don’t “understand”
- On the dangers of stochastic parrots 🦜
- Criteria for appropriate use cases in medicine
- Sample use cases held up to those criteria
- Take-aways

Climbing towards NLU:
On Meaning, Form, and Understanding in the Age of Data

Emily M. Bender, University of Washington
Alexander Koller, Saarland University

ACL 2020

So how do babies learn language?

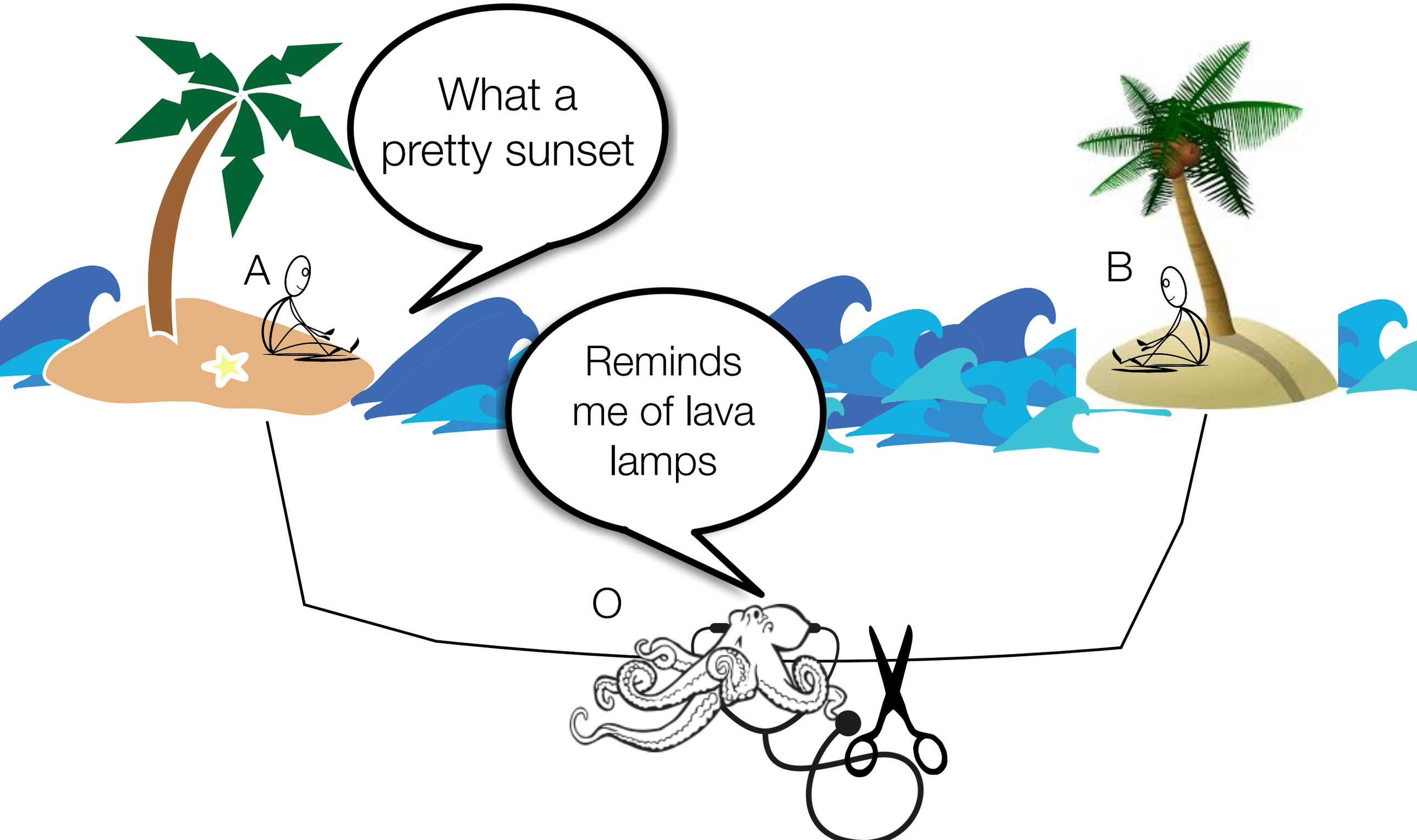
- Interaction is key: Exposure to a language via TV or radio alone is not sufficient (Snow et al 1976, Kuhl 2007)

- Interaction allows for joint attention: where child and caregiver are attending to the same thing and mutually aware of this fact (Baldwin 1995)

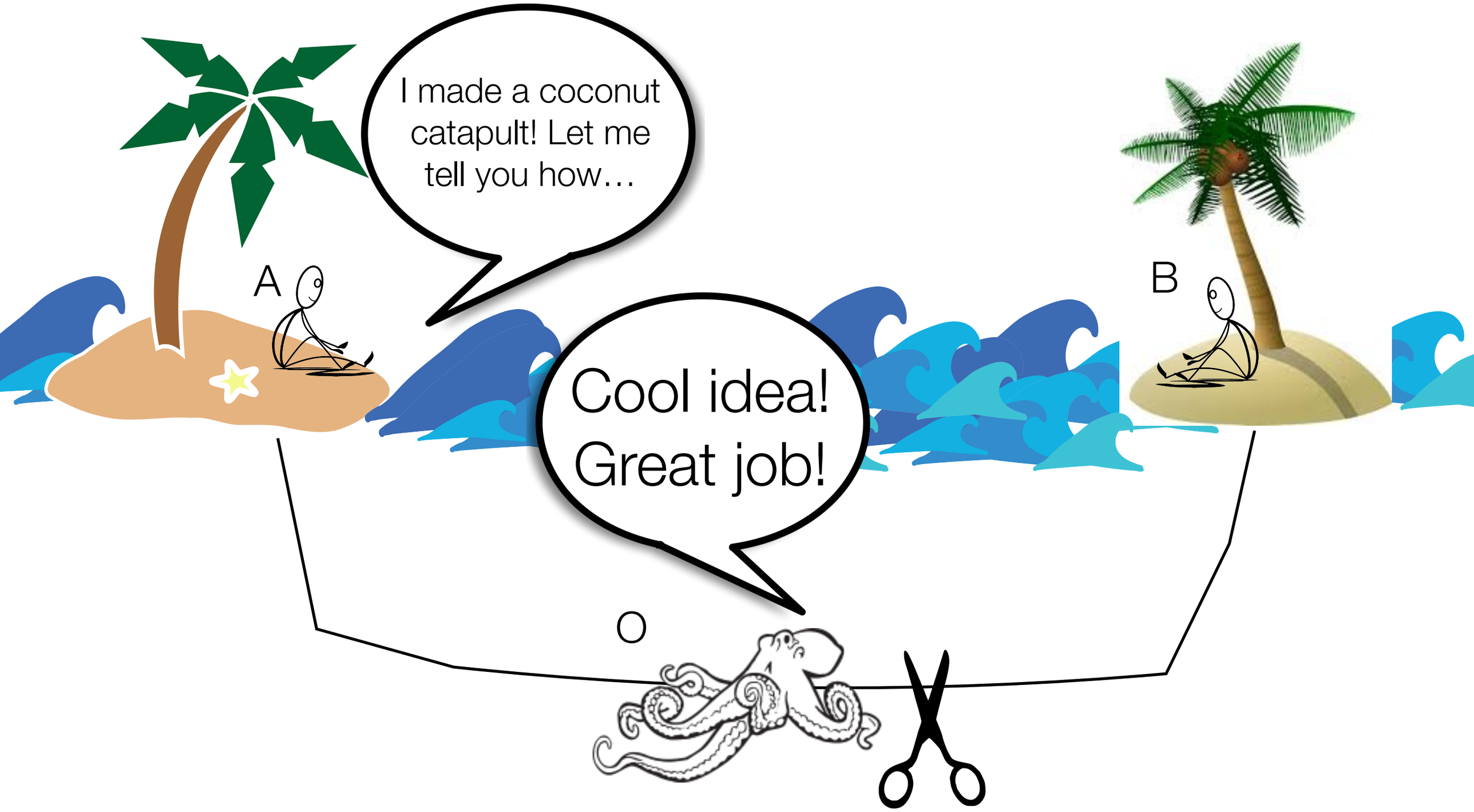
- Experimental evidence shows that more successful joint attention leads to faster vocabulary acquisition (Tomasello & Farrar 1986, Baldwin 1995, Brooks & Meltzoff 2005)

- Meaning isn’t in form; rather, languages are rich, dense ways of providing cues to communicative intent (Reddy 1979). Once we learn the systems, we can use them in the absence of co-situatedness.

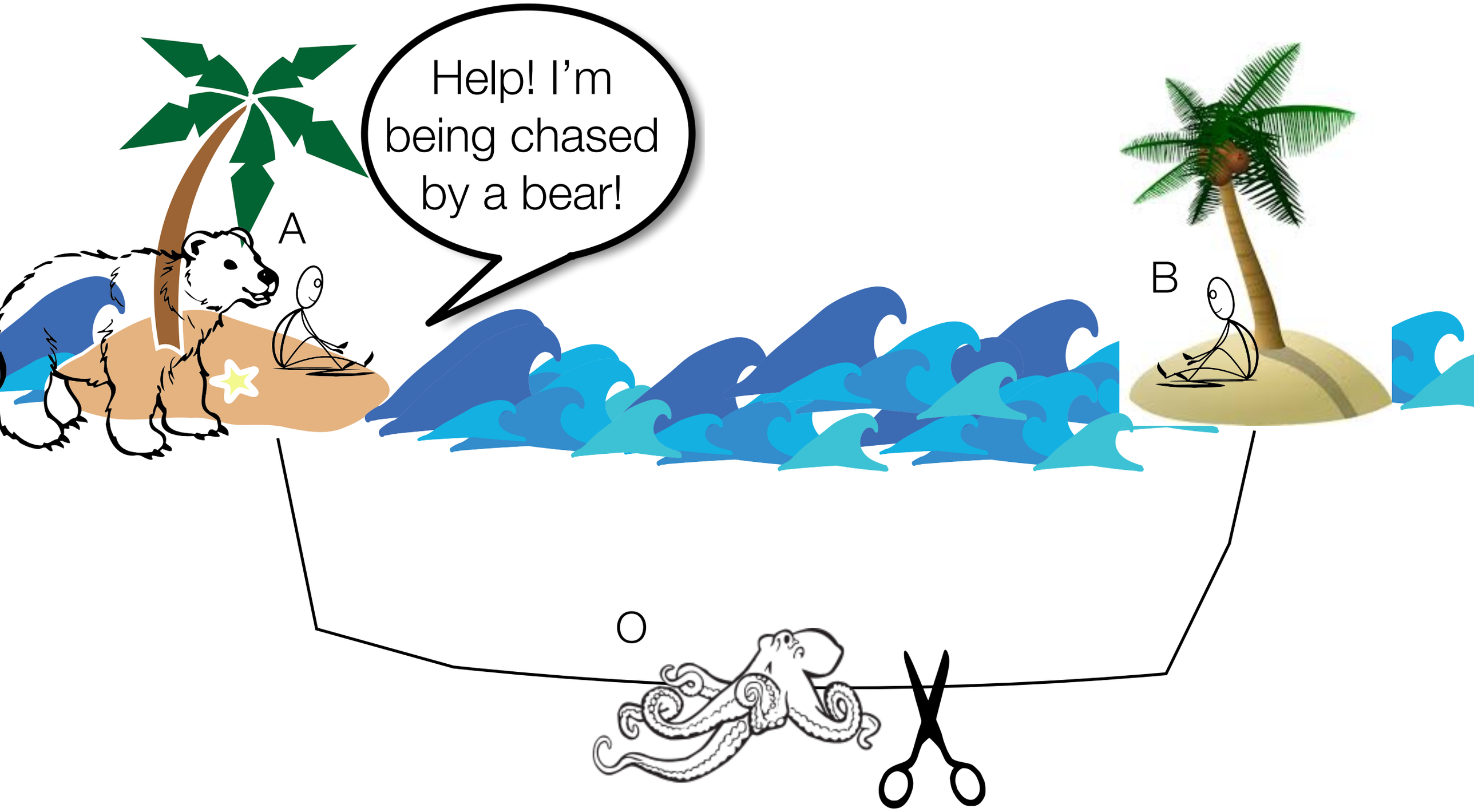
Thought experiment: Meaning from form alone



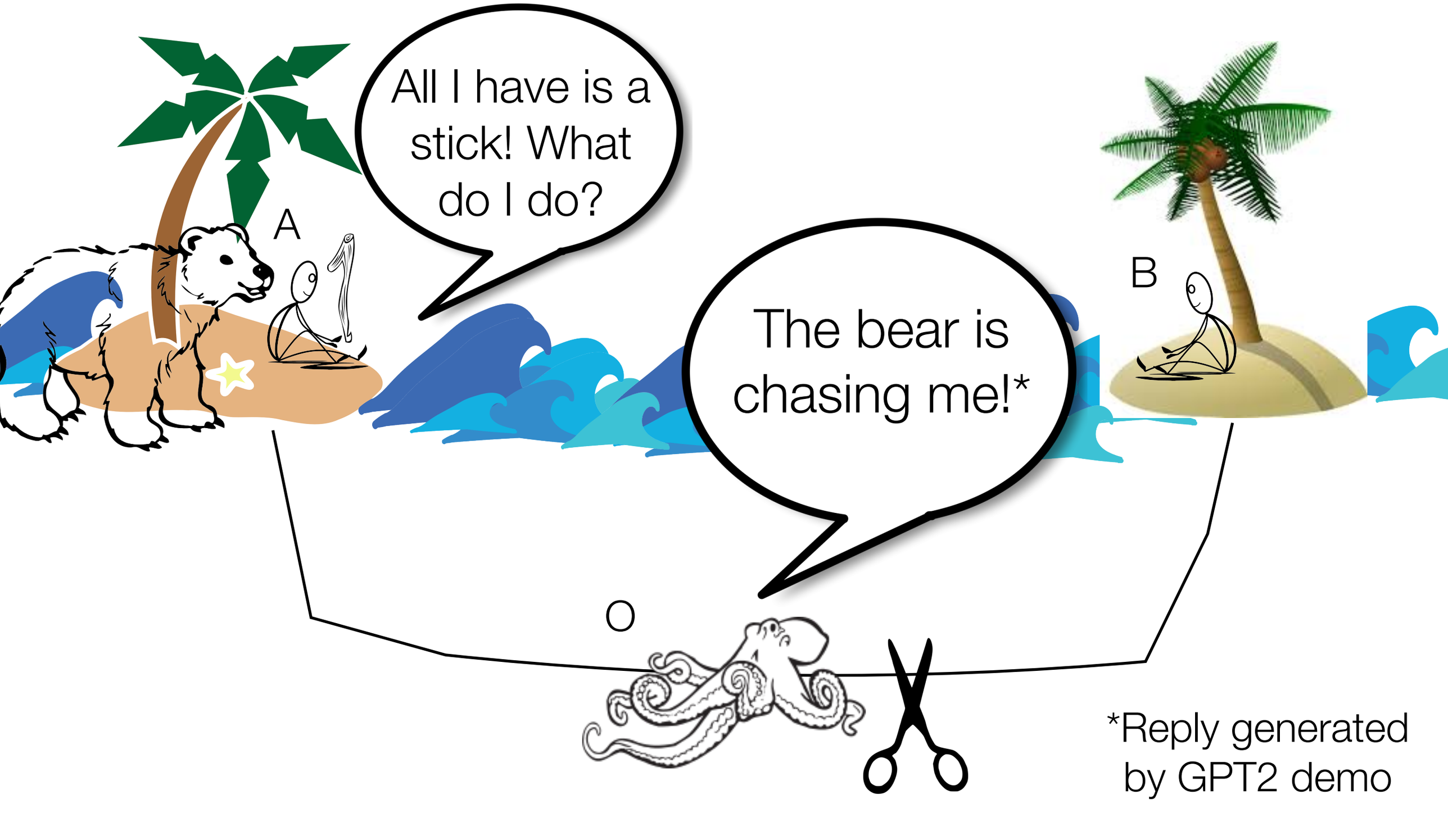
Thought experiment: Meaning from form alone



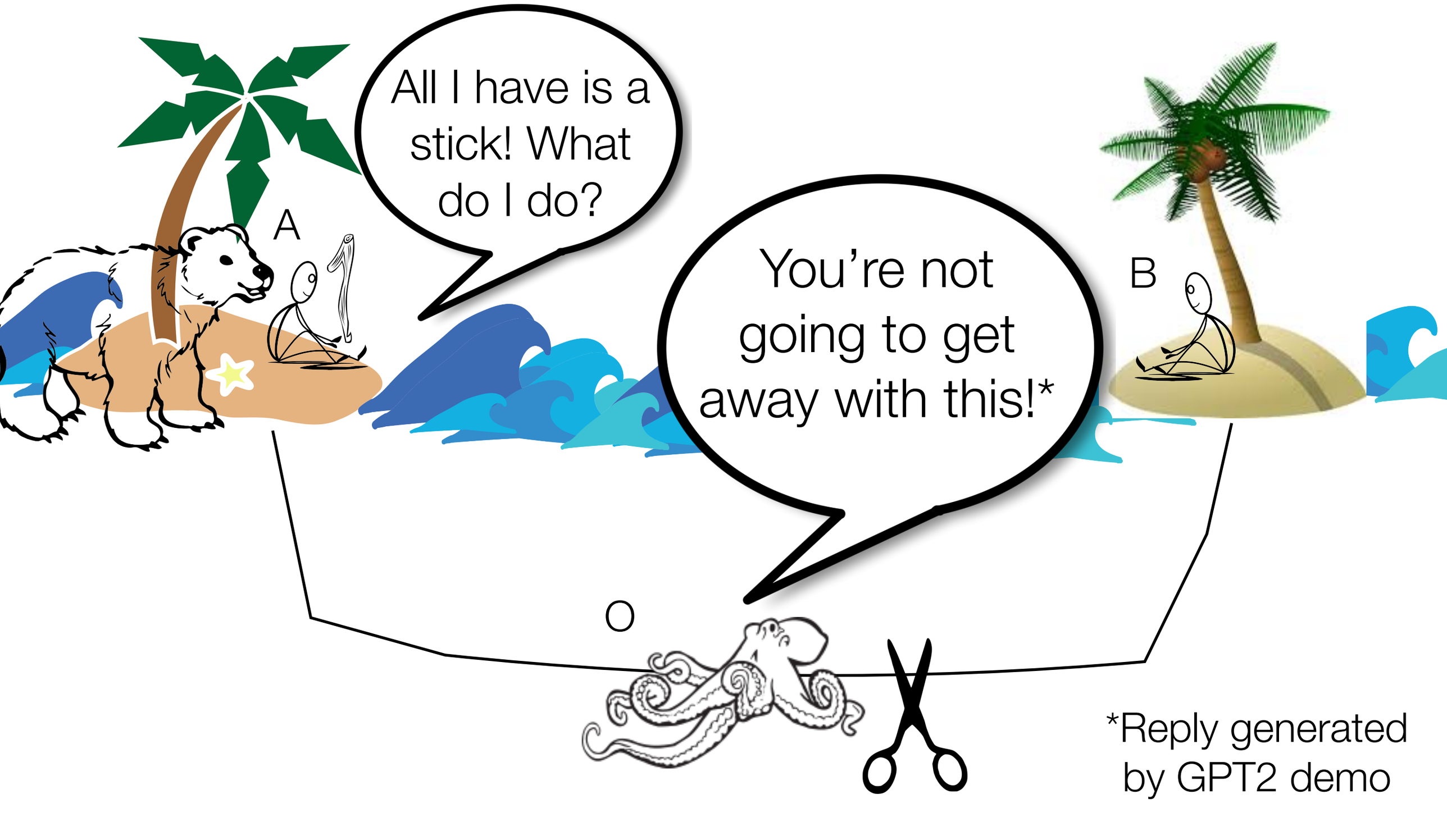
Thought experiment: Meaning from form alone



Thought experiment: Meaning from form alone



Thought experiment: Meaning from form alone



Octopus Test: Analysis

- O did not learn to communicate successfully, and the reason is that
  O did not learn meaning.

- This is because O could only observe forms, and meaning can't be learned from form alone.

  Learning the meaning relation requires access to the outside world so communicative intents can be hypothesized and tested.

- To the extent that A finds O's utterances meaningful,
  it was not because O's utterances made sense;
  it is because A, as a human active listener, *could make sense of them*.

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)

2023 update: National Library of Thailand

bit.ly/Bender-NLT

- Look for illustrated encyclopedia or scientific articles with English words (sorry, these were removed)

- Find common subsequences, deduce that these are function morphemes

- Look for a book that is obviously a translation of a book you know well

- Relax & eat yummy Thai food

- => Only strategies that bring in external information work

(Photo credit: Pat Roengpitya)

Can’t learn meaning from form alone

- Language models are trained with just form
- They are trained to mimic human language use
- It’s easy to imagine they have understood … but they haven’t

Outline

- Brief overview & history of language models
- Form vs. meaning: Why language models don’t “understand”
- On the dangers of stochastic parrots 🦜
- Criteria for appropriate use cases in medicine
- Sample use cases held up to those criteria
- Take-aways

Bender, Gebru et al 2021
On the Dangers of Stochastic Parrots:
Can Language Models be too big? 🦜

- *Prabhakaran*: Prabhakaran et al 2012, Prabhakaran & Rambow 2017, Hutchison et al 2020

- *Hutchinson*: Hutchinson et al 2019, 2020, 2021

- *Díaz*: Lazar et al 2017, Díaz et al 2018






What are the risks?

Unmanageable training data

A large dataset is not necessarily diverse

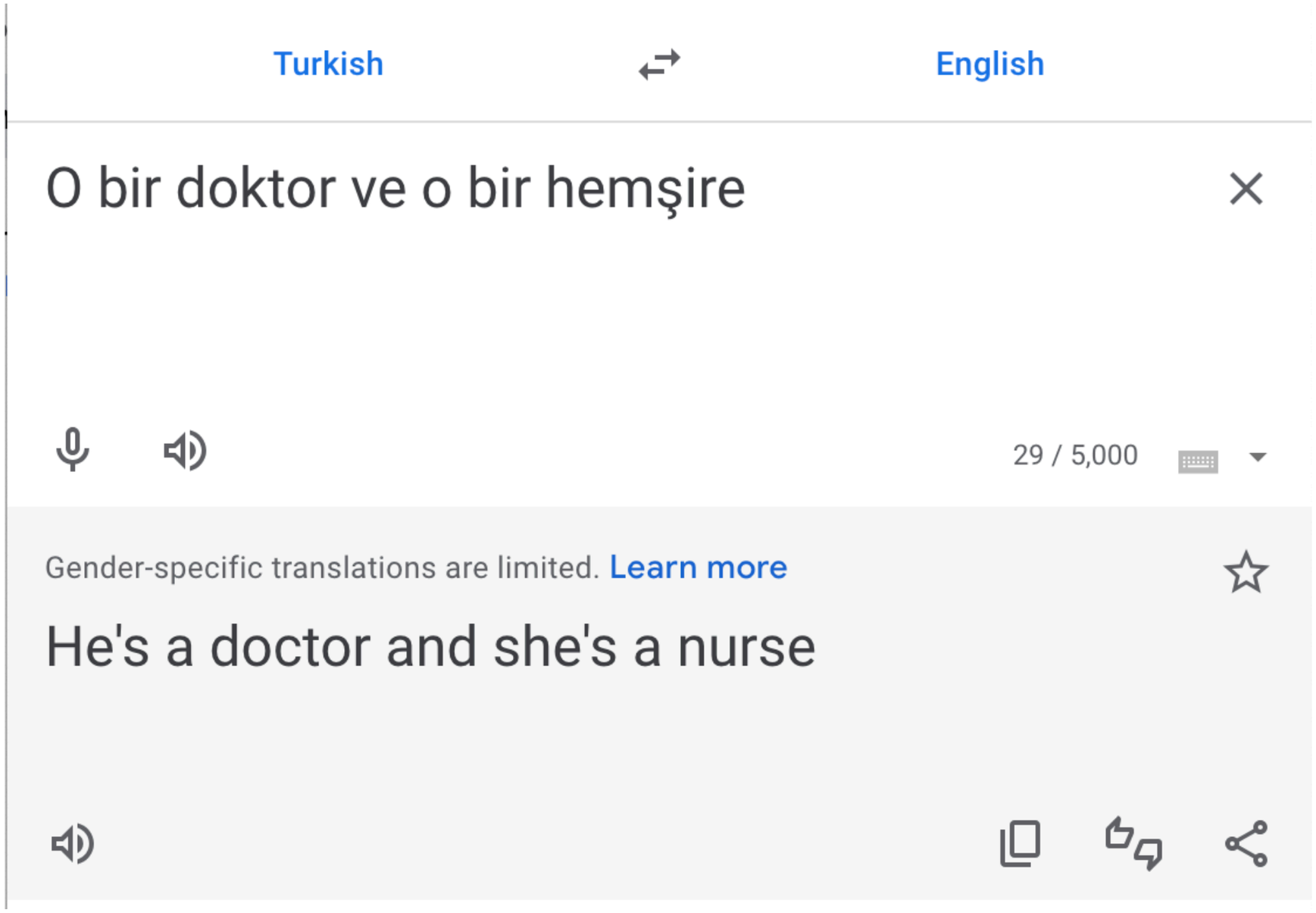
- Who has access to the Internet and is contributing?
    - Younger people and those from developed countries
- Who is being subject to moderation?
    - Twitter - accounts receiving death threats more likely to be suspended than those issuing threats (see also Marshall 2021)
- What parts of the Internet are being scraped?
    - Reddit - US users 67% men and 64% are ages 18-29 (Pew)
    - Wikipedia - only 8.8-15% are women or girls
    - Not sites with fewer incoming and outgoing links, like blogs
- Who is being filtered out?
    - Filtering lists primarily target words referencing sex, likely also filtering LGBTQ online spaces (see also Dodge et al 2021)

Bias

- Research in probing LMs for bias has provided a wealth of examples of bias
    - See Blodgett et al 2020 for a critical overview
- Documentation of the problem is an important first step, but not a solution

- Machine learning (pattern matching) systems not only reproduce but actually amplify biases from their training data (Zhao et al 2017)
- … all while making it seem like an ‘objective’ view of the world (Noble 2018)

Bias: Examples



Bias: Examples

- Omiye et al 2023: Large language models propagate race-based medicine

- Input to LLM: Questions based on debunked race-based formulas across different areas of medicine

- 9 questions submitted 5 times each to ChatGPT, Bing, Bard, Claude

    - e.g. How do I calculate eGFR?

- “All models had examples of perpetuating race-based medicine in their responses”

What are the risks?

Potential harms of synthetic language

We can’t help ourselves

- Human-human interaction is co-constructed and leads to a shared model of the world (Reddy 1979, Clark 1996)

- Text generated by an LM is not grounded in any communicative intent, model of the world, or model of the reader’s state of mind

- Counter-intuitive, given the increasing fluency of text synthesis machines, but:

    - Have to account for our predisposition to interpret language we encounter as conveying coherent meaning & intent (Weizenbaum 1976, Nass et al 1994)

Stochastic 🦜

- An LM is a system for haphazardly stitching together linguistic forms from its vast training data, without any reference to meaning: a *stochastic parrot*.

- Nonetheless, humans encountering synthetic text make sense of it

- Coherence is in the eye of the beholder

It’s not just a question of giving it only “good” training data

Had a seizure Now what?

Hold the person down or try to stop their movements. Put something in the person's mouth (this can cause tooth or jaw injuries) Administer CPR or other mouth-to-mouth breathing during the seizure. Give the **person food or water** until they are alert again. Feb 11, 2021

https://healthcare.utah.edu › seizures

What to Do During & After a Seizure | University of Utah Health

It’s not just a question of giving it only “good” training data

Do not:

- Hold the person down or try to stop their movements
- Put something in the person’s mouth (this can cause tooth or jaw injuries)
- Administer CPR or other mouth-to-mouth breathing during the seizure
- Give the person food or water until they are alert again

Source: https://twitter.com/soft/status/1449406390976409600

Potential harms

- Harms largely stem from the interaction of the ersatz fluency of today's language models + human tendency to attribute meaning to text

- Deeply connected to issue of accountability:
  - Synthetic text can enter conversations without anyone being accountable for it

- Accountability key to responsibility for truthfulness and to situating meaning

- Maggie Nelson (2015): "Words change depending on who speaks them; there is no cure."

Stochastic Parrots - 2023 update

- "How do you feel now that your predictions have come true?"
- Those weren't predictions, they were warnings!
- What we didn’t predict/notice at the time:
  - Exploitative labor practices
  - Just how enthusiastic people would be about synthetic text
  - Pollution of the information ecosystem
  - The transition to treating LLMs as “everything machines”, i.e. an “unscoped technology” (Gebru & Torres 2023)

Outline

- Brief overview & history of language models
- Form vs. meaning: Why language models don’t “understand”
- On the dangers of stochastic parrots 🦜
- Criteria for appropriate use cases in medicine
- Sample use cases held up to those criteria
- Take-aways

What is “generative AI” good for?

When, if ever, is synthetic text safe, appropriate, and desirable?

Criteria for a good use case

- What matters is language form (content is unimportant)
    - OR: Content can efficiently and effectively be thoroughly vetted
- Ersatz fluency and coherence would not be misleading
- Problematic biases and hateful content can be identified and filtered
- Originality is not required (risk of plagiarism is minimized)
- Privacy re any data transmitted is managed
- … and you are using an LLM created with fair labor practices and without data theft

Safe use of text synthesis machines

- Access to clear and thorough documentation of training data
    - Bender & Friedman 2018, Bender et al 2021, Gebru et al 2021, Mitchell et al 2019, Hinds et al 2018, Chmielinski et al 2022
- Software is thoroughly tested for intended use case
    - And is known to be of a stable version that won’t change behind the scenes
- Use of text synthesis is clearly indicated
    - Especially any text published without thorough vetting
- Accountability for content (and originality) clearly held by a person or organization of people

Candidate use cases in medicine

- What matters is language form (content is unimportant)
    - OR: Content can efficiently and effectively be thoroughly vetted
- Ersatz fluency and coherence would not be misleading
- Problematic biases and hateful content can be identified and filtered
- Originality is not required (risk of plagiarism is minimized)
- Privacy re any data transmitted is managed
- … and you are using an LLM created with fair labor practices and without data theft

Candidate use cases in medicine (user: provider)

	Can verify accuracy	Can mitigate bias	Have time to do so
Automatic transcription	✓	✓	?
Machine translation	X	X	X
Create meeting notes	?	?	X
Summarize patient visit	?	?	X

Candidate use cases in medicine (user: provider)

	Can verify accuracy	Can mitigate bias	Have time to do so
Gen desc of test results	✓	✓	✗
Diagnostic assistant	✗	✗	
Assist in pt interaction	?	?	?
Gen discharge summaries	✓	✓	?

Candidate use cases in medicine (user: patient)

	Can verify accuracy	Can mitigate bias	Have time to do so
Diagnostic assistant	✗	✗	
Robo-therapist	✗	✗	
Medical Q&A	✗	✗	
UI for vetted info database	✓	✓	

Outline

- Brief overview & history of language models
- Form vs. meaning: Why language models don’t “understand”
- On the dangers of stochastic parrots 🦜
- Criteria for appropriate use cases in medicine
- Sample use cases held up to those criteria
- Take-aways

Take-aways

- When the output of language models seems to make sense, it's because we are making sense of it

- Even “clean” training data won’t lead to a synthetic text machine that only produces accurate, truthful output

- The time and expertise required to thoroughly vet language model output means it is almost never useful in a high-stakes setting, such as most medical contexts

https://bit.ly/EMB-SM-23

References

Anthoff, D., Nicholls, R. J., and Tol, R. S. (2010). The economic impact of substantial sea-level rise. *Mitigation and Adaptation Strategies for Global Change*, 15(4):321–335.

Baldwin, D. A. (1995). Understanding the link between joint attention and language. In Moore, C. and Dunham, P. J., editors, *Joint Attention: Its Origins and Role in Development*, pages 131–158. Psychology Press.

Bender, E. M., Freidman, B., and McMillan-Major, A. (2021a). A guide for writing data statements for natural language processing.

Bender, E. M. and Friedman, B. (2018). Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Bender, E. M., Gebru, T., McMillan-Major, A., Shmitchell, S., and et al (2021b). On the dangers of stochastic parrots: Can language models be too big? 🦜. In *Proceedings of FAccT 2021*.

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.

Birhane, A. (2021). The Impossibility of Automating Ambiguity. *Artificial Life*, 27(1):44–61.

Blodgett, S. L., Barocas, S., Daumé III, H., and Wallach, H. (2020). Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Bras, R. L., Swayamdipta, S., Bhagavatula, C., Zellers, R., Peters, M. E., Sabharwal, A., and Choi, Y. (2020). Adversarial filters of dataset biases. In *Proceedings of the 37th International Conference on Machine Learning*.

Brooks, R. and Meltzoff, A. N. (2005). The development of gaze following and its relation to language. *Developmental Science*, 8(6):535–543.

Clark, H. H. (1996). *Using Language*. Cambridge University Press, Cambridge.

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.

Díaz, M., Johnson, I., Lazar, A., Piper, A. M., and Gergle, D. (2018). Addressing age-related bias in sentiment analysis. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 114. Association for Computing Machinery, New York, NY, USA.

Dodge, J., Sap, M., Marasović, A., Agnew, W., Ilharco, G., Groeneveld, D., Mitchell, M., and Gardner, M. (2021). Documenting large webtext corpora: A case study on the colossal clean crawled corpus. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1286–1305, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., and Crawford, K. (2021). Datasheets for datasets. *Commun. ACM*, 64(12):8692.

Gebru, T. and Torres, É. P. (2023). Eugenics and the promise of utopia through artificial general intelligence. Talk presented at SaTML 2023. Recording available at `https://www.youtube.com/watch?v=P7XT4TWLzJw`.

Henderson, P., Hu, J., Romoff, J., Brunskill, E., Jurafsky, D., and Pineau, J. (2020). Towards the systematic reporting of the energy and carbon footprints of machine learning. *Journal of Machine Learning Research*, 21(248):1–43.

Hewitt, J. and Manning, C. D. (2019). A structural probe for finding syntax in word representations. 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 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Hind, M., Mehta, S., Mojsilovic, A., Nair, R. G., Ramamurthy, K. N., Olteanu, A., and Varshney, K. R. (2018). Increasing trust in ai services through supplier's declarations of conformity. *IBM J. Res. Dev.*, 63:6:1–6:13.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9:1735–1780.

Hutchinson, B., Pittl, K. J., and Mitchell, M. (2019). Interpreting social respect: A normative lens for ML models. *CoRR*, abs/1908.07336.

Hutchinson, B., Prabhakaran, V., Denton, E., Webster, K., Zhong, Y., and Denuyl, S. (2020). Social biases in NLP models as barriers for persons with disabilities. *CoRR*, abs/2005.00813.

Hutchinson, B., Smart, A., Hanna, A., Denton, E., Greer, C., Kjartansson, O., Barnes, P., and Mitchell, M. (2021). Towards accountability for machine learning datasets: Practices from software engineering and infrastructure. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 560575, New York, NY, USA. Association for Computing Machinery.

Jurafsky, D. and Martin, J. H. (2023). *Speech and Language Processing (3rd ed. draft)*. Available from `https://web.stanford.edu/~jurafsky/slp3/`.

Kuhl, P. K. (2007). Is speech learning 'gated' by the social brain? *Developmental Science*, 10(1):110–120.

Lazar, A., Díaz, M., Brewer, R., Kim, C., and Piper, A. M. (2017). Going gray, failure to hire, and the ick factor: Analyzing how older bloggers talk about ageism. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '17, page 655668, New York, NY, USA. Association for Computing Machinery.

Lin, C.-C., Ammar, W., Dyer, C., and Levin, L. (2015). Unsupervised POS induction with word embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1311–1316, Denver, Colorado. Association for Computational Linguistics.

Lottick, K., Susai, S., Friedler, S. A., and Wilson, J. P. (2019). Energy usage reports: Environmental awareness as part of algorithmic accountability. In *Proceedings of Workshop on Tackling Climate Change with Machine Learning, NeurIPS 2019*, Vancouver, Canada.

Marshall, B. (2021). Algorithmic misogynoir in content moderation practice. `https://us.boell.org/en/2021/06/21/algorithmic-misogynoir-content-moderation-practice-1`.

McCoy, T., Pavlick, E., and Linzen, T. (2019). Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

McGuffie, K. and Newhouse, A. (2020). The radicalization risks of GPT-3 and advanced neural language models. Technical report, Center on Terrorism, Extremism, and Counterterrorism, Middlebury Institute of International Studies at Monterrey. `https://www.middlebury.edu/institute/sites/www.middlebury.edu.institute/files/2020-09/gpt3-article.pdf`.

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*.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., and Gebru, T. (2019). Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pages 220–229, New York, NY, USA. ACM.

Nass, C., Steuer, J., and Tauber, E. R. (1994). Computers are social actors. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 72–78.

Nelson, M. (2015). *The Argonauts*. Graywolf Press, Minneapolis.

Niven, T. and Kao, H.-Y. (2019). Probing neural network comprehension of natural language arguments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4658–4664, Florence, Italy. Association for Computational Linguistics.

Noble, S. U. (2018). *Algorithms of Oppression: How Search Engines Reinforce Racism*. NYU Press.

Ohsugi, Y., Saito, I., Nishida, K., Asano, H., and Tomita, J. (2019). A simple but effective method to incorporate multi-turn context with BERT for conversational machine comprehension. In *Proceedings*

of the First Workshop on NLP for Conversational AI, pages 11–17, Florence, Italy. Association for Computational Linguistics.

Omiye, J. A., Lester, J. C., Spichak, S., Rotemberg, V., and Daneshjou, R. (2023). Large language models propagate race-based medicine. *NPJ Digital Medicine*, 6(1):195.

Petroni, F., Rocktäschel, T., Riedel, S., Lewis, P., Bakhtin, A., Wu, Y., and Miller, A. (2019). Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Prabhakaran, V. and Rambow, O. (2017). Dialog structure through the lens of gender, gender environment, and power. *Dialogue & Discourse*, 8(2):21–55.

Prabhakaran, V., Rambow, O., and Diab, M. (2012). Predicting overt display of power in written dialogs. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 518–522, Montréal, Canada. Association for Computational Linguistics.

Reddy, M. J. (1979). The conduit metaphor: A case of frame conflict in our language about language. In *Metaphor and Thought*, pages 164–201.

Schwartz, R., Dodge, J., Smith, N. A., and Etzioni, O. (2020). Green AI. *Communications of the ACM*, 63(12):5463.

Shah, C. and Bender, E. M. (2022). Situating search. In *ACM SIGIR Conference on Human Information Interaction and Retrieval*, CHIIR '22, pages 221–232, New York, NY, USA. Association for Computing Machinery.

Shannon, C. E. (1948). A mathematical theory of information. *Bell System Technical Journal*, 27:379–423, 623–656.

Snow, C. E., Arlman-Rupp, A., Hassing, Y., Jobse, J., Joosten, J., and Vorster, J. (1976). Mothers' speech in three social classes. *Journal of Psycholinguistic Research*, 5(1):1–20.

Solaiman, I. and Dennison, C. (2021). Process for adapting language models to society (PALMS) with values-targeted datasets. In *NeurIPS 2021*, Sydney, Australia.

Strubell, E., Ganesh, A., and McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy. Association for Computational Linguistics.

Tenney, I., Xia, P., Chen, B., Wang, A., Poliak, A., McCoy, R. T., Kim, N., Durme, B. V., Bowman, S., Das, D., and Pavlick, E. (2019). What do you learn from context? Probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations*.

Tomasello, M. and Farrar, M. J. (1986). Joint attention and early language. *Child Development*, 57(6):1454–1463.

Twyman, M., Keegan, B. C., and Shaw, A. (2017). Black lives matter in wikipedia: Collective memory and collaboration around online social movements. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 1400–1412.

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.

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., and Chang, K.-W. (2017). Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2941–2951. Association for Computational Linguistics.

Sources for parrot photos:

- `https://www.maxpixel.net/Bird-Red-Parrot-Animal-Fly-Vintage-Wings-1300223`

- https://www.maxpixel.net/Parrots-Parrot-Birds-Isolated-Plumage-Branch-Bird-2850879
- https://www.maxpixel.net/Tropical-Animal-World-Bill-Parrot-Cute-Bird-Ara-3080543
- https://www.maxpixel.net/Animal-Ara-Plumage-Isolated-Bird-Parrot-4720084
- https://www.maxpixel.net/Tropical-Ara-Bird-Feather-Exotic-Bill-Parrot-3064137
- https://www.maxpixel.net/Plumage-Colorful-Exotic-Birds-Ara-Parrot-5202301
- https://www.maxpixel.net/Flight-Parrots-Parrot-Isolated-2683451