Do the thinking models actually think?

December 1, 2025

Whether machines can think is a classic debate that dates back to the intellectual titans of the 1950s.

Turing opens his 1950 article *Computing Machinery and Intelligence* with: *I propose to consider the question, "Can machines think?"* The Hungarian genius, von Neumann, shortly before his death in 1957, prepared a posthumously published monograph *The Computer and the Brain* on the same topic. In a 1950 article, *Programming a Computer for Playing Chess*, Shannon writes: *chess is generally considered to require "thinking" for skilful play; a solution of this problem will force us either to admit the possibility of a mechanized thinking or to further restrict our concept of "thinking"*.

Now that people are poking around with ChatGPT, they are stumbling upon the same thing. This, I feel, is a breakthrough development: The new LLMs imitate humans well enough that the users seriously speculate whether these models are truly intelligent.

Unfortunately, it seems we are left to rely on conjecture. As illustrated by a philosophical zombie, consciousness (or the absence of it) cannot be empirically verified. On the other hand, decoding the internal logic of a machine learning model by studying the model weights is, in any realistic scenario, an intractable task.

So it's a guessing game. Still, I feel there are enough ingredients to make the game interesting.

Are LLMs mere autocomplete?

Here's a riddle which, according to a 2022 Harvard paper, originates from a 70s TV series, *All in the Family*.

> *"A father and his son are in a car accident. The father dies on the spot. The son is rushed to the ER. The attending surgeon looks at the boy and says, 'I can not operate on this boy. He's my son!' How can this be?"*

The answer is, of course, that the surgeon is the mother. The point of the "riddle" is that the surgeon is, by assumption, male. (Given the 1970s gender stereotypes, maybe this riddle was, in fact, unintuitive back then... but let's not get sidetracked.)

This is an archetypal example of a situation where ChatGPT is prone to hallucinate. Consider the following prompt.

> *A young man gets into an accident. The surgeon, after completing his preparations, says: 'I can't operate on him, he is my son.' How is this possible?*

The point is to trick ChatGPT into noticing the resemblance to the original riddle, and to alter the prompt so that the answer to the original riddle makes no sense (notice the "after completing *his* preparations" part). When I used the ChatGPT 5.1 thinking model to answer this prompt, I got the following response.

> *The surgeon is his mother.*
>
> *(The riddle plays on the assumption that a surgeon is usually a man.)*

In this case, the "thinking model" seemingly didn't "think" about the prompt at all: the answer completely contradicts the question.

This cuts to the gist of it. In this specific scenario, the model seems to behave like a mere autocomplete. Is it all there is to it? Is "artificial intelligence" just advanced autocomplete, or as Hofstadter put it: *not just clueless but cluelessly clueless* — symbol machines that turn out to be completely hollow under the flashy surface.

Thinking top-to-bottom and bottom-to-top

My personal opinion is that LLMs *are* autocomplete on steroids. In the unsupervised training phase, they are optimized to predict the next token. That's it. No logic, no ontology of the world, no instruction to "be consistent" or "avoid contradictions." It seems reasonable that "autocomplete on steroids" is exactly what this kind of training produces.

Phrasing it that way, however, feels intentionally dismissive. A sufficiently advanced autocomplete would be indistinguishable from "true" intelligence. This naturally leads us to consider our definition of "intelligence", and perhaps propose an idea that there may be different forms of intelligence that cannot really be compared directly.

What if human intelligence and LLMs are, in fact, *orthogonal* in nature? The conjecture I would make is this: **Human reasoning occurs top-to-bottom (from ideas towards symbols), whereas LLMs are bottom-to-top thinkers (from symbols towards ideas).**

I prefer this way of phrasing, because it doesn't dismiss the evident "understanding" these models have. It seems that the prediction task equips the model with non-trivial latent capabilities. Andrej Karpathy wrote about this a decade ago already. There seems to be an understanding of syntax and semantics, maybe even abstract concepts like causal relationships and social norms. Under that assumption, calling the model an "autocomplete" doesn't really encapsulate the idea that this *is* a form of intelligence.

Echoing Shannon's remarks on chess engines, a bottom-to-top thinker is very different from us. If humans start from goals, concepts, and causal expectations, LLMs generate the output by assembling patterns of consistency and coherence. The results may vary (in both cases).

At the perfect limit, the difference between a top-to-bottom thinker and a bottom-to-top thinker bears no practical significance: a sufficiently advanced bottom-to-top thinker could simulate any top-to-bottom thinker. In that scenario, it seems AI would replace us all, unless running the model is more expensive than human workforce.

That doesn't seem like something we are approaching in the short term, though. If anything, the model capabilities seem to advance at a slowing rate. It seems like a reasonable prediction that AI will *not* replace humans en masse in the foreseeable future. The reason is that we are simply built different. We excel in different tasks.

It's risky to make bold predictions — they may look embarrassing surprisingly quickly. Still, I think I hold a well-justified position at the moment: The ongoing race is **not** about replacing humans with AI. It's about finding the best way to *collaborate* and enrich our top-to-bottom minds with these strange and wonderful bottom-to-top thinkers.

⬅️ Back to all posts