CAN YOU TRUST A TEXT GENERATED BY ARTIFICIAL INTELLIGENCE?

Charles Ollion[1] and Anna Bonnet[2]

[1]Surfrider Foundation Europe, Biarritz, France
[2]Laboratoire de Probabilités, Statistique et Modélisation (LPSM), Sorbonne Université, Paris, France

ChatGPT and other artificial intelligence (AI) applications seem to understand everything and give convincing answers, even for tough exam questions. But can we always trust these answers? In this article, we'll take a closer look at these AI systems: how they're trained, what makes them smart, and how we can chat with them. We will see that they produce answers that seem truthful, but they may not always care providing us with factual information.

INTRODUCTION

When you use an internet search engine (Google or Bing for example), the results you get are web pages or answers that have been created beforehand, for example a Wikipedia page, usually written by humans. When you ask a question to platforms like ChatGPT, the process is very different: the result is generated on the fly by an artificial intelligence (💻) engine. Try it yourself: if you ask the same question several times (📝), the answers may be different. Here's a real-life example of two ChatGPT answers to my question:

📝 What is Science? Give a short and funny answer.

💻 Science is essentially asking a lot of questions to nature and trying not to interrupt it when it gives really long and complicated answers!

💻 Science is like trying to solve a giant cosmic puzzle, but the pieces are sometimes invisible and the instructions are written in the form of riddles!

ARTIFICIAL INTELLIGENCE. The ability of a computer program to reason and learn. Today, we call ChatGPT "artificial intelligence", even though experts and scientists don't consider it to be truly "smart".

ChatGPT is a web platform launched in late 2022 by a company called OpenAI. It has attracted a lot of media attention as well as great interest in the scientific world. Since then, many other applications have been developed (Bard, Ernie, LLaMa), using the same fundamental principles. By the way, GPT stands for Generative Pre-Trained Transformer, but what does this mean? Let's use the definition of each of these words to understand how ChatGPT functions. Then, we will have the tools to answer the question at hand: can we trust AI engines?

THE GENERATIVE ENGINE: THE LANGUAGE MODEL

The AI engine uses what we call a language model [1] to understand the data entered by the user and generate text as an output. The way it works is as follows: the engine creates a sentence word by word, predicting which new word will best fit with the previous ones. Consider an example of sentence generation (Figure 1). The input sentence is divided into words (left), and the role of the engine is to predict which word, from a predefined vocabulary of all possible words (right), would fit best. It does this by assigning a score to every possible word: most words will have a score of 0, which means they are unlikely to be chosen, and a few will score higher (represented by blue lines).

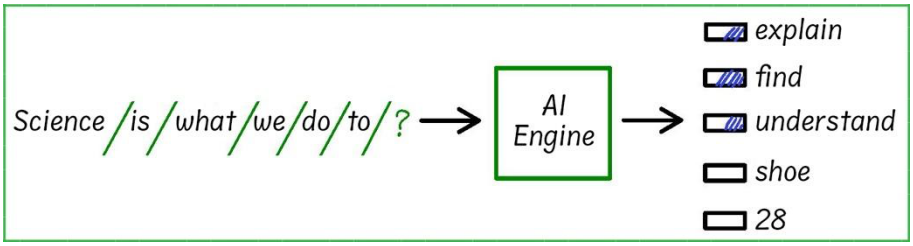

Figure 1. The AI engine seeks to predict the next word.

In the example above, the word "find" will have the highest score, followed by "understand" and "explain" while a word like "shoe" will have a score of 0. Then, the engine chooses a word at random from those with the highest scores: in this case, the word "find" is more likely to be chosen, but other words with a score other than 0 can also be chosen. This word is added to the sentence, and we can repeat the whole process to find the next one, until a complete answer is produced!

When a question is fed to ChatGPT, or when we chat with it for hours, the AI engine always follows the principle of the Language model: try to predict the next word. Some words are written by you, others generated previously by the engine, so that at any given moment ChatGPT can continue generating the most likely words. This gives the most plausible phrases or answers, given the entire context of the conversation.

HOW TO (PRE-)TRAIN THE SEARCH ENGINE?

In order for the engine to be able to predict the next right words, it must be trained beforehand. This "pre-training" is carried out by engineers and researchers over several months. The engine is then frozen and used. This means that ChatGPT and other AI systems are no longer learning anything when you talk to them!

But what exactly does it mean to train an AI engine? In the beginning, before training, the engine is not very smart. At this stage, it usually produces the same low scores for each word.

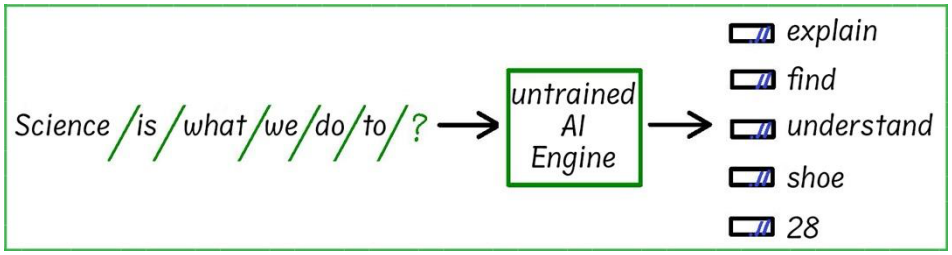

Figure 2. A beginner (untrained) AI engine is not able to predict the next word correctly.

If you ask a beginner AI engine a question (Figure 2), it will produce a random sequence of words that will make no sense.

Engineers and researchers have created algorithms to train these motors by giving the beginner AI lots of existing texts [2]. The algorithm is a computer program that looks like this:

ALGORITHM. A set of instructions to follow to complete a task. For example, a recipe can be an algorithm for making a cake. Computer algorithms are created by programing using a code.

1. Choose a sentence to show the untrained engine, for example, this one: "Science is a way of finding explanations to understand how the world works as well as its phenomena."

2. Ask the engine to score the words that follow this part of the sentence: "Science is a way to".

3. The lower the score of the right word - in this case "find" - the greater the engine error. At first, when the engine is not trained, this error is significant. The algorithm uses this error to tweak the engine slightly, so the next time the engine is asked for scores with this type of input, it will give a better score to "find", and therefore make fewer mistakes.

When we repeat this process many times, with billions of different sentences, the engine learns and starts gradually to make fewer and fewer mistakes and therefore produces higher scores for probable words. It takes many fast computers and weeks of training to train the large motors that power ChatGPT.

Depending on the texts chosen to show the engine during training, it will assign different scores to the words when they are used afterwards: for example, if we select mainly training texts in French,

it will give very low marks to the English words, but high scores to some of the French words

In reality though, the training process is quite a bit more complex, and to fully train an AI engine, it takes several steps.

WHY IS THE TRANSFORMER SO EFFICIENT?

Why is ChatGPT so successful? It's a mystery, and scientists are still trying to explain it, but there are several elements in the design of these AI engines that are very important, so let's look at a few.

ARTIFICIAL NEURAL NETWORK. A type of computer program that is loosely inspired by the way the brains of animals work. It is used to solve complex problems

One of the secret weapons is the way the engine is designed: it uses an artificial neural network architecture called a transformer [3]. A transformer is an efficient computer that receives the input sentence in order to produce the scores. It has a special feature that allows it to focus on certain parts of the sentence that are most important to produce the best possible results. For example, in order to produce a good score for "find" in the example above, the transformer will likely focus on the word "of," which suggests that the next word is surely a verb, as well as the word "science" which indicates the overall meaning of the sentence. It's like a detective's magnifying glass that helps the engine read between the lines and identify what is important. This is very useful, especially when you give the engine a very long text as input!

One of the main reasons it works so well is the scale of the training data: it's hard to imagine how many sentences these motors are trained on. GPT-3 (the third version of the language model that powers ChatGPT) was trained on about 500 billion words, which is about 5 million different books. Another important reason is the diversity of texts ChatGPT was trained on Not only were the texts written in many different languages, they also included programming languages such as Python, as well as instructions and recipes.

The way the vocabulary is chosen is also very important. Remember that the AI engine's job is to assign a score to each of the possible words. ChatGPT seems to understand every possible English word, as well as misspelled words like "helo", or any word from another language, such as Japanese words like ひらがな, which makes billions of possibilities! Among its vocabulary, there are also punctuation marks, which contribute to the coherent structure of the text, or line breaks, which allow ChatGPT to write a poem in verse! To keep the vocabulary relatively small (about 50000 possible words), the trick is to use not only words, but also parts of words, called tokens, and letters or characters, to allow ChatGPT to produce any word by combining them (Figure 3).

For example, ChatGPT might separate the word "converted" into two tokens: "convert" and "ed". The first contains the root of the

word, common with words of the same family such as “convert” or “converted” while the ending defines the nature of the word, here a past participle.

```
Simple words are just converted to a single token, but a name such as
 ChatGPT is broken down into a few.
```

Figure 3. This example shows how ChatGPT breaks down text into tokens. Each color change indicates where ChatGPT separates the text. You can play cutting a text into tokens at the following address https://platform.openai.com/tokenizer

This explains why ChatGPT is able to provide amusing explanations for made-up words:

> Give a plausible definition for this invented word: superheropathy
>
> The term "superheropathy" could refer to an imaginary disease whose main symptom is pretending to have superpowers, like comic book heroes. Patients act as if they can fly, have superhuman strength, or do magical things, even if they can't really do it. They can wear costumes and do impersonations by imagining that they are superheroes. All this for fun and imagination!

CAN AI ENGINES BE TRUSTED?

Since AI engines are able to understand any data and their answers are impressive, we tend to believe that they are "smart" and trust them. But trusting them is a big mistake, because they produce answers that are merely plausible, but can often be wrong. Let's use math as an example:

> 23 * 7 - 109
>
> The result of the expression 23 * 7 - 109 is 2.

ChatGPT confidently provides a wrong answer of 2, the correct answer being 52. More importantly, when asked the question again, it gives a different incorrect answer.

> 23 * 7 - 109
>
> The result of the expression 23 * 7 - 109 is 42.

To understand what's going on, it's essential to remember how ChatGPT works: the AI engine doesn’t follow the rules of arithmetic to respond. All it does is identify that a small number was a likely answer rather than a higher number or a token that isn't a number.

Knowing this, it can still be useful to interact with an AI engine while knowing its weaknesses. By understanding how it works, it is even possible to make it more useful. For example, instead of asking for a numerical response directly, it's best to do the following:

✍ How much is 23 * 7 – 109 worth? Can you describe the different steps of calculation before giving the answer?

💻 Of course, I would be happy to help you do the different steps of this calculation:
Step 1: Multiply 23 by 7: 23 * 7 = 161
Step 2: Subtract 109 from the result of the multiplication: 161 - 109 = 52
So the answer is 52.

We forced the engine to perform an intermediate reasoning: think step by step to solve the problem! Note that this doesn't stop the engine from making mistakes, but it typically helps provide a better response. This way of thinking for AI engines is called a chain of thought [4], and can also be used to provide interesting examples during training. In the latest version of ChatGPT, it tends to detail its calculations by default, precisely to avoid making mistakes. On the other hand, it is always possible to "trap" the AI engine by asking it to answer with a number only, which often leads to it making mistakes…

CONCLUSION

AI engines such as ChatGPT provide impressive results, but they are not trained to tell the truth, so they should never be trusted blindly. However, they are useful because they can be used in applications and will continue to be used to create personal assistants or improve web searches. Though they are practical, their use raises many ethical questions: for example, AI engines can significantly change the way we write a text, without giving credit to the authors of the data on which they have been trained. Also, if we start using these engines too much, it would be difficult to know which texts or conversations are written by humans and which are artificial. Another aspect to consider is that ChatGPT can write stories, essays, and other things for us, but if we rely on it to accomplish all our tasks, we risk not learning, not being creative, not developing, and not thinking for ourselves. Even though scientists know a lot about how ChatGPT works, they still don't know why it works so well! This means we can use it while understanding its strengths and limitations, but we should never take ChatGPT's answers at face value and we should always question them.

REFERENCES

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Kukasz Kaiser, Illia Polosukhin. "Attention is all you need." Advances in neural information processing systems 30 (NIPS 2017).

[2] David E. Rumelhart, Geoffrey E. Hinton, Ronald J. Williams. Learning representations by back-propagating errors. Nature 323.6088 (1986): 533-536.

[3] Yoshua Bengio, Réjean Ducharme, Pascal Vincent. "A neural probabilistic language model. Advances in neural information processing systems 13 (NIPS 2000).

[4] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, DennyZhou. Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems 35 (2022): 24824-24837.

ORIGINAL ARTICLE (ENGLISH VERSION)

SUBMITTED March 1, 2024; **ACCEPTED** November 2024
PUBLISHED: January 2025

EDITORS: Catherine Braun-Breton & Ula Hibner, Association Jeunes Francophones et la Science

SCIENTIFIC MENTORS: Catherine Braun-Breton & Ula Hibner

CITATION: Ollion C., Bonnet A. (2025). Can we trust a text generated by artificial intelligence? Jeunes Francophones et la Science (https://www.jeunesfrancophonesetlascience.fr/).

CONFLICT OF INTEREST: The authors state that the research was conducted in the absence of any commercial or financial relationship that could be construed as a potential conflict of interest.

YOUNG REVIEWER

CHARLIE, AGE: 16

Charlie is French-American and lives in Virginia in the United States. She loves dance, science, piano and cello.

AUTHORS

CHARLES OLLION

Charles does research in artificial intelligence (AI) and its applications. He holds a PhD in Computer Science and has taught AI for Machine Vision and Language Understanding at the Institut Polytechnique de Paris. He currently works for the NGO Surfrider Foundation Europe where he uses AI to detect plastic waste. If he's not behind his computer, he's probably climbing the nearest mountain!

ANNA BONNET

Anna is a lecturer and researcher in statistics at Sorbonne University, specializing in mathematical modeling for the life sciences. She collaborates with ecologists, doctors and biologists to develop and apply new data analysis tools.