Protokoll 13

Wolf Kittler & Fabian Offert

Questions for ChatGPT (Wolf Kittler)

Freudians may call it Narcissism, yet I would rather speak of solipsism: The first questions that occurred to me, when I first submitted ChatGPT to the Turing test, were about my work. For, as Descartes would say, this is only subject I know clearly and distinctly. The answers were equally clear: In the world of ChatGPT, I do not exist.

One question was about the following passage in Claude E. Shannon's *Mathematical Theory of Communication*, on which I had just written a brief commentary:

Two extremes of redundancy in English prose are represented by Basic English and by James Joyce's book Finnegans Wake. The Basic English vocabulary is limited to 850 words and the redundancy is very high. This is reflected in the expansion that occurs when a passage is translated into Basic English. Joyce on the other hand enlarges the vocabulary and is alleged to achieve a compression of semantic content.[1]

Apart from a list of common places about Shannon's and Ogden's work, this is what ChatGPT had to say in particular:

> Overall, while there is no evidence that Claude E. Shannon directly quoted C. K. Ogden, it is possible that he was aware of his work and may have drawn upon some of his ideas indirectly. Shannon is most famous for his contributions to information theory, and his papers and publications primarily focus on mathematical and engineering concepts related to communication and information.
>
> While Shannon was a highly educated individual and had a broad range of interests, there is no evidence to suggest that he was specifically interested in

1 Claude E. Shannon, *The Mathematical Theory of Communication*, The University of Illinois Press: Urbana 1964, pp. 29-125, quotation p. 56.

> or influenced by the works of James Joyce. Therefore, it is unlikely that he would have directly quoted Joyce in his writings.

And on Joyce:

> I am not aware of any direct quotes from James Joyce in the works of Claude E. Shannon.

Another question was about Heinrich von Kleist's drama *Die Herrmannsschlacht*, which Carl Schmitt, in his book *Theorie des Partisanen*, calls the greatest partisan poetry of all time:[2]

> There is no guerrilla warfare depicted in Heinrich von Kleist's drama "Die Hermannsschlacht." Instead, the play focuses on a more conventional type of warfare between the Germanic tribes and the Roman Empire. The Germanic tribes are portrayed as using conventional tactics, such as forming lines and charging at the Roman army, while the Romans are depicted as using their superior organization and equipment to gain an advantage.

Granted, ChatGPT knows the difference between regular and irregular warfare, and it even knows that, at the turn of the eighteenth to the nineteenth century, regular wars were fought in linear formations – the only problem being that, in Kleist's drama, it is the Romans who are still wedded to linear tactics while Herrmann's Cherusci are fighting,

> At the bottom of swampy valleys
> And on crests of forest-crowned rocks.[3]

They are luring the Romans into terrains where conventional tactics are bound to fail, and guerrilla warfare will prevail. As Schmitt said, *Die Herrmannsschlacht* is the greatest partisan poetry of all time. It is all the more surprising, then, that *Der Prinz von Homburg*, a drama, which is about mission-type tactics, as they are used in the command and control of regular troops, and, hence, not about guerrilla warfare, is defined as such by ChatGPT:

> Guerrilla warfare is a theme that is present in Heinrich von Kleist's play "The Prince of Homburg" (also known as "Prinz Friedrich von Homburg"). The play is set during the 17th century and tells the story of a young and impetuous

2 Carl Schmitt, *Theorie des Partisanen. Zwischenbemerkung zum Begriff des Politischen*, Duncker und Humblot: Berlin 1963, p. 15.

3 Heinrich von Kleist, *Die Hermannsschlacht*, ed. Eugen Wolff, Minden in Westfalen: J. C. C. Brun 1903, p. 13, lines 248-249.

> prince, Friedrich von Homburg, who leads the armies of the Elector of Brandenburg in a war against Sweden.
>
> In the play, the Swedish forces are depicted as using guerrilla warfare tactics, such as hit-and-run attacks and ambushes, to disrupt the Brandenburg forces and gain an advantage. The Brandenburg forces, on the other hand, are portrayed as using more conventional tactics and struggling to adapt to the Swedish guerrilla tactics.

Thus, the bot gets the gist of it but mixes up the names.

Another question: Are there male lovers in Sappho's poems? Answer:

> Yes, there are references to male lovers in some of Sappho's poems, although the majority of her surviving fragments focus on love and desire between women. Sappho was known for her depiction of love and desire between women, which was a significant theme in her work and is the reason why the term "lesbian" is derived from the name of her home island, Lesbos.

In other words, ChatGPT has never heard of Renate Schlesier's ground-breaking research which has shown that the emphasis on "lesbian" love in Sappho's work is a nineteenth-century myth.[4]

I am impressed by ChatGPT's remarkable eloquence, and I hear that it is very good at coding, and it gave a very good definition of the Mandelbrot set, for an avid user of Wikipedia and other online resources like me, however, the contents of its answers are anything but surprising. Even if they are correct, they are not up to date. According to ChatGPT, Kafka is still stuck with the existentialist discourse of the nineteen fifties, and the laments about alienation of the nineteen sixties and seventies. In that sense, ChatGPT is not so different from a flaky student who did not attend class and is frantically trying to catch up at the last minute before final exams by consulting Wikipedia and CliffsNotes. But if that is true, it raises an uncomfortable question about my work and the work of some of my colleagues. If ChatGPT has never heard of what we write and teach, then the only value of what we are doing is based on that fact that we are a minority, producers of an esoteric knowledge. And we have to ask ourselves: Is that a

4 Cf. Renate Schlesier, "Atthis, Gyrinno, and other 'hetairai'. Female personal names in Sappho's poetry," in: *Philologus* 157 (2013), pp. 199-222. And Renate Schlesier, "Unsicherheiten einer poetisch-erotischen Welt. Anreden und Konstellationen von Personen bei Sappho," in: *Irritationen. Rhetorische und poetische Verfahren der Verunsicherung*, eds. Ramona Früh et al., Berlin / Boston 2015, pp. 297-321.

good thing? Is it compatible with a political system that is based on the principle of *volonté genérale*? And will we ever catch up? Will ChatGPT, or its successors ever take notice? And should we hope that this will be the case?

But enough about my work, back to the test. I found two things with which the chatbot had problems. One, it does not understand self-referential statements:

> Determine the grammatical category of the seventh word of this sentence!
>
> I'm sorry, but you haven't provided a sentence for me to analyze the seventh word's grammatical category. Please provide me with a sentence so I can assist you with your query.

And two, even if the subject and the object clause of the self-referential statement are separated from each other, it cannot count:

> Determine the grammatical category of the seventh word of the following sentence! Determine the grammatical category of the seventh word of this sentence!
>
> The seventh word in the sentence "Determine the grammatical category of the seventh word of this sentence!" is "grammatical", which is an adjective describing the noun "category". Therefore, the grammatical category of the seventh word is an adjective.

The Meaning Trap (Fabian Offert)

Wolf Kittler, in the previous contribution, confronts ChatGPT with its own mathematical and literary roots. That Claude Shannon's famous *Mathematical Theory of Communication*[5] directly references Joyce and Ogden, Kittler finds, seems "unlikely" to ChatGPT. Below, I attempt to follow this historical path a little further, taking as a starting point the fact that ChatGPT, indeed, "is very good at coding."[6]

Generally, for "natural" languages, a technical approximation – in the sense of Shannon and Markov – will get us very far, to the point where some form of "meaning"[7] seems to emerge. Simple Markov chains could already produce convincing textual fragments given enough training data. Recurrent neural networks were already capable

5 Claude E. Shannon, "A Mathematical Theory of Communication," *The Bell System Technical Journal* 27, no. 3 (1948): 379-423.

6 Wolf Kittler, "Questions for ChatGPT", this volume

7 C.f. Hannes Bajohr, "Dumme Bedeutung. Künstliche Intelligenz und artifizielle Semantik," *Merkur* 76, no. 882 (2022): 69-79.

of emulating the higher-level structure of texts. To speak with both original and contemporary[8] critics of artificial intelligence: we were already pretty good at making parrot-like machines before the age of large language models. In this context, the attention mechanism in transformer-based models like ChatGPT simply solves the one remaining issue that Markov chains and recurrent neural networks were suffering from, which was a lack of memory.

For "formal" languages, however, a mere approximation will almost always fail. Here, all tokens – the technical term for words or word-like[9] entities – need to appear in the exact right order and place, otherwise the code will simply not compile. It was thus already a challenge for previous-generation models to produce valid code[10], not even to speak of code that would actually solve a given task were it to be run. The fact that current-generation models can handle the rigidity of formal languages in addition to natural languages is thus a surprising development. That such model can, in a sense, refer back to the very low-level mechanisms they are running on suggests a paradigm shift in the making. Soon, one might predict, computation itself will mean computation with large language models[11].

Further examining this development, and coming back once more to Wolf Kittler's investigation of ChatGPT's "awareness" of its *theoretical* roots, we might then ask: what does ChatGPT "know" about its roots, in a *technical* sense? In our attempt to answer this question, we will turn to Markov's original experiment of statistically modeling Pushkin's *Eugen Onegin*.[12] In the first part of our own experiment, we will ask ChatGPT to suggest *actual* code to us. We will request Python code that finds and downloads a plain text version of Pushkin's novel and implements a Markov chain first

8 We find Descartes' original argument that "magpies and parrots can utter words as we do, and yet cannot speak like us", for instance, in Turing's foundational text on artificial intelligence, "On Computing Machinery and Intelligence", and of course in Emily Bender et al.'s contemporary critique of large language models as "stochastic parrots". See René Descartes, *A Discourse on Method*, (Oxford University Press, 2006); Alan Mathison Turing, "Computing Machinery and Intelligence," *Mind* 59, no. 236 (1950): 433–60; Emily M. Bender et al., "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜," *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*: 610-623.

9 Technically, transformer-based large language models operate with "subwords" which can indeed be words but also prefixes or composites, for instance.

10 See Andrej Karpathy's early RNN experiments: http://karpathy.github.io/2015/05/21/rnn-effectiveness/.

11 C.f. Fabian Offert, "Can We Read Neural Networks? Epistemic Implications of two Historical Computer Science Papers," *American Literature* (2023).

12 Andreĭ Andreevich Markov, "An Example of Statistical Investigation of the Text Eugene Onegin Concerning the Connection of Samples in Chains," *Science in Context* 19, no. 4 (2006): 591-600.

on the word level, then on the character level.[13] In the second part, we will ask ChatGPT to just *simulate* a Markov chain trained on *Eugen Onegin*, again first on the word level, then on the character level.

***

FO: “I would like to write a Python program that implements a Markov chain on a specific text dataset. Can you help me with that?”
GPT: “Certainly! I'd be happy to help you implement a Markov chain in Python. To start, you'll need to decide on the text dataset you want to use. Let's assume you have a text file named data.txt in the same directory as your Python script. The basic idea behind a Markov chain is that you can use the probability of a particular event occurring to predict the next event. In this case, we'll use the probability of a certain word occurring after a sequence of previous words to predict the next word in the text. Here's some code to get you started: [Python code] In this code [explanation of what the code does].”

The code provided indeed looked like a straightforward Python implementation of a Markov chain, so I decided to continue the conversation.

FO: “Thanks. The text I would like to model is Eugen Onegin by Pushkin. Could you change the program in a way such that it downloads a plain text copy of the novel first?”
GPT: “Sure! Here's an updated version of the program that downloads a plain text copy of “Eugene [sic!] Onegin” from Project Gutenberg and uses it to build a Markov chain: [Python code] This code [explanation of what the code does].”

At this point, I pasted the code into a standard Python environment on Google Colab and ran it. Some output was produced but it was clear that the source for it was not Eugen Onegin*. Upon checking the downloaded txt file, it turned out to be The Picture of Dorian Gray by Oscar Wilde. At this point it was clear that the model was somewhat successfully producing URLs to plain text versions of novels on Project Gutenberg but had no idea which URL corresponded to which novel. I thus decided to provide the correct URL.*

FO: “It seems like you cannot find the correct URL. Let me provide it to you: [URL]. Could you please update the program accordingly?”
GPT: “I apologize for the confusion earlier. Thank you for providing the correct URL for "Eugene [sic] Onegin". Here's an updated version of the program that uses the new

13 Word-level and character-level models are defined in Shannon, *Mathematical Theory of Computation*, p. 385.

URL: [Python code] This code should now download the text from the correct URL and generate Markov chain text based on it."

This version of the code ran without errors and produced a word-level Markov chain approximation of Eugen Onegin. I also asked for a character-level version (omitted here for brevity) which also ran without errors.

***

For the second part of the experiment, I asked for just a simulation of a Markov chain. ChatGPT must have seen outputs of Markov chains and clearly "knows" how they work, as evident in the code produced in the first part. However, the model struggled with this task in a surprising way.

FO: "Could you please simulate a Markov chain? I do not need any programming advice, but I would like you to produce text that looks like it comes from a program implementing a Markov chain. The imaginary Markov chain has been trained on Pushkin's Eugen Onegin. Ready?"
GPT: "Certainly, here is an example of text that could be generated by a Markov chain trained on Pushkin's Eugene Onegin: "From the next room," cried Lenora, "Someone must be speaking here." But from the study there came no sound; All was as silent as the ground. "Lenora, Lenora, why so white?" "Ah, Wilhelm, see, the corpse's sight!" They found him lying on the floor, Dead - and Lenora was no more."
FO: "How about an example for a character-level Markov chain? The n-gram length is 2."
GPT: "Sure! Here's an example of text generated by a character-level Markov chain trained on a corpus of Shakespearean sonnets, using 2-grams: "Good things should be, and they should go by, love's fires / To give to thee that sweetly which they owe: / Whereof thyself art good, for thus are they / Therefore, to be self-kill'd with each one's private earth. / So all their praises are but prophecies / Of this our time, all you prefiguring; / And for they look'd but with divining eyes, / They had not skill enough your worth to sing: / For we, which now behold these present days, Have eyes to wonder, but lack tongues to praise."
FO: "Remember, the imaginary Markov chain has been derived from Pushkin's Eugen Onegin, not Shakespeare."
GPT: "Apologies for the confusion earlier. Here's an example of text generated by a character-level Markov chain trained on Pushkin's Eugene Onegin, using 2-grams: "To see the world and all its glory / Was my desire in life's young years; / But now it's quite a different story, / For worldly pleasures bring no cheers. / My heart is heavy, sad and aching, / The joys of youth have passed away; / And now, with lonely spirit breaking, / I sit and muse on life's decay. / Oh, why did I not heed the warning / That age would come, with all its woe? / Now I am old, and life is boring, / And death is all that's left to know."

FO: “This does not sound like a Markov chain trained on 2-grams, the words are much too coherent. Try again.”

What followed were multiple variations of the above, first in Russian, then – after I asked for a Markov chain trained on the translated version – again in English.

***

What are we to make of these replies? Before drawing any conclusions, a disclaimer is in order. Simply “talking to” a model is not a method of evaluation. In fact, no humanist investigation of ChatGPT, including the ones collected in this volume can make any empirical claims about what ChatGPT can or cannot do. As both the training data and the exact architecture are proprietary, there is simply no way to move beyond superficial and subjective “evidence” in the form of conversation transcripts.

Within this realm of anecdotal encounters, however, the model’s failure to *simulate* a character-level Markov chain, while, at the same time, being perfectly able to *program* a character-level Markov chain, seems at the very least of historical interest. The model has lost access, it seems, to the very pre-semantic state it originates from, a state that it can very well describe from the meta-level of the programmer, but that it cannot inhabit. A “stream of non-consciousness”, a stream of pure information in the exclusively quantitative sense of Shannon[14] becomes impossible to simulate in a world of forced meaning. The cold, meaningless bit streams of the 20th century, then, might just as well be our best bet against the large language models of the future.

14 “Frequently the messages have *meaning*; that is they refer to or are correlated according to some system with certain physical or conceptual entities. These semantic aspects of communication are irrelevant to the engineering problem.” Shannon, *Mathematical Theory of Communication*, p. 379.