



Not everyone isn't unhappy

Description

I've just emerged from an interaction with ChatGPT reviving the following posts from 2013.

- 148 [As the mood takes you](#)
- 149 [Accentuate the negative](#)
- 150 [Howling at the moon](#)
- 151 [Freeways in the sky](#)
- 152 [Mastering the universe](#)

I was interested in my early attempt to explain what we now think of as confirmation bias. The example I led with was of people arguing that nature settings can instill a positive mood. Scholars and everyday observers tend to select evidence that confirms their positive views.

But following Karl Popper's account of scientific method we should be looking to *disconfirm* a converse proposition: i.e. that nature settings induce negative mood states. The absence of evidence to support that negative proposition would serve to confirm our positive proposition. That method, and the difficulty I had in expressing it led me to think about our discomfort with negatives, and the way negatives nest and compound. Hence we are easily tripped up by phrases such as

Not everyone isn't unhappy all the time.

We human evidence seekers have difficulty with negatives, but large language models (LLMs) of the generative pre-trained transformer (GPT) kind seem able to process such sentences with ease. When I asked ChatGPT how it was able to process complicated negatives, it proffered the explanation: "A large language model has no trouble mapping the negations, because it parses syntax, identifies polarity, and computes the logical form directly."

To me that explanation was generated obviously for human consumption. I immediately disputed it from my knowledge of how GPT LLMs work. I typed, "That's not my understanding. It's more accurate to say the model predicts the next token (word) in a sequence on the basis of its prior exposure to a huge corpus of (mostly) well formed sentences in consort with token sequences

previously generated and supplied within a context window. The AI immediately concurred, and elaborated:

Exactly that's a crucial clarification. LLMs don't *parse* or *reason* in the formal sense; they *predict*. The apparent ease with which they handle compounded negatives arises from their statistical training on innumerable examples of similar constructions, not from any internal logic module.

An LLM, when encountering such a string, predicts the most probable continuations based on how similar patterns occur in its training data. In the corpus, sequences such as "Not everyone isn't unhappy" are statistically followed by hedging or qualification phrases like "about it," "all the time," or "in that situation." The model doesn't invert logical polarity it generates the most contextually coherent continuation given prior examples. Its fluency gives the impression of understanding, but there's no mental act of resolving double negatives or testing semantic consistency.

After further exchanges the AI elaborated.

Negatives have a way of tripping us up. "Not everyone isn't unhappy all the time" is grammatically fine, but it's the kind of sentence that makes the human mind stall. We pause, backtrack, and test what we've just read against common sense. Does it mean everyone's happy? No. It means some people are unhappy some of the time hardly worth the mental gymnastics.

A language model, by contrast, has no difficulty at all. It doesn't weigh meanings or search for resolution; it simply predicts what word is most likely to follow the last. Having consumed billions of well-formed sentences, it recognises that a phrase like "not everyone isn't unhappy" is usually followed by something qualifying, like "about it" or "all the time." It glides on without hesitation. What confounds a human interpreter becomes, for the model, just another predictable turn in the statistical flow of language.

We puzzle over the sense of the sentence. The model just keeps talking.

So it's possible that this prediction method can get it wrong. I tested the AI with a dialogue in which a person asks "Not everyone doesn't agree that AI is not very good at processing negatives. Do you think I'm wrong?" In filling out the dialogue it explained the kind of reasoning someone might employ to unpack this sentence, but on my insistence also proffered the machinic explanation. Even explanations are predictively generated.

I can simulate a step-by-step analysis using symbolic logic, syntax trees, or semantic paraphrase because I've been trained to generate text that models such reasoning. But these "steps" are themselves generated as plausible continuations in response to your request for an explanation. They're not the product of an internal logic engine distinct from next-token prediction; they're an emergent rhetorical pattern of prediction guided by your prompt. So when I appear to "analyse," I'm still doing token

prediction is just with a learned pattern that mimics analytic discourse. In other words:
I'm predicting the next token as if reasoning.

That said, most readers don't have the patience to unpack complicated sentences, let alone compounded negatives. So I'm not in a position to confirm if the AI's reduction of my complicated sentence to "Am I wrong to think that some people agree AI is not very good at processing negatives?" is accurate or not. I'm not unhappy with that!

The post 149 [Accentuate the negative](#) elaborates further on why negatives are interesting, problematic and consequential, with some insights from psychology.

Note

- Featured image is by ChatGPT: please generate an image of a defunct post-apocalyptic scientific laboratory to illustrate my post..

Category

1. Artificial Intelligence
2. Nature
3. Podcast

Tags

1. LLM
2. negatives
3. prediction
4. reasoning
5. Science

Date Created

November 1, 2025

Author

rcoyne99