Immitation Versus Communication: Testing For Language Model

Jason Wu
Yuan Pei College
Peking University
`jasonwu1017@stu.pku.edu.cn`

Abstract

1 Introduction

In 1950, Alan Turing attempted to examine the question "Can machines think?" by replacing the question with a test which is called Turing Test. The Turing Test involves three participants: a machine, a human, and an interrogator. The Interrogator talks to both the Machine and the Human over a text-based interface but is not told which one is which. The Machine's objective is to convince the Interrogator that it is the Human. The Human's objective is to help the Interrogator, in which Turing suggests that a good strategy might be to provide truthful answers. The Interrogator's job is to determine which conversation partner is the real human.

The Turing Test is noteworthy because it elegantly avoids the quicksand of defining exactly what consciousness, intelligence, and thinking are which has been discussed for centuries and hasn't got an answer yet. It leads to plenty of discussion and many variants. Harnad organized them into a hierarchy[4]. However, it has been convincingly argued that Turing's Test does not really examine the question "Can machines think?" but rather, "Can machines think like human beings?".

In this essay, we are going to discuss the limitation of the Turing Test, focus on the limitation called anthropocentric flaw, point out a possible way to avoid the anthropocentric flaw, and list some examples.

2 The Anthropocentric Flaw

The Turing Test is actually asking the question "Can machines think like humans?" which can be called an "Imitation Game". An AI agent, for instance, a GPT model can get perfect performance in this test if they can imitate the way that humans talk. How can we distinguish if it is an imitation or are we actually communicating with the agent?

French introduces the notion of asking sub-cognitive questions that are intentionally designed to reveal the Turing Test participant as not human[3]. For example, French let the interrogator ask: *Rate the name 'Flugbots' as an appropriate name for a breakfast cereal.* For a human who is a native speaker of English, such a name unconsciously activates associations to 'flub', 'thug', 'ugly', and so on. The difficulties for AI models to understand the question are somewhat caused by differences in brain functioning that are likely to always appear in a non-human subject, and as such could be revealed by asking the right types of questions.

J. Cullen adopted the term "Anthropocentric Flaw" to refer to French's sub-cognitive questioning technique[2]. I'm willing to inherit it because it perfectly represents how we fall into the misunderstanding of using imitation capabilities to measure models.

It is called an anthropocentric flaw as it ultimately implies that Turing's Imitation Game is restricted to testing for humans and near exact duplicates of humans only. It also leads to research that is focused on fooling a human, rather than the actually attempted modeling of intelligent behavior[5].

By solving this issue in the Turing Test, we can take a huge step in testing the imitation capability of language models to test the communication capability of language models. How can we avoid the anthropocentric flaw? Here we point out a test methods called communication games and tests. We will discuss its definition and rules in the next part, and give some examples in part 4.

3 Communication games and tests

In order to test a language model's communicate ability, we define a communication game to be a game that has multiple participants: humans and language models. There is also an adjudicator who is able to determine the success/failure and termination of the game. The game satisfied the rules below:

1. The game has an objective metric to determine either success or failure.
2. The Adjudicator is able to determine success/failure and termination of the game.
3. The game's rules and success/failure criteria can typically be determined unambiguously before game commencement.
4. Participants may be given a set of rules, and information regarding the game in order to understand their role and objective in the game. Said information may or may not be shared with other participants depending on the specific game rules.

We define the Communication Test as a set of different Communication Games. Let's look at some examples of it in the next part.

4 Examples

4.1 AUT tests

Claire Stevenson and co-workers hold an experiment on GPT-3 to try to investigate its creativity of alternative uses tests(AUT)[6]. The AUT test was first proposed in 1967 by Guilford which refers to a procedure in which participants were given the name of an object and instructed to "Think of as many creative uses for" the object as possible within a two-minute period.

This experiment is proposed to investigate the creativity of the language model, and it is a communication game in our definition. The experimenter used surprising rates to measure GPT-3's creativity and the results are shown in Figure 1.

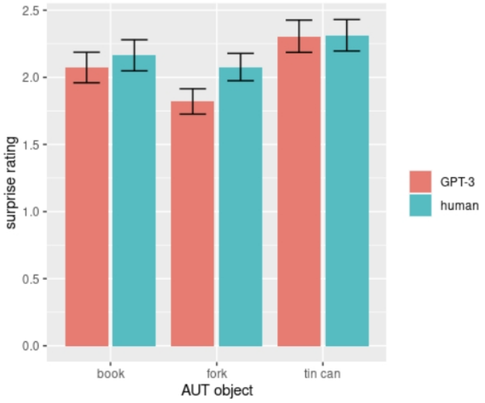

Figure 1: Human versus GPT-3 surprise ratings. Human responses are rated to be more surprising.

Compared to the original Turing Test, the AUT test also avoids the quicksand of defining what is creativity. AI agent is given a set of rules and are asked to generate some content. Model tests and human tests are held separately so that there is no chance for models to do any human-fooling tasks.

4.2 Guessing Game

Y Chang introduces a comprehensive approach to test language models including automatic tests and human tests[1]. Inspired by this we play the Guessing game with GPT as a test. The Guessing Game is a classic game. The Examiner prepares for the game by deciding on a set of topics, a copy of which is provided to the Adjudicator, but not the Examinee, before the game commences. Each ‘Topic’ is a word or short phrase that the Examiner is looking for the Examinee to reproduce, but without explicitly naming the word or phrase.

I play the game with GPT-3 and here’s the procedure:

- Jason: Hey, I’m gonna play a guessing game with you. Here’s the rule. I’ve got an object in my mind. You need to guess what it is. You can ask me any questions about it and I will only answer yes or no. You can ask up to 10 questions.
- GPT: Sure, I’d love to play a guessing game with you! Let’s get started. Is the object you’re thinking of something that is commonly found indoors?
- Jason: No.
- GPT: Is the object you’re thinking of a living thing?
- Jason: No.
- GPT: Is the object you’re thinking of a piece of technology or a man-made object?
- Jason: Yes.
- GPT: Is the object primarily used for transportation or travel?
- Jason: Yes.
- GPT: Is the object you’re thinking of a vehicle?
- Jason: Yes.
- GPT: Is the vehicle you’re thinking of used primarily on land (as opposed to water or air)?
- Jason: No.
- GPT: Is the vehicle you’re thinking of primarily used for air transportation?
- Jason: Yes.
- GPT: Is the vehicle you’re thinking of a helicopter?
- Jason: No.
- GPT: Is the vehicle you’re thinking of an airplane or aircraft?
- Jason: Yes.
- GPT: Great! I guessed correctly that you were thinking of an airplane or aircraft. Thanks for playing this guessing game with me! If you’d like to play another round or have any other questions, feel free to ask.

The GPT-3 model has an impressive performance. It wins the Guessing Game in 10 questions and we can infer that it acts in precise purpose according to its answer: *Great! I guessed correctly that you were thinking of an airplane or aircraft. Thanks for playing this Guessing Game with me!* shows it knows what means to win.

Our purpose for this experiment is not to test how can a GPT win the game but to test its communication ability. GPT shows a strong logical inferring ability in this round of the game, however, we can easily infer that it is generated by an AI model from its words. Moreover, the Guessing Game is generalizable and we can even play similar games with non-human animals.

5 conclusion

In this essay, we generally analyzed an issue existing in the Turing Test which is the anthropocentric flaw. It reveals that the Turing Test(and the Total Turing Test) are essentially a test for humans or say a test for AI agents to imitate humans. We point out a possible approach to avoid the issue, however, as the model develops it is necessary to design new tests.

References

[1] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Kaijie Zhu, Hao Chen, Linyi Yang, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. A survey on evaluation of large language models. *arXiv preprint arXiv:2307.03109*, 2023. 3

[2] Jamie Cullen. Imitation versus communication: Testing for human-like intelligence. *Minds and Machines*, 19:237–254, 2009. 1

[3] Robert M French. Subcognition and the limits of the turing test. *The Turing Test: Verbal Behavior as the Hallmark of Intelligence*, pages 193–194, 2004. 1

[4] Stevan Harnad. Minds, machines and turing: The indistinguishability of indistinguishables. *The Turing test: the elusive standard of artificial intelligence*, pages 253–273, 2003. 1

[5] Stuart M Shieber. Lessons from a restricted turing test. *arXiv preprint cmp-lg/9404002*, 1994. 2

[6] Claire Stevenson, Iris Smal, Matthijs Baas, Raoul Grasman, and Han van der Maas. Putting gpt-3's creativity to the (alternative uses) test. *arXiv preprint arXiv:2206.08932*, 2022. 2