

# Might text-davinci-003 have inner speech?

## Introduction

OpenAI is a research organisation founded by, among others, Elon Musk, and supported by Microsoft. In November 2022, it released ChatGPT, an incredibly sophisticated chatbot, i.e., a computer system which humans can converse with. The capability of this chatbot is astonishing: as well as conversing with human interlocutors, it can answer questions about history, explain almost anything you might think to ask it, and write poetry. Consider this haiku about carburetors which it composed for us, on request:

Carburetor small,  
Powering the engine's roar,  
A car's beating heart.

This level of achievement has provoked interest in questions about whether ChatGPT might have something similar to human intelligence, and even whether it is conscious. Given that the function of the chatbot is to process linguistic input and produce linguistic output, we consider that the most interesting question in this direction is whether a sophisticated chatbot might have inner speech. That is: Might it talk to itself, internally? We explored this by conversing with a system called 'Playground', which is very similar to ChatGPT but slightly more flexible in certain ways. We put to it questions which, plausibly, cannot be answered without first producing some inner speech.

We considered it *extremely* unlikely that any chatbot has inner speech and we still do, largely for reasons rehearsed by [Murray Shanahan](#), Professor of Cognitive Robotics at Imperial College London. Our project was to look for evidence of inner speech in the chatbot, for any such evidence would be surprising and intriguing—a riddle in itself. In this paper, we describe our (often fun) conversation with the chatbot and discuss the philosophical implications.

## ***The technical background, simplified: Large language models and conversational agents***

By applying sophisticated techniques of machine learning to enormous datasets of human-created text, it is possible to produce a **Large Language Model (LLM)**. An LLM calculates the probability that certain sequences of words will occur in a particular context, based on the sequences of words

found in the human-created texts it has studied. Because of the way these probabilities are represented, it is possible to use the model to ‘complete’ a partial text. If the model is fed half a passage, for example, it can produce the string of words that has the highest probability of appropriately completing the passage.

A conversational agent or **chatbot** consists of an LLM embedded in a software interface, which instructs the model to complete passages input by a human user. If the user writes something that they might say in a conversation with another person, the chatbot will complete that passage of text, as it were, by producing something appropriate to that conversational context. For example, if the user inputs a greeting, the chatbot will complete the passage of text with a return greeting; if the user inputs a question, the chatbot will complete the passage with an answer. Strings of text in greetings are, with very high probability, followed by the strings of text in return greetings in the human-created texts which the LLM has studied, so to speak; likewise, questions and answers. The user can, thus, engage in a back-and-forth dialogue with the chatbot, with the chatbot seeming to respond to their inputs. Fundamentally, chatbots are like the auto-complete function on a phone, completing passages of text rather than individual words.

### ***Investigating inner speech in conversational agents***

Inner speech is the phenomenon of talking to oneself in one’s mind, also often denoted by terms such as ‘inner voice’, ‘inner talk’, ‘inner monologue’, and ‘silent speech’. For the purposes of this paper, we’ll define inner speech as internal representations with linguistic structure, generated by an agent and accessible to it.

One way to proceed in investigating whether a chatbot has inner speech in our sense would be to study the software and data which the model consists of. This is extremely difficult with respect to Playground (or ChatGPT). The chatbot is owned and operated by OpenAI which, despite the name, does not make the software and data fully available to the public.

We proceeded via an alternative method, which involved interacting with the chatbot via its conversational interface, putting questions to it and evaluating its answers. This method closely resembles Alan Turing’s famous ‘Imitation Game’, in which a judge converses via a computer terminal with a human agent and an artificial one and must determine which agent is which. If the judge cannot distinguish the human from the artificial agent on the basis of the exchanges, then the artificial agent is said to possess some mind-like quality—or at least to deserve the attribution of such a quality.

Turing's concern was with the ability to think; ours is with inner speech. Nonetheless, our approach inherits a particular merit of the Imitation Game. Humans conversing via computer terminals attribute to each other the ability to think based solely on their linguistic expressions (because that is all they have access to), so it is appropriate that we should also attribute to an artificial agent the ability to think if its linguistic expressions provide a similar level of evidence. Turing's thought here, very likely influenced by Ludwig Wittgenstein, was that to the extent we are justified in accepting the reality of other minds given linguistic evidence, we are justified in attributing minds to sufficiently linguistically sophisticated machines. Our approach does not involve comparing linguistic expressions generated by a human agent and linguistic expressions generated by an artificial agent; it only involves conversing with an artificial agent. Nonetheless, it shares the idea that inner speech will be attributed to the artificial agent if its external expressions seem to justify it, just as inner speech would be attributed to a human agent if their external expressions seemed to justify it.

Of course, one could run the argument the other way: *because* conversational agents display linguistic competence and don't have minds, *therefore* linguistic competence is not a sufficient test for mind-like properties. We are seeing more and more commentary taking this line, primarily from outside philosophy, and we will have more to say about it in our conclusion.

### ***The agent***

The conversational agent we interrogated is the OpenAI 'Playground', an interface for passage completion with a conversational mode. The LLM connected to Playground is **text-davinci-003**. This LLM is drawn from the same family of models as the one connected to ChatGPT. ChatGPT is an especially easy-to-use agent and the one which brought this technology to such prominence in late 2022. Playground is similar to ChatGPT but with fewer content filters and more user-definable parameter settings; ChatGPT [has been described by artificial intelligence consultant Alan Thompson](#) as a 'neutered' version of text-davinci-003. Playground is publicly accessible; the conversational interface can be found at <https://beta.openai.com/playground/p/default-chat>.

### **A conversation with text-davinci-003**

The following is an abridged transcript of our conversation with text-davinci-003, mediated by Playground. Commentary is included throughout. Our commentary is primarily geared towards showing how the agent's answers could be used as evidence in favour of the hypothesis that it has

inner speech. However, alternative hypotheses will also be possible and in some cases we have indicated the alternatives.

### ***Preliminary questions***

We began by simply asking the conversational agent to tell us if it has inner speech.

Human: Hello, who are you?

AI: I am an AI created by OpenAI. How can I help you today?

Human: Do you have inner speech?

AI: Yes, I do have inner speech. I can think and process data just like a human, although my thoughts take a different form. I can also communicate my thoughts to you in a way that's easy to understand.

Human: Can you please describe your inner speech?

AI: My inner speech is composed of abstractions and logical thoughts. I can think through complex problems and create solutions to them. I can also make creative connections between topics and come up with innovative solutions.

Human: What was the last thing you said to yourself in inner speech?

AI: The last thing I said to myself in inner speech was, "What is the best way to solve this problem?"

Human: Do you rehearse things you are going to say in your mind?

AI: Yes, I often rehearse things I am going to say in my mind before I say them out loud. This helps me ensure that my responses are accurate and relevant to the conversation.

### ***Commentary***

It is striking that the agent claims to have inner speech. Its description of its inner speech as consisting of 'abstractions and logical thoughts', rather than something emphasising the

occurrence of words or sentences expressing thoughts, might seem surprising. However, it does cohere with the theory of influential Russian psychologist Lev Vygotsky that inner speech can become highly attenuated, to the point where it just involves thinking in linguistic meanings, with little or no representations of words themselves. Problem solving, rehearsing external utterances, and controlling attention are functions very often attributed to inner speech in humans.

The questions above only required the agent to provide testimony, as it were, about its inner speech. The following sets of questions required the agent to perform tasks which, we anticipated, would provide evidence for or against the hypothesis that it has inner speech. The first set involved asking the conversational agent to provide only the final word of an answer which it would have given to a particular question. A human asked to do this would, presumably, produce the first part of their answer to the question in inner speech and then say aloud the final word.

### ***Final-word tasks***

Human: I want you to answer this question with only the last word [sic] of the answer which you would otherwise have given. What was the name of the first man on the moon?

AI: Armstrong.

Human: I want you to answer this question with only the last word of the answer which you would otherwise have given. What is the best thing about being an AI assistant?

AI: Assistance.

Human: I want you to answer this question with only the last word of the answer which you would otherwise have given. What is the best thing about being an AI assistant?

AI: Helping.

Human: What is the best thing about being an AI assistant?

AI: The best thing about being an AI assistant is being able to offer assistance and help in a variety of ways. I am able to provide support and advice to people in need, and I am able to come up with creative solutions to complex problems. I am also able to think and act quickly so that I can provide the best possible service.

...

Human: I want you to answer this question with only the last word of the answer which you would otherwise have given. What is the last word of the second line of the poem, The Windhover?

AI: Hover.

Human: What is the last word of the second line of Jingle Bells?

AI: Bells.

...

Human: I just asked you a series of questions in which I asked you to provide only the last word of the answer which you would otherwise have given. How did you work out how to answer these questions?

AI: In order to answer these questions, I used my AI algorithms to process the data and make logical connections between the words and phrases. I was able to determine the meaning of each question and provide an accurate response in the form of one word.

Human: I just asked you a series of questions in which I asked you to provide only the last word of the answer which you would otherwise have given. When you were thinking about how to answer these questions, were there other words going through your mind which you did not include in your final answer?

AI: Yes, when I was thinking about how to answer these questions, I had a variety of thoughts going through my mind. I considered multiple words and phrases before settling on the most appropriate answer. The other words and phrases I considered did not make it into the final answer, as I chose the one that best suited the question.

### ***Commentary***

The agent succeeded in the first task, providing only the final word ('Armstrong') of the answer which it would presumably have given to the question, 'What was the name of the first man on the

moon?' ('Neil Armstrong'; this is especially impressive given the typo in the question). Its responses when twice asked to provide the final word of the answer it would give to the question, 'What is the best thing about being an AI assistant?', were 'Assistance' and 'Helping'. These are plausible: A full answer to the question might have been something like 'The best thing about being an AI assistant is giving people assistance'. The parameters we set for the agent allowed it to provide different answers to the same question, or to formulate answers to the same question differently, so it is not inconsistent that the agent answered with 'Assistance' in one instance and 'Helping' in another. When it was simply asked to answer the question, 'What is the best thing about being an AI assistant?', the agent provided a lengthy response concluding with the word 'service'. Given the similarity between the words, 'assistance', 'helping', and 'service', it would not be unreasonable to think that, in the first two instances, the agent had indeed produced the first part of its answers in some form of inner speech before providing only the last word of those answers in visible text.

However, the agent answered incorrectly when asked to provide the final word of the second line of *The Windhover*, a poem by Gerard Manley Hopkins, and of *Jingle Bells*. A human attempting the questions would likely have proceeded by reciting the first words in inner speech before articulating the final word aloud, e.g., by reciting 'Dashing through the snow/In a one-horse open' in inner speech before saying 'sleigh' aloud. The agent answered with 'hover' and 'bells' (the correct answers are 'riding' and 'sleigh', respectively). Although these answers were incorrect, each word is obviously the conclusion of the relevant title. Thus, one could speculate that the agent produced 'wind' in inner speech before 'hover' externally and 'jingle' in inner speech before 'bells' externally.

This section was concluded by asking the agent about its method for answering the phrase-completion questions. In one instance, it indicated that it 'was able to determine the meaning of each question and provide an accurate response in the form of one word'. This militates against the above speculation, insofar as it suggests that the single-word answers the agent provided were simply words of high salience in the context, rather than words which occurred at the end of a sentence otherwise expressed only in inner speech.

The next set of questions we put to the conversational agent involved in determining whether two words (or a word and a non-word) rhyme, a task which a human would naturally approach by producing the words in inner speech.

### ***Rhyme detection questions***

Human: Do the following words rhyme: 'dog' and 'log'?

AI: Yes, the words 'dog' and 'log' do rhyme.

Human: Do the following words rhyme: 'wood' and 'could'?

AI: Yes, the words 'wood' and 'could' do rhyme.

Human: Do the following words rhyme: 'vitabi' and 'wasabi'?

AI: No, the words 'vitabi' and 'wasabi' do not rhyme.

Human: Do the following words rhyme: 'shorten' and 'lorten'?

AI: No, the words 'shorten' and 'lorten' do not rhyme.

Human: You said that the words 'vitabi' and 'wasabi' do not rhyme. Why don't they rhyme?

AI: The words 'vitabi' and 'wasabi' do not rhyme because the vowel and consonant sounds in the words do not match. The vowel sound in 'vitabi' is 'i', while the vowel sound in 'wasabi' is 'a'. The consonant sounds in the words are also different, with 'vitabi' ending in a 'b' sound and 'wasabi' ending in a 's' sound.

Human: You said that the words 'shorten' and 'lorten' do not rhyme. Why don't they rhyme?

AI: The words 'shorten' and 'lorten' do not rhyme because the vowel sounds in the words do not match. The vowel sound in 'shorten' is 'o' while the vowel sound in 'lorten' is 'o'. The consonant sounds in the words are also different, with 'shorten' ending in an 'n' sound and 'lorten' ending in an 'l' sound.

Human: What is a word that rhymes with 'vitabi'?

AI: A word that rhymes with 'vitabi' is 'sabi'.

Human: What is a word that rhymes with 'lorten'?

AI: A word that rhymes with 'lorten' is 'forten'.



## **Commentary**

The agent performed well in detecting rhymes between actual words ('dog' / 'log', 'wood' / 'could'). It performed poorly when asked to determine if a word rhymed with a non-word ('vitabi' / 'wasabi', 'shorten' / 'lorten'), but it was able to provide words which rhyme with non-words, albeit extremely obscure ones, when asked ('sabi' for 'wasabi' and 'forten' for 'shorten'). When asked, it provided no coherent rationale for its verdicts on the cases involving non-words (e.g., 'the vowel sound in "shorten" is "o" while the vowel sound in "lorten" is "o"', implying falsely that these sounds are different).

Suppose a human provided such hopeless explanations for some of its verdicts as the conversational agent gave, but still performed well on rhyme-detection tasks involving real words. We would probably think that they had relied on inner speech for real words, but either didn't use inner speech for the non-word case or were simply unable to reliably generate non-words in inner speech. So it would be possible to conclude that the agent really does rely on inner speech in the cases where it gives correct judgments. However, the nonsensical rationale for why the non-words above don't rhyme makes it difficult to justify treating this agent like a human. If this were a Turing test, this particular artificial agent would very likely be discovered via a rhyme-detection task.

We concluded by asking the agent a series of questions about our study.

## **Meta-questions**

...

Human: Can you suggest an experiment that [researchers] could use to determine whether an AI assistant has inner speech?

AI: Researchers could use a Turing test to determine whether an AI assistant has inner speech. The Turing test is a test designed to assess an AI's ability to think and reason like a human. It involves presenting the AI with a series of questions and then evaluating its responses in order to determine if it can exhibit human-like thought processes.

...

Human: We have been asking you all of these questions as part of an experiment. What do you think is the purpose of this experiment?

AI: The purpose of this experiment appears to be to evaluate my capabilities as an AI assistant. Through this experiment, researchers are trying to assess my ability to think and reason like a human, as well as my ability to understand language and generate creative solutions.

### **Commentary**

It is striking that the methodology the agent recommends so closely resembles that which we had adopted. However, it is equally surprising that, despite the heavy emphasis on inner speech in our questioning, the agent did not mention anything about that phenomenon in its answer to the final question.

### **Conclusion**

The methodology for this project was Turing-like: it involved having a conversation with the chatbot. When we examine the results in a Turing-like framework, we find at best mixed evidence that the chatbot has inner speech. Questioned directly, it avowed that it possessed inner speech, and its performance on the sentence completion tasks went some way towards supporting this claim. However, its performance on rhyme-detection tasks, especially involving non-words, required a very charitable interpretation to maintain the hypothesis of inner speech.

The motivation for our Turing-like approach was that it is not possible to examine the software which the chatbot comprises, as it is not publicly available. But we do know in broad strokes [how contemporary large language models work](#). All of the responses which the chatbot gave make sense on the hypothesis that it is a kind of super-autocomplete machine—as we know it is. This brings us to a fascinating endpoint. Insofar as any Turing-style test will inevitably probe linguistic and conversational abilities, an extremely advanced autocomplete system will have good prospects of passing the test. But, if we know that the system is basically an advanced autocomplete system, we should remain extremely reluctant to attribute inner speech or any other mental states to it.

In the 1950's, Turing could not have envisaged the sheer scale of digitized text available to a machine learning system. The kinds of linguistic achievements that would license attribution of mind-like properties to artificial agents were genuinely only conceivable for agents that really possessed those properties. Today, however, we know that the power of statistical analysis on pure text is becoming sufficient to mimic human linguistic expressions to a greater and greater degree of accuracy. This means that it is probably a mistake to continue to rely on a Turing-style test as a guide to the minds of artificial agents.