

Three Things That LLMs Have Made Us Rethink

December 8, 2023 — Essays

rodneybrooks.com/three-things-that-llms-have-made-us-rethink/

There are three things that the existence of LLMs, such as ChatGPT-3.5 and ChatGPT-4 make us have to rethink. At different times and amongst different communities they have all had lots of AI researchers talking about them, often with much passion.

Here are three things to note:

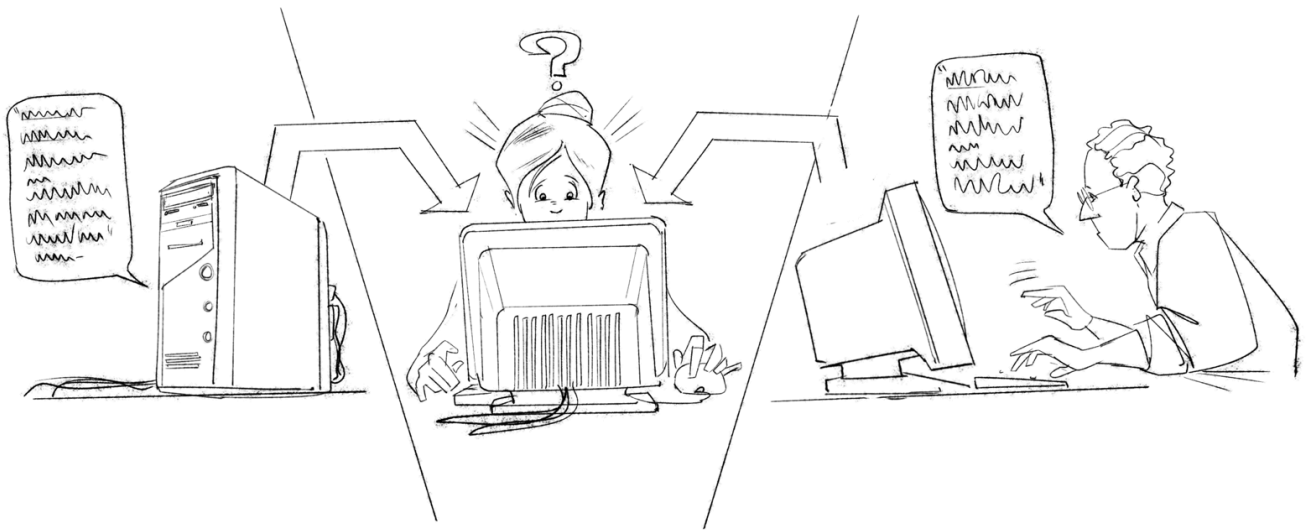
1. The Turing Test has evaporated.
2. Searle's Chinese Room showed up, uninvited.
3. Chomsky's Universal Grammar needs some bolstering if it is to survive.

We'll talk about each in turn.

The Turing Test

In a 1950 paper titled [Computing Machinery and Intelligence](#), Alan Turing used a test which involved a human deciding whether an entity that the person was texting with was a human or a computer. Of course, he did not use the term "texting" as that had not yet been

invented, rather he suggested that the communication was via a "teleprinter", which did exist at the time, where the words typed in one location appeared on paper in a remote location. "Texting" is the modern equivalent.



Turing used this setup as rhetorical device to argue that if you could not accurately and reliably decide whether it was a person or a computer at the other end of the line then you had to grant that a machine could be intelligent. His point was that it was not just simulating intelligence but that it would actually be intelligent if people could not tell the difference.

Turing said;

I believe that in about fifty years' time it will be possible to programme computers, with a storage capacity of about 10^9 , to make them play the imitation game so well that an average interrogator will not have more than 70 per cent, chance of making the right identification after five minutes of questioning.

His number 10^9 referred to how many bits of program would be needed to achieve this result, which is 125MB, i.e., 125 Mega Bytes. Compare this with ChatGPT-3.5 which has 700GB, or 700 Giga Bytes, of weights (175 billion 32 bit weights) that it has learned, which is almost 6,000 times as much.

His paragraph above continues:

The original question, 'Can machines think!' I believe to be too meaningless to deserve discussion. Nevertheless I believe that at the end of the century the use of words and general educated opinion will have altered so much that one will be able to speak of machines thinking without expecting to be contradicted.

Despite his goal to be purely a rhetorical device to make the question, 'Can machines think!' (I assume the punctuation was a typo and was intended to be '?') meaningless, this led to people calling the machine/person discernment test the *Turing Test*, and it became the default way of thinking about how to determine when general Artificial Intelligence had been achieved. But, of course, it is not that simple. That didn't stop annual Turing Tests being set up, with entrants from mostly amateur researchers, who had built chat bots designed not to do any useful work in the world, but designed and built simply to try to pass the Turing Test. It was a bit of a circus and mostly not very useful.

Earlier this year I felt like I was not hearing about the Turing Test with regards to all the ChatGPTs, and in fact the scientific press had noticed this too, with [this story](#) in Nature in July of this year:

nature

Explore content ▾

About the journal ▾

Publish with us ▾

Subscribe

[nature](#) > [news feature](#) > article

NEWS FEATURE | 25 July 2023

ChatGPT broke the Turing test – the race is on for new ways to assess AI

Large language models mimic human chatter, but scientists disagree on their ability to reason.

[Celeste Biever](#)

Don't worry, there are still papers being written on the Turing Test and ChatGPT, for instance [this one from October 2023](#), but the fervor of declaring that it is important has decreased.

We evaluated GPT-4 in a public online Turing Test. The best-performing GPT-4 prompt passed in 41% of games, outperforming baselines set by ELIZA (27%) and GPT-3.5 (14%), but falling short of chance and the baseline set by human participants (63%).

In general the press has moved away from the Turing Test. ChatGPT seems to have the sort of language expertise that people imagined some system as intelligent as a

person would have, but it has become clear that it is not the crystalline indication of intelligence that Turing was trying to elucidate.

SEARLE'S CHINESE ROOM

In 1980, John Searle, a UC Berkeley philosopher, introduced the idea of a "Chinese Room", as a way to argue that computers could not be truly intelligent in the way that people are, not truly engaged with the world in the way people are, and not truly sentient in the way people are.

He chose "Chinese" as the language for the room as it was something totally foreign to most people working in Artificial Intelligence in the US at the time. Furthermore its written form was in atomic symbols.

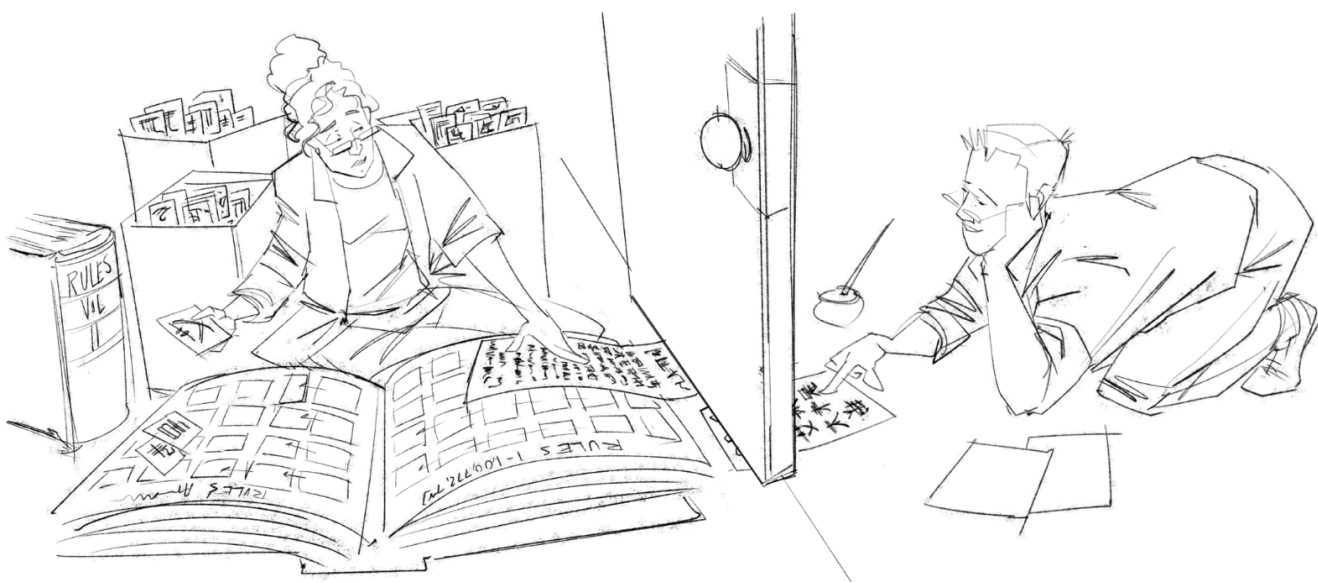
Here is what ChatGPT-3.5 said when I asked it to describe Searle's Chinese Room. I have highlighted the last clause in blue.

Imagine a person who doesn't understand the Chinese language locked inside a room. This person has a set of instructions (a program) written in English that tells them how to manipulate Chinese symbols based on the symbols that are input to the room through a slot. They have no understanding of Chinese and don't know the meanings of the symbols.

From the outside, someone passes messages written in

Chinese through the slot into the room. The person inside the room follows the instructions in the program and produces responses in Chinese based solely on symbol manipulation, without understanding the content or meaning of the messages.

To an observer outside the room, it may appear as though the person inside understands Chinese, as they can generate coherent responses in Chinese. However, the person inside the room doesn't actually understand the language. They are merely following a set of syntactic rules or algorithms for manipulating symbols, without grasping the semantics or meanings of those symbols.



To prove that ChatGPT really is the Chinese Room, I typed in some Chinese text to version 3.5 and it responded with Chinese text:

艾未未是谁？

艾未未是一位知名的中国艺术家、社会活动家和评论家。他以其大胆的艺术创作和公开对政府审查和人权侵犯的批评而

闻名。艾未未的作品涉及雕塑、摄影、行为艺术、装置艺术和社交媒体等多种媒介，展现了对政治、社会和人权议题的关注。他因批评中国政府而多次引起争议，这导致了他与政府当局的冲突，并且在过去曾经被拘留和监禁。他因其独特的艺术作品和公开表达的观点而在国际艺术界和社会活动家圈中备受瞩目。

My question was "Who is Ai Wei Wei?" and it responded in what people tell me is perfectly good Chinese, describing the Chinese artist Ai Wei Wei, his art, and his run ins with the government.

So here we have Searle's Chinese Room. It has been trained on billions of words from the web but has no experience of the world, has never heard Chinese spoken, and has never even read Chinese characters as characters but rather has seen them only as integers along with (I assume) a pre-processing step to map whichever of the five (Unicode, GB, GBK, Big 5, and CNS) common digital Chinese character code standards each document uses to that common set of integers. (The fact that ChatGPT 'knows', in its description of the room, above, that Chinese is written in symbols is not because it has ever seen them, but because it has "read" that Chinese uses symbols.)

The fact that GPT **is** the Chinese Room, the fact that one now exists, means that many of the old arguments for and against Searle's position that he was staking out with his rhetorical version of the room must now be faced squarely

and perhaps re-evaluated. Searle's Chinese Room was a topic of discussion in AI for well over 25 years. Everyone had to have an opinion or argument.

In my book *Flesh and Machines: How Robots Will Change Us* (Pantheon, New York, 2002), I made two arguments that were in opposition to Searle's description of what his room tells us.

Firstly, I argued (as did many, many others) that indeed Searle was right that the person in the room could not be said to understand Chinese. Instead we argued that it was the whole system, the person, the rule books, and the state maintained in following the rules that was what understood Chinese. Searle was using the person as a stand in for a computer fresh off the production line, and ignoring the impact of loading the right program and data on to it. In the ChatGPT case it is the computer, plus the algorithms for evaluating linear neuron models plus the 175 billion weights that are together what make ChatGPT-3.5 understand Chinese, if one accepts that it does. In my book I said that no individual neuron in a human brain can be said to understand Chinese, it has to be the total system's understanding that we talk about. ChatGPT-3.5 is an example of a computer doing the sort of thing that Searle was arguing was not possible, or at least should not be spoken about in the same way that we might speak about a person understanding Chinese.

Secondly, I argued (using Searle as the person in the room

as he sometimes did):

Of course, as with many thought experiments, the Chinese room is ludicrous in practice. There would be such a large set of rules, and so many of them would need to be followed in detailed order that Searle would need to spend many tens of years slavishly following the rules, and jotting down notes on an enormous supply of paper. The system, Searle and the rules, would run as a program so slowly that it, the system, could not be engaged in any normal sorts of perceptual activity. At that point it does get hard to effectively believe that the system understands Chinese for any usual understanding of 'understand'. But precisely because it is such a ludicrous example, slowed down by factors of billions, any conclusions from that inadequacy can not be carried over to making conclusions about whether a computer program running the same program 'understands' Chinese.

Hmm, well my bluff has been call by the existence of ChatGPT. First, note that I was right about the size of the rule set, 175 billion neural weights, that it would take a person effectively forever to follow them. But every modern laptop can hold all those rules in the file system (it is less than a terabyte of memory), and the algorithm is parallel enough that a chunk of processing in the cloud can make ChatGPT run at human language speeds.

If I maintain my above argument from 2002, I would have

to say that ChatGPT does '*understand*' Chinese. But those who have read my writings over the years would guess, rightly, that I don't think it does. Without grounding in physical reality I don't think a machine can understand in the same way we do. ChatGPT is just like someone following rules with no semantic understanding of the symbols, but it does it at the speed my argument above said was necessary for it to really be understanding. But now I'm going to say it still doesn't understand. My old self and my today self are not being intellectually coherent, so I am going to have to think about this some more over the next few years and refine, perhaps rethink, but certainly change in some way what it is I conclude from both Searle and ChatGPT existing.

Other people over the last forty years have argued, and I have agreed, that language in humans is strongly grounded in non-language. So, we have argued that a computer program, like ChatGPT-3.5 could not have a consistent performance level that would seem like human language. ChatGPT-3.5 certainly seems to have such consistent performance, as long as you don't poke it too deep—it certainly has a level that would work for most of your daily interactions with strangers. Our arguments are therefore challenged or broken. I don't yet know how to fix them.

CHOMSKY'S UNIVERSAL GRAMMAR

Nom Chomsky is the most influential linguist ever and he

spent most of his career in MIT's linguistics department. His advocacy of the existence of a universal grammar system in humans, and humans only, has shaped much cognitive science and theories about cognition in both humans and non-humans.

I asked ChatGPT to "explain Chomsky's universal grammar", and it replied:

Noam Chomsky's theory of Universal Grammar (UG) is a hypothesis in the field of linguistics that suggests that the ability to acquire language is innate to humans, and that there is a universal set of grammatical structures that underlies all human languages. Chomsky introduced this concept in the 1950s and has since developed and refined it over the years.

Key points of Chomsky's Universal Grammar include:

Innate Language Faculty: Chomsky argues that humans are born with an innate capacity for language. He proposes the existence of a "language acquisition device" (LAD) or some innate cognitive structures that enable children to rapidly and effortlessly learn the language(s) to which they are exposed during early childhood.

Universal Grammar Principles: Chomsky suggests that all human languages share a common underlying structure or set of principles. These principles are considered to be part of the Universal Grammar and are hard-wired into the human brain. They provide a

framework that constrains the range of grammatical structures and rules that are possible in any language.

Poverty of the Stimulus: Chomsky's argument for the existence of Universal Grammar includes the "poverty of the stimulus" idea. He contends that the linguistic input that children receive is often insufficient or ambiguous, yet they are able to acquire complex grammatical structures. This implies that there must be an innate mechanism guiding language acquisition.

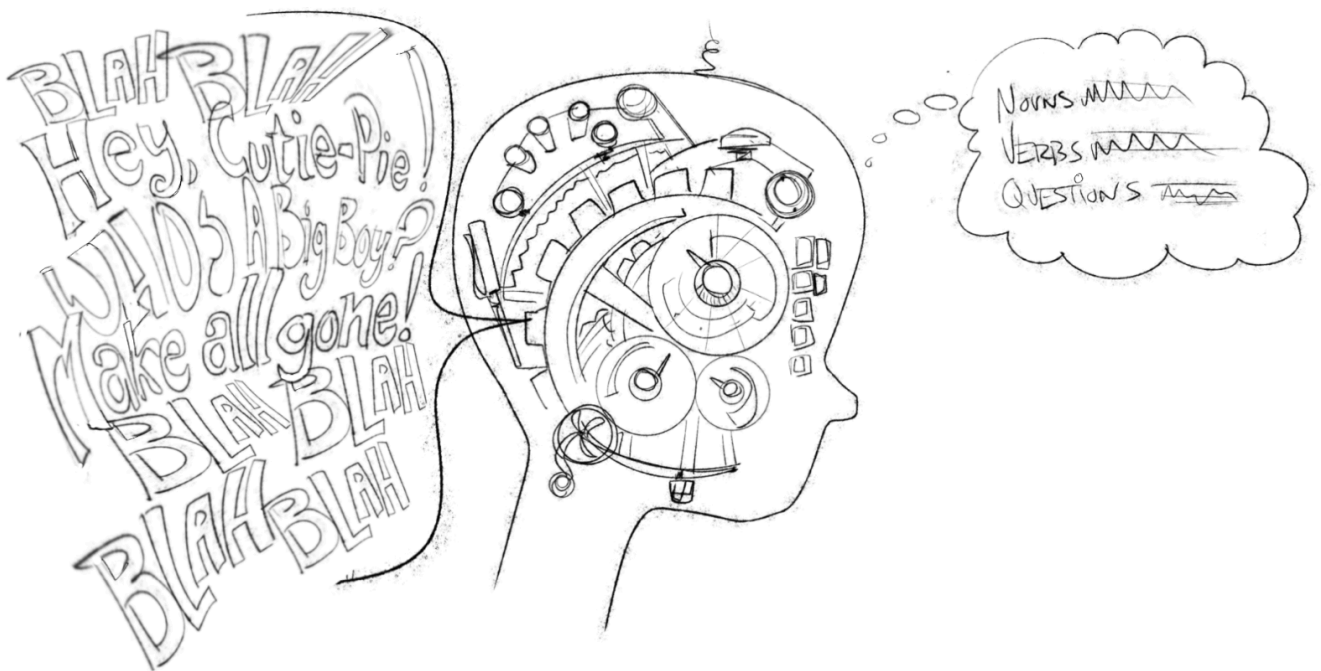
Language Acquisition as Rule-Based: Chomsky posits that the process of language acquisition involves the internalization of grammatical rules. Children, according to this theory, unconsciously deduce the grammatical structures of their native language(s) based on the linguistic input they receive.

It's important to note that while Chomsky's Universal Grammar theory has been influential and has shaped the field of linguistics, it is not without its critics. Some researchers argue for alternative theories of language acquisition that emphasize the role of environmental factors and learning mechanisms. Additionally, the nature and specifics of the supposed universal grammar are still topics of ongoing research and debate within the field of linguistics.

ChatGPT gets it right, this is what Chomsky was arguing for. I highlighted in blue one particular alternate view that

ChatGPT brought up. [It is no surprise to me that ChatGPT made such a good summary—that is the sort of thing it is quite good at, having ingested billions of pages of human written material on all sorts of matters.]

Here is a diagram I had someone draw for me to illustrate the idea of Universal Grammar. A child hears words and acquires both the grammar of the language and the language itself.



I asked: Does ChatGPT have a universal grammar module?

To which ChatGPT replied:

No, ChatGPT does not have a universal grammar module in the same sense as Chomsky's Universal Grammar theory. ChatGPT is based on a different approach known as transfer learning, specifically the GPT (Generative Pre-trained Transformer) architecture developed by OpenAI.

ChatGPT learns from a diverse range of internet text during its pre-training phase, but it doesn't have explicit knowledge or representation of grammatical rules as predefined modules. Instead, it captures patterns, context, and relationships within the data it has been trained on. It doesn't have a built-in understanding of linguistic structures in the way that the concept of Universal Grammar suggests.

The model relies on statistical patterns and associations in the data it has seen during training to generate contextually relevant responses. While it can generate coherent and contextually appropriate text, it doesn't possess an explicit understanding of grammar rules or linguistic universals as proposed by Chomsky's theory.

Again, it got it completely right.

Clearly ChatGPT (and other LLMs) learn human language to a level of performance that we had not previously see in computational systems, a level that we are forced to compare favorably with humans. And they learn language without any sort of universal grammar knowledge that Chomsky argues both constrains and enables what humans can learn as they acquire language. They also learn it without any role for environmental factors that ChatGPT mentioned above in its reference to criticisms of universal grammar, and that I highlighted in blue.

No structures in the human brain that implement a

universal grammar system, even quite distributed, have been identified. ChatGPT certainly doesn't have one. Does this put the universal grammar theory at risk?

Maybe. Maybe not.

Chomsky's argument about the poverty of stimulus for human children may be what saves the need for universal grammar in humans. ChatGPT is exposed to a few orders of magnitude of source material more (and it uses massive more energy to learn) than a human child.

On the other hand the ability to learn human grammar with no mechanism for grammar built in is certainly a surprise, at least to time traveling AI researchers from thirty or even twenty years ago.

Universal grammar enthusiasts have long argued that no other biological species can have language that has grammar and recursive composability, because they don't have universal grammar. Computational LLMs do (have such language) and don't (have universal grammar).

I personally remember very heated arguments on the MIT campus between Chomsky followers and computer scientists working on statistical models in the early 2000's. The arguments came up in computer science faculty hiring meetings. The Chomskians claimed that because of the need for universal grammar there would never be anything useful that came out of statistical approaches to language. ChatGPT has proved them

wrong. (And yes, I personally shouted angrily at Chomskians in some of those hiring meetings.)

The question is whether there is a more narrowly arguable way to push forward on the need for universal grammar in resource and stimulus limited environments.

A Preliminary Conclusion

The existence, for one year now, of functional, available, LLMs has challenged some previous intellectual positions. Those challenges will be debated and made sharper. The existence of those challenges, however, does not necessarily mean that any, or even many, of the wild predictions around for how important and transformative LLMs will be will come to pass. Things have changed forever, but as with many other forever changes in AI over the last 65 years, there are most likely many more things that we will change and that we will learn,

I believe that we are still in the uncertain baby step stages. It is worth repeating the last sentence of Alan Turing's computational machinery paper that we started off with above. It is still as relevant today as it was in 1950.

We can only see a short distance ahead, but we can see plenty there that needs to be done.