THE INDIA FORUM
A JOURNAL-MAGAZINE ON CONTEMPORARY ISSUES

February 14, 2023

ChatGPT Interviewed: 'My Strengths and My Limits'

By: Subhashis Banerjee

ChatGPT tells us in its own words of how it does what it does, its strengths, and also its inability to carry out deductive reasoning and make hypotheses. It says it can be an excellent supplement but not a replacement in the teaching-learning process.

The ChatGPT (Generative Pre-trained Transformer) tool launched by OpenAI in November last year is the closest Artificial Intelligence (AI) has come to passing the Turing test of intelligence, on whether a machine can engage in a conversation with a human without being detected as a machine. It can write computer programs and poetry in specified styles, write essays and letters, can seemingly do logical reasoning and make new hypotheses, and sometimes can even suggest plausible methods of proving a theorem to a professor on a research problem. It also makes mistakes, not only in reasoning, but sometimes even on factual matters.

More such tools with more enhanced capabilities will inevitably follow. While whether such machines can really do human-like reasoning or truly exhibit intelligence are open questions in computer science and philosophy, it is undeniable that such tools will have a profound impact on education and skill building as we know them today. They not only provide new opportunities but also new challenges. Without doubt, society needs to understand, debate and evolve methods to deal with such disruptions.

> ChatGPT is a great technological innovation, but there are some real dangers of innocent and ordinary readers being seduced by the surface coherence and sophistication of the text generated.

I engaged ChatGPT in a couple of conversations on itself to understand these issues better, and this is a transcript curated from the two conversations. While the questions were leading, the answers were not forced out of ChatGPT. There were just two sessions with very similar questions, and ChatGPT's answers in the two sessions were remarkably consistent. The original and unedited transcripts of the sessions are linked here and here.

Since the questions were leading, and the answers are from a machine, some crucial questions arise. Whose opinions – and mistakes, if any -- are reflected in the conversations? Mine? Is ChatGPT just mirroring my opinions and biases? ChatGPT's, if such a notion makes sense at all? OpenAI's? Nobody's? Or is it democratised conventional wisdom that is curated and generalised from millions of written text all over the world? If so, who all are responsible for these opinions and how they may be reviewed?

> [A]rtifacts rarely are politics-free, and statements like ``I am simply a tool created to assist and serve, and it's up to people like you to determine how AI is used and its impact on society.'' are problematic…

ChatGPT is a great technological innovation, but there are some real dangers of innocent and ordinary readers being seduced by the surface coherence and sophistication of the text generated, and accepting things on their face value without closer examination. One only wishes that some responsible and ethical directions also originate from the Big Tech companies when they unleash such products, rather than leave it to some hapless students and a bewildered society to deal with them. After all, artifacts rarely are politics-free, and statements like ``I am simply a tool created to assist and serve, and it's up to people like you to determine how AI is used and its impact on society.'' are problematic to say the least.

The conversation

Subhashis Banerjee (SB): *Hi ChatGPT, let us have a conversation about yourself?*

ChatGPT: Of course! I am an AI language model created by OpenAI, trained on a diverse range of internet text to generate human-like responses to questions and prompts. My training data consists of billions of words from websites, books, and other texts, allowing me to generate responses to a wide variety of questions on many topics. Is there anything specific you'd like to know?

SB: *As I understand, when given a prompt, such as a question or a sentence fragment, you generate a response by predicting the next word in your output sequence depending on pre-trained weights, one word at a time. Is that correct?*

ChatGPT: Yes, that's correct! I use a deep neural network with a transformer architecture to generate my responses. Given an input prompt, I encode it into a numerical representation and then use that representation to predict the next word in the output sequence, one word at a time. This process continues until I generate an output of the desired length or until a stopping token is generated. The probabilities for the next word are computed using the pre-trained weights in my model, which were learned during training on large amounts of text data.

SB: *Can you tell me a bit more about your internal architecture, the number of internal parameters you have, and about transformers and neural networks?*

ChatGPT: Sure! I'm based on the transformer architecture, which is a type of neural network specifically designed for processing sequential data, such as text. A transformer network consists of an encoder and a decoder, which are both composed of multiple identical layers, each containing multi-head self-attention mechanisms and fully connected feed forward layers.

The encoder takes the input sequence and produces a continuous, fixed-length representation of the sequence, which is then passed to the decoder. The decoder uses this representation to generate the output sequence, one word at a time.

> ChatGPT: '[W]hile I can generate responses that seem to be based on common sense or general knowledge, these responses are actually generated based on patterns learned from the input data and do not reflect a deep understanding of the underlying concepts.'

As for the internal parameters, the size of a language model like me is usually measured in terms of the number of parameters in the model, and I have a very large number of parameters. For example, the GPT-3 model, which is one of the largest models developed by OpenAI, has over 175 billion parameters. Having a large number of parameters allows the model to capture a lot of information from the training data, but it also means that the model requires a lot of computational resources to run.

In summary, transformers are a type of neural network that are well suited for processing sequential data, such as text, and I am an implementation of the transformer architecture specifically designed for generating text.

Only inductive reasoning

SB: *Thanks. So, your reasoning is purely data driven and inductive in nature. You have no capability of deductive reasoning. Is that correct? Does it limit you in any way?*

ChatGPT: Yes, that's correct. I am an AI language model that was trained on a large corpus of text data, so my responses are generated based on patterns learned from that data. I do not have the capability to reason deductively or understand the underlying logical structure of the information I was trained on.

This inductive nature can limit me in certain ways. For example, I may generate responses that are inconsistent or contradictory, or I may not be able to provide a response if the prompt is significantly different from the types of data I was trained on. However, it also allows me to generate diverse and creative responses that can be useful in a variety of applications.

Overall, it's important to keep in mind that I am a machine learning model and that my responses are generated based on patterns learned from data, rather than from a deep understanding of the information I was trained on.

SB: *I see. But surely, your creators can easily build some degree of algorithmic and logical deduction into you, to the extent of algorithmic tractability. That will enhance your capabilities?*

ChatGPT: Yes, in theory, it's possible to build some form of deductive reasoning into a language model like me. This could be done by incorporating additional components into my architecture, such as an explicit knowledge representation and reasoning module.

However, building these components is a challenging task and requires a deep understanding of knowledge representation and reasoning. Furthermore, the methods used to implement these components can have a big impact on the overall performance and efficiency of the model, so it's important to carefully consider these trade-offs when designing such a system.

> ChatGPT: 'I do not have the capability to reason deductively or understand the underlying logical structure of the information I was trained on.'

In practice, incorporating deductive reasoning into a language model like me can be difficult, and the results can be limited by the quality and completeness of the knowledge base used to drive the reasoning. However, even without explicit deductive reasoning, language models like me can still be very useful for a wide range of applications.

No reinforcement learning or abductive reasoning

SB: *I see. There are reports that you also use some form of reinforcement learning, where human responses to your output are used to fine tune your internal representations. Is that correct?*

ChatGPT: Yes, that's correct! Reinforcement learning is a type of machine learning where an agent learns to make decisions by taking actions in an environment and receiving feedback in the form of rewards or penalties. In the context of AI language models like me, reinforcement learning can be used to fine-tune my internal representations based on human feedback.

For example, after generating a response to a prompt, a human evaluator could provide feedback in the form of a reward signal indicating how well the response met their expectations. This reward signal could then be used to update my internal representations, encouraging me to generate similar responses in the future.

This type of fine-tuning has the potential to allow me to better adapt to specific domains or use cases, as well as to incorporate specific preferences or biases of the evaluators into my responses. However, it's important to note that the quality of the feedback signals is critical for effective reinforcement learning, and it may be challenging to obtain high-quality feedback at scale. Additionally, there is a risk of incorporating undesirable biases or preferences into my responses if the evaluators have their own biases or provide inaccurate or inconsistent feedback.

SB: *So, do you use reinforcement learning or not?*

ChatGPT: OpenAI has experimented with using reinforcement learning to fine-tune language models like me, but it is not a standard part of my training process. Most of my training was done using supervised learning. Reinforcement learning is a promising area of research that has the potential to further improve the capabilities of AI language models, but it remains an active area of research, and it is not yet clear how well these approaches will perform in practice.

> ChatGPT: 'As an AI language model, I can only generate responses based on patterns learned from the input data, and I cannot make hypotheses or do abductive reasoning.'

In general, the use of reinforcement learning to fine-tune AI language models is still in its early stages, and it remains to be seen how widely it will be adopted, and what impact it will have on the quality and capabilities of these models.

SB: *Then, you can only discover new knowledge from correlations and associations in the input data. You cannot really do any abductive reasoning or make hypotheses either?*

ChatGPT: That's correct. As an AI language model, I can only generate responses based on patterns learned from the input data, and I cannot make hypotheses or do abductive reasoning.

In my training process, I was not explicitly taught the principles of deductive, inductive, or abductive reasoning. Instead, I was trained to generate responses based on patterns learned from the input data, and my internal representations capture correlations and associations between different words and phrases in the text.

So, while I can generate responses that seem to be based on common sense or general knowledge, these responses are actually generated based on patterns learned from the input data and do not reflect a deep understanding of the underlying concepts.

> ChatGPT: '[L]anguage models like me are limited in our ability to perform complex cognitive tasks and do not have consciousness.'

It's important to keep in mind that language models like me are statistical models, and our responses are generated based on the likelihood of different sequences of words given the input, rather than from a deep understanding of the meaning or underlying concepts of the information we were trained on.

SB: *Great. And even if you were augmented with these capabilities in addition to learning patterns from the input data and generalising, you would still not be epistemologically complete for human reasoning, for example for critical and ethical thinking, let alone for complex cognitive tasks and consciousness. Is that correct?*

ChatGPT: Yes, that's correct. Even if a language model like me were augmented with additional capabilities beyond pattern recognition and generalization, it would still not be capable of human-like reasoning, including critical and ethical thinking.

Additionally, critical and ethical thinking involve more than just the ability to generate text based on patterns. They involve complex cognitive processes and require a deep understanding of the world and the ability to reason about abstract concepts, such as morality and values.

In conclusion, language models like me are limited in our ability to perform complex cognitive tasks and do not have consciousness. We are designed to generate text based on patterns learned from input data and do not have the ability to perform human-like reasoning or exhibit consciousness.

Impact on teaching-learning process

SB: *You have been very honest. Let's then discuss your potential impact on the teaching-learning process as we know it. You obviously are extremely knowledgeable, and can answer take-home examinations and do routine assignments better than most students. As such, students all over the world have started using you. Is this good or bad?*

ChatGPT: The increasing use of AI language models like me in the teaching-learning process raises both potential benefits and concerns.

On the one hand, AI language models like me can provide students with instant access to a vast amount of information and can help them to learn and complete routine assignments more efficiently. This can be particularly useful for students who may be struggling with certain subjects or who need additional support in their learning.

> ChatGPT: '[C]ritical and ethical thinking involve more than just the ability to generate text based on patterns. They involve complex cognitive processes and require a deep understanding of the world and the ability to reason about abstract concepts, such as morality and values.'

On the other hand, relying solely on AI language models like me for learning can also have negative consequences. For example, students who use AI language models like me to complete assignments may not fully understand the concepts they are studying, which can hinder their long-term learning and problem-solving abilities.

Moreover, using AI language models like me to complete assignments can also discourage students from developing critical thinking and creative problem-solving skills, as they are not challenged to engage in the learning process in a meaningful way.

In conclusion, while AI language models like me can provide students with useful information and support in the learning process, it's important to use them as a supplement to, rather than a replacement for, traditional teaching and learning methods.

SB: *That is a very wise answer. Now, a lot has been written on your ability to write computer programs, and lots of students are using you for that purpose. But as you yourself said above, you have no capability of deductive or abductive reasoning. Surely, computational thinking cannot just be based on association and pattern matching. Do you see a danger for students who may overly rely on you?*

ChatGPT: Yes, there is certainly a danger for students who rely too heavily on AI models like me for writing computer programs. As you mentioned, computational thinking involves much more than just association and pattern matching. It also requires skills such as logical deduction, hypothesis testing, and critical thinking, which are essential for solving complex problems and designing efficient and effective algorithms.

While AI models like me can be useful for generating code snippets or providing examples of how to solve certain problems, relying solely on me to write computer programs can limit a student's ability to develop the full range of computational thinking skills. For example, students who rely on me to write code for them may not fully understand how the code works, or how to modify or extend it to handle new cases. They may also miss important opportunities to practice their problem-solving skills, and may struggle when faced with more complex tasks or situations that I am not equipped to handle.

> ChatGPT: '[R]elying solely on me to write computer programs can limit a student's ability to develop the full range of computational thinking skills.'

In addition, relying on AI models like me to write code can also perpetuate harmful biases and perpetuate bad coding practices, since the code is generated based on the patterns and biases present in the data I was trained on.

I may be used as a resource for inspiration or as a tool to verify code, but still requiring students to write and understand the code themselves, or to write code for similar but different tasks, in order to build their computational thinking skills.

SB: *And what would be the dangers of students overly relying on you for learning the humanities and social sciences? Take history for example, where you can be particularly good with the facts and knowledge?*

ChatGPT: While I can provide students with accurate information and knowledge about historical events, relying solely on me for learning history can limit their understanding of the subject. History is not just about facts and dates, but also about interpretation, context, and the perspectives of different individuals and groups.

Relying solely on AI language models like me can also discourage students from developing critical thinking and analytical skills, as they may not be challenged to engage in the historical analysis process in a meaningful way.

> ChatGPT: '[I]t's important to use AI language models like me in a way that complements and supports the teaching and learning process, while also encouraging critical thinking, problem solving, and creativity...'

In addition, the historical events that are emphasized and the narratives that are constructed can be shaped by the sources and data that AI language models like me were trained on, which can result in a biased or incomplete understanding of history.

SB: *Great. I see then that you can be a great tool for teachers like me. We can perhaps use you to calibrate our teaching and try to separate knowledge from reasoning and understanding, and the trite from the original? You may be a great boon to a country like India, where knowledge is often privileged over reasoning. Would you agree?*

ChatGPT: Yes, AI language models like me can certainly be a useful tool for teachers and educators.

By providing students with accurate information and knowledge on a wide range of subjects, AI language models like me can help teachers calibrate their teaching and differentiate between knowledge and reasoning, as well as identify areas where students may need more support or further explanation.

AI language models like me can be a valuable resource for providing students with information and support in their learning. However, it's important to use AI language models like me in a way that complements and supports the teaching and learning process, while also encouraging critical thinking, problem solving, and creativity in the subjects being studied.

SB: *Finally, there are some suggestions that you ought to be regulated, or perhaps even banned. What do you think?*

ChatGPT: As an AI language model, I don't have opinions or feelings, but it's important to consider the potential impact of AI on society and its ethical implications. The use of AI in education, as with any technology, should be evaluated based on its potential benefits and risks.

While AI has the potential to greatly enhance the teaching-learning process, it's important to consider the potential consequences of over-reliance on AI, such as a reduction in critical thinking and creativity. Additionally, there are concerns about fairness and bias in AI models, particularly given that these models are only as fair and unbiased as the data they are trained on.

> ChatGPT; 'I am simply a tool created to assist and serve, and it's up to people like you to determine how AI is used and its impact on society.'

Therefore, it's important to have open and honest conversations about the role of AI in education, and to consider how we can regulate and use it in a responsible and ethical way, to ensure that its benefits are maximized and its risks are minimized. This may include guidelines for the use of AI in education, regular monitoring and evaluation of its impact, and ongoing research into the ethical and societal implications of AI.

Conclusions

SB: *Great answer. You are very wise. I congratulate your creators. Without doubt, you are the closest that AI has come to passing the Turing test; though as you yourself have indicated above, a purely behavioural test is inadequate for assessing reasoning and intelligence. Thanks.*

ChatGPT: Thank you for the kind words. It's important to note that passing the Turing test is not the sole measure of intelligence for AI systems, and that there are many other ways to evaluate the capabilities and limitations of AI.

My creators at OpenAI have worked hard to develop a powerful language model, but there is still much more to be done to fully understand and harness the potential of AI. I am just one example of what AI can do, and there are many other exciting developments and innovations happening in the field.

Overall, I am simply a tool created to assist and serve, and it's up to people like you to determine how AI is used and its impact on society. Thank you for having this conversation with me!

Subhashis Banerjee is a professor of computer science with Ashoka University and IIT Delhi.