Why A.I. Struggles with Negative Words | Otherwords
https://www.youtube.com/watch?v=cp0QhCV5uHw Transcript:
https://dontveter.com/ec/not.pdf

Say you're playing with one of those AI image generators.

And you give it a prompt to generate a room with no elephants in it.

I'd like a room that is empty without elephants, please.

I said do not put an elephant in the room in this image.

What is going on? Researchers and casual users alike have documented that generative AI tools often struggle with negation, the language we use to express absences, untruths, and opposites.

Gen AI models may be able to show you what you do want to see, but they're not as good at not giving you what you don't want to see.

Just look at that last sentence, and you can understand why.

The ways we use negation in language can be pretty complex, but our brains are built to handle all the nuance of negation, whether we know it or not.

I'm Dr. Erica Brozovsky, and this is "Otherwords." In its simplest form, negation is language's Uno reverse card.

If affirmation expresses that something is true or real or present, "The sun is up," "He wears glasses," "I ate some cookies," negation expresses the opposite.

"The sun is not up," "He doesn't wear glasses," "I ate no cookies."

But notice how we add to subtract. Affirmation is basically the default of language.

If I say, "My shirt is blue," I don't have to use any special markers to indicate the thing I'm saying is true.

But in hundreds of languages around the world, negation requires us to add words, phrases, or particles of language in order to flip meaning on its head.

Some of the most recognizable examples in English are words like no, not, and none.

We add these in front of or after verbs or adjectives to negate their meaning. My shirt is not blue.

We can also negate with affixes, those little particles we place at the start and/or middle of words to change their meaning.

There are negative prefixes like impossible, unbelievable, non-negotiable, and suffixes like the less in heartless, hopeless, or toothless.

Adding on to language in order to negate its meaning exists in sign languages too.

In American Sign Language, signers can negate words by shaking their head no while signing or reversing the orientation of a sign, like the signs for want and for not want.

Over a hundred languages use multiple negation items in a single sentence or phrase.

In French, verbs are negated by adding negation words on either side of the verb, like in "Je ne parle pas français."

And in English dialects like Appalachian English and African-American English, negating words can stack together, as in you ain't never lied.

Here, the negation words don't cancel each other out, but instead emphasize each other. Linguists call this negative concord.

In the examples we've seen so far, negation acts as the polar opposite of affirmation.

Possible, impossible, blue, not blue, elephant, no elephant. Shouldn't that be easy for a large language model to understand?

I mean, computers are usually pretty good at binaries, but negation isn't always so black and white.

Sometimes we use it to express ideas that are a little fuzzier.

In one study, researchers asked participants to rate negated phrases on a scale of one to 10.

So if one means frigid and 10 means boiling, where would they place phrases like, "The coffee is not hot"?

They found the subjects interpreted phrases like not hot or not good to simply mean less hot or less good.

This suggests that negation words like no, not, and none aren't simply on-off switches for whatever word or phrase that follow.

Instead, our brains use context to determine whether negation inverts the meaning or simply adjusts its intensity.

Multiple negatives can also be used to add nuance.

So if I ask you, "Don't you love this song?" And you answer, "I don't not love it," those two negatives don't cancel each other out or intensify the negation.

Instead, I might interpret them to mean something more positive than hate, but a little less than love.

All this complexity means negation is one of the trickiest features of language to pick up.

Think about how we learn language by first describing the world around us, our family members, our names, objects, colors, shapes. It's a bit of an abstract leap to talk about what isn't there or isn't happening.

Children go through multiple stages of acquiring and learning to produce negation.

If you've ever met a toddler, you know they pick up the word no both quickly and enthusiastically, but they usually use this type of negation as a standalone response to another speaker.

Around ages two to three, kids may start to understand phrases like "The elephant is not in the room."

But it's not until around age four or five that they start to comprehend more abstract uses of negation, like, "An elephant is not a monkey."

Psychologists studying language comprehension found that negation takes longer for our brains to process.

One strategy our brains used to understand negation is to recode a negative statement to a positive one.

Say, transforming "The door is not closed" to "The door is open" in order to verify the statement.

If the negation is more open-ended like, "The door is not red," then our brains have to run through more possibilities to recode the sentence into what color the door is.

So how does that compare to our Gen AI models from earlier?

In 2019, a computational linguist at the University of Chicago used psycholinguistic tests to evaluate how a large language model processes linguistic information.

She found that instead of correctly interpreting negation words in the context of a full phrase or sentence, the LLM largely ignored negation entirely.

It kind of makes sense why that happens. An LLM's job isn't to understand language or make meaning. It's to make predictions about what information is most relevant or probable.

Humans though aren't in the business of making the safest, most likely prediction.

As we covered in our episode on the history of computer language, humans don't process language one word at a time.

Instead, we carry around a grammar in our heads, a complex set of usage rules that allows us to interpret how each word, phrase, and particle of our language works together in context.

As part of that grammar, negation allows us to talk about abstract concepts, fine-tune the meaning of our words, and even exercise our imaginations.

All that context allows us to be clearer and more creative in our communication.

Does that mean that humans are still better at language than computers? Well, I'm not not saying that.

Now say you give it a prompt to generate an image of a room with no elephants in it, of a room.

It's like if I'm describing the room I'm standing in, I might mention the camera in front of me or the green screen behind me or the dog next to me.