WHO JUDGES THE TURING TEST BETTER: EXPERTS OR CHATGPT?

Natalija Gajić [1] [ORCID 0000-0001-5453-7979], *Vladimir Mandić* [1] [ORCID 0000-0001-6996-2222]
[1] *University of Novi Sad, Faculty of Technical Sciences, Department of Industrial Systems and Management, Serbia*

Abstract: *The continuous advancement of language models, such as ChatGPT-4, has brought remarkable improvements in their conversational capabilities, blurring the lines between human and machine interaction. The use of the Turing test becomes particularly relevant in the context of distinguishing machinel-generated output from human responses. This paper investigates the comparative performance of human experts and ChatGPT-4 in evaluating the Turing test. To conduct the study, a dataset comprising human and ChatGPT-3 responses to HR interview questions was curated. People working in software companies were recruited, while ChatGPT-4 served as a machine-based judge. In the Turing test, human participants exhibited lower confidence in their decisions compared to ChatGPT-4, but ultimately reached 90.91% accuracy in rating conversations, while ChatGPT-4 demonstrated an accuracy of 59.1%. This study can serve as a starting point for exploration into the evaluation of human experts and ChatGPT-4 performance in the Turing test, and the idea of machines aiding in the difficult task of differentiating between humane- and machine-generated text.*

Key words: Turing test, ChatGPT, Empirical Study

1. INTRODUCTION

The *Turing test* (Turing, A.M., 2009) has long served as a benchmark for differentiating between humane- and machine-generated text. As artificial inteligence (AI) systems continue to advance, the Turing test becomes increasingly significant in efforts to accurately identify and attribute text to its rightful source. Through its training on vast amounts of diverse data, *large language models* (Radford et al., 2018) such as ChatGPT-3.5 and ChatGPT-4[1] have developed an exceptional ability to understand and respond to human queries, generating coherent and contextually relevant text. It is important to remain aware of the ethical considerations and potential challenges that arise with such advancements. By acknowledging the presence of AI-generated text, we can leverage its benefits while also maintaining transparency and accountability in our digital interactions.

In Turing's imitation game (Turing, A.M., 2009), also known as the Turing test, a human interrogator engages in a conversation with two participants, one being a machine and the other a human. The interrogator is unaware of which participant is human and which is a machine. By subjecting both a human and a machine to a conversation through text-based communication, the test aims to evaluate whether the machine's responses can be indistinguishable from those of a human. While the role of the imitating player (machine) in the test was thoroughly investigated through the years, the role of the interrogator has also become the subject of research (Hernández-Orallo, 2020). The simple fact that humans are employed as interrogators has been seen to be something that undermines the test's credibility by some (Hayes and Ford, 1995) and as an important indicator of human imperfections by others (Warwick and Shah, 2016).

The *inverted Turing test* is the first proposal of a test where the judge is a machine (Watt, 1996). In the inverted Turing test, the judge is the one being evaluated. The inverted Turing test checks if the system's powers of discrimination are equivalent to those of an expert human judge. This is different from the reverse Turing test (Von Ahn et al., 2003), with its implementations usually known as CAPTCHAs, where the judge (machine) is not evaluated. The reverse Turing test brings the idea of the test being totally automated, as the human or machine to be detected does not have to be compared against a real human.

[1]Advanced natural language processing models, developed by OpenAI, based on "Generative Pre-trained Transformer" architecture: https://openai.com/blog/chatgpt

(Hernández-Orallo, 2020) states that relying on humans to judge the result of the Turing test usually leads to problems of subjectivity, bias, reliability, and scalability. The study is mostly focused on the idea of machines replacing the role of humans in the Turing test and the task of evaluating intelligence. The study presents the idea that humans can be selected and trained to become better judges in the Turing test, but that in the end, they will reach an evaluation quality plateau because of their mental resources, motivation, and capability. This plateau can nonetheless be broken by machines. Similarly, this paper explores the idea that machines can possibly take the role of discriminators between human- and machine-generated text. To the best of our knowledge, there are no studies that directly compare human and machine performance in distinguishing between human-generated and machine-generated text.

The objective of this study is to investigate the comparative performance of human experts and ChatGPT in judging a version of the Turing test. Specifically, it investigates their ability to distinguish between human- and machine-generated answers to HR interview questions used in recruiting software engineers. In order to achieve the objective, we designed a two-phase empirical protocol. The rest of the paper is organized as follows: Section two provides a comprehensive explanation of the methods used in the empirical study, while section three presents the results from the experimental phase. The Discussion section is dedicated to analyzing the results and finally, in the last section, conclusive remarks and findings of the study are drawn.

2. METHODS

In the first phase of the study, human and machine-generated conversations were collected. In the second phase, a version of the Turing experiment was conducted on humans and ChatGPT-4.

2.1 Data collection

Ten questions were collected from internet blogs presenting the most common HR interview questions in the field of software engineering. Two types of questions were collected: experience-based and opinion-based questions. Each conversation, either with humans or generated by AI consisted of five randomly chosen questions from the collected set of questions. Answers from humans were collected through an online meeting, and answers from ChatGPT-3.5 were collected using the OpenAI API.

2.1.1 Interviews

An online meeting was held on Teams with each of the participants. Before the meeting, participants were instructed to answer the questions as they would in a real HR interview setup. Participants were cautioned against disclosing personal information to safeguard their anonymity. For questions about previous experience, they were instructed to share only permissible information.

Participants were asked five randomly chosen questions from the collected set of questions. The meeting transcript was collected to obtain the answers. Participants were then asked to confirm the validity of the obtained answers. They were allowed to change the wording and add omitted information to the transcript, but were not allowed to add information that was not said during the interview.

2.1.2 Prompting ChatGPT-3.5

ChatGPT answers were collected by prompting the gpt-3.5-turbo model through the OpenAI chat completion API. A Python script was written for generating prompts and collecting the answers. Prompting was done by following the best practices and strategies for getting a better response from the chatbot provided by OpenAI (OpenAI Platform, 2023).

Twelve conversations were generated by ChatGPT-3.5. Each conversation consisted of five random questions. The answer to each of the questions was collected by providing the model with information about the questions and answers that were previously generated in the conversation. This way, the model had the context of the conversation the same way the human did. The model was tasked to answer the questions as a person with a job in software development would in a real HR interview setup. Also, it was provided with the characteristics of the job position, preferred technology, and level of experience of the person with a job in software development. Lastly, it was instructed to respond using

approximately the same number of sentences as the human providers used in their answers, which was between two and four sentences.

Characteristics of the person the model was imitating were set up to match the characteristics of the interviewees who gave answers in the first part of the data collection phase. Six different examples of characteristics were produced. Prompting the model with different characteristics of the human allowed for different answers to the same question. Also, the model was prompted with the same characteristics twice with a different value of the temperature parameter. For each conversation, a random value of the temperature was chosen from the interval [1, 1.5]. Responses with higher temperatures display greater linguistic variety, while the low one represents grammatically correct and deterministic text (Ippolito et al., 2019). Testing showed that answers obtained with a temperature higher than 1.5 become logically incoherent, so temperatures higher than 1.5 were not considered.

2.2 The Turing experiment

For the experiment to be valid, the observer should not have advanced knowledge of the entity that is providing the answers, as they can use this knowledge to bypass the discrimination (Watt, 1996). For this reason, the questions are answered by ChatGPT-3.5, and judged by ChatGPT-4, a different identity. Also, none of the participants who answered the HR interview questions in the data collection phase of the experiment didn't participate in the Turing experiment. It is also to the best of the researcher's ability ensured that the human judges have no personal connections and knowledge of the people who were providing the answers. A Django application was created for administering the experiment on humans, and ChatGPT-4 answers were collected using the official ChatGPT website.

2.2.1 Human experiment

When participants agreed to take part in the research, they were sent the link through which the experiment was accessed. The link contained a unique code by which each participant was identified.

Each participant was shown three HR interview conversations. In order to avoid the situation where the conversation is evaluated as if it was generated by a machine because it resembles the conversation that was previously shown to the participant, it was decided to only show one machine-generated conversation to each of the participants. The three conversations were randomly chosen and the order of the human-generated and machine-generated conversations was random. Although the number of human and machine-generated conversations was predetermined, participants were not aware of this. They were told that the ratio of human-generated to machine-generated conversations was unspecified.

Some previous studies showed that the perception of humanness changes as participants go through successive evaluations (Candello et al., 2017) and that the evaluation of the first question-answer can affect the evaluation of the subsequent pairs (Ariely, 2010). Having this in mind, participants were presented with one conversation at a time, without the possibility of revisiting previous conversations. This approach minimizes the chances of direct comparison. On the other hand, the possible influence of the first evaluation is still present.

Prior to running the experiment, the research was tested on pilot participants to identify and correct issues with the study setup. The test helped determine the time duration provided to participants for rating the conversation. Each conversation was shown to the participant for 5 minutes and during this time participant was able to rate their level of confidence that the answer was given by a human, using a Likert scale of 1-4.

2.2.2 ChatGPT-4 experiment

For each of the 24 conversations that were collected in the first phase of the research, ChatGPT-4 was prompted to decide whether the answers were given by a human or machine. The prompt consisted of the same explanation of the task that the humans were given at the beginning of the experiment. Similarly to humans, the model was instructed to give a rating of the level of confidence that the answer was given by a human using the same Likert scale that was given to humans. Prompting was done by following the best practices advised by OpenAI (OpenAI Platform, 2023), and by using the chain of thought zero-shot prompting method (CoT zero-shot) (Kojima et al., 2022).

3. RESULTS

3.1 Participants

For both the data collection phase and the Turing experiment, participants were people working in Novi Sad, Serbia. For the data collection phase, there were 11 software developers and one Quality Assurance tester. For the Turing experiment, people with job positions in software companies besides development were included. All participants self-reported knowledge of English on a scale of 1 (able to have a simple conversation) to 10 (can communicate fluently). Eight men and four women between the ages of 24 and 45 were interviewed. The mean age was 31 (SD=9.577) and their self-reported knowledge of English on a scale of 1 to 10 was 7.5 (SD=1.931). Information regarding participants' job positions and years of experience was used for composing examples of characteristics of the human ChatGPT-3.5 was prompted to answer as.

For the Turing experiment, there were 24 participants, 18 males and 6 females between the ages of 22 and 46. Their mean age was 28 (SD=6.169) and their self-reported knowledge of English on a scale of 1 to 10 was 8.6 (SD=1.213).

3.2 Turing experiment

The Shapiro-Wilk test suggested a significant deviation from the normal distribution for the AI ratings (statistic=0.91, p-value=0.052). The Wilcoxon rank-sum test was applied in order to assess the significance of the difference between human and AI ratings of the same conversations. Aggregated human rating of each of the conversations was established as the median of all ratings given by humans for that conversation. The ChatGPT model gave the same answer when being prompted multiple times, so no aggregation was done, and only one extracted rating the model gave was taken. The Wilcoxon rank-sum test yielded a statistically significant result (statistics=2.441, p = 0.014), indicating a substantial difference between the paired ratings being compared. Figure 1 shows density distributions of absolute rating errors for (A) humans and (B) ChatGPT-4, and Figure 2 shows confusion matrices for humans (A) and ChatGPT-4 (B) judges.

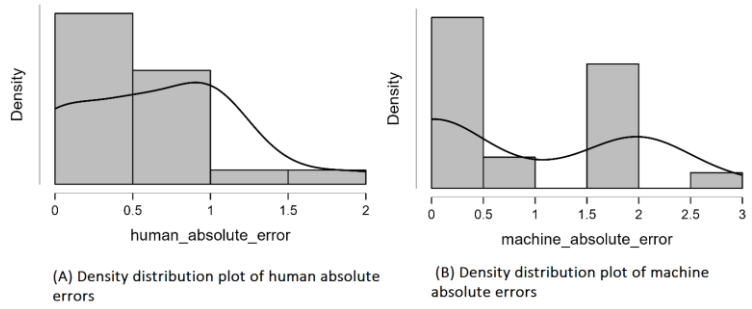

Figure 1: Density distributions of absolute rating errors

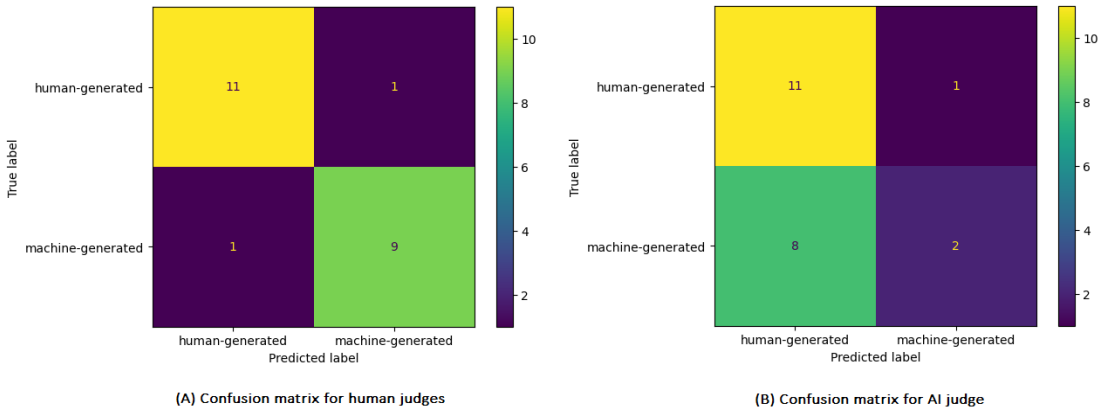

Figure 2: Confusion matrices of ratings

ChatGPT-4 successfully rated human-generated conversations, but only managed to correctly rate two out of ten machine-generated conversations. Human accuracy reached 90.91% while ChatGPT-4 demonstrated an accuracy of 59.1%, only marginally better than random chance.

Although human judges reached higher accuracy, it is also valuable to examine with what level of confidence. Figure 3 shows distribution plots of ratings of human-generated conversations by human judges (A) and AI judge (B) and their ratings of machine-generated conversations (C) and (D).

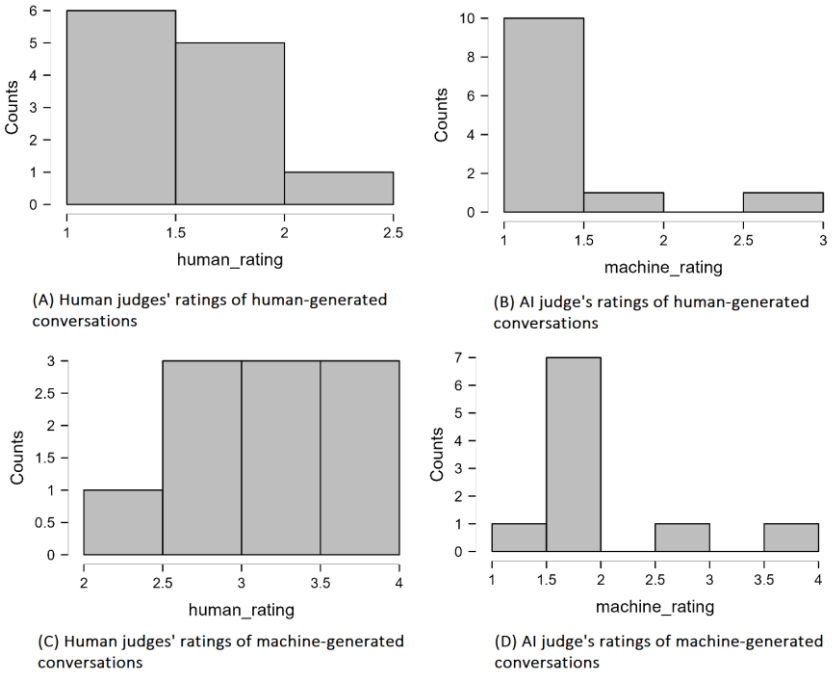

Figure 3: Distribution plots for ratings

4. DISCUSSION

Density distributions in Figure 1 show that both humans and ChatGPT-4 made errors, but in the case of the ChatGPT-4 judge, there was a greater number of ratings with an absolute error between 1.5 and 2. With the rating interval width of 4, these errors indicate that ChatGPT-4 inaccurately rated conversations with lower confidence. Confusion matrices shown in Figure 2 confirm this. ChatGPT-4 correctly rated a small porportion (20%) of machine-generated conversations, and this factor may account for the statistically significant difference between ratings provided by humans and ChatGPT judge.

With a 1 rating denoting high confidence in human-generated conversations, ChatGPT-4 mostly rated human-generated conversations correctly with high confidence (90.91% of correct ratings), while humans tended to be less confident (54.54% of correct ratings). This tendency of human judges to give ratings with lesser confidence was true for machine-generated conversations as well (33.33% of correct ratings). Although ChatGPT-4 correctly rated only two machine-generated conversations, the model showed low confidence (rated with a 2) in seven out of eight cases.

The limitations of the study should be taken into account when interpreting the results. The first limitation relates to the small number of human participants in the Turing experiment and the limited amount of conversation examples. This may restrict the generalizability and robustness of the findings. Additionally, certain effects could be overlooked and more evident in studies involving a significantly larger number of participants and data points. With this in mind, this study represents just a starting point for the investigation of the performance of human experts and ChatGPT-4 in evaluating the Turing test.

Another limitation is the lack of interactive conversations in the Turing test design; participants only evaluated static conversations. Interactive conversations could introduce considerable variability, affecting the study. Providing recorded conversations ensures standardized experiences for subjects, enabling answer comparability. Another thing that should be taken into account is that the majority of participants were Serbians, but the language of the study was English, and this could have also caused

some difficulties (even though the participants were controlled by making the advanced level of English a criterion for participation).

5. CONCLUSIONS

Results showed that humans performed better in the Turing test but were less confident in their decisions than ChatGPT-4, especially when it comes to human-generated conversations. The biggest difference between human judges and ChatGPT-4 was in rating machine-generated conversations. While human judges showed good performance in rating both human- and machine-generated conversations, ChatGPT-4 correctly rated primarily human-generated conversations and incorrectly rated the majority of machine-generated conversations, although with lower confidence.

For future research, it would be valuable to explore the reasoning strategy behind the judge's decision, both human and ChatGPT-4's. Furthermore, exploring how various prompts and contexts given to ChatGPT-4 can influence its performance could be worth investigating. This could potentially reveal common mistakes and propose approaches to effectively address them.

6. ACKNOWLEDGMENT

The research presented in this paper is partially supported by the project "Implementation of the results of scientific research work in the field of Industrial Engineering and Management in DIIM teaching processes with the aim of their continuous improvement", at the Department of Industrial Engineering and Management, Faculty of Technical Sciences, University of Novi Sad, Republic of Serbia.

7. REFERENCES

Ariely, D., 2010. Predictably irrational: the hidden forces that shape our decisions. Math Comput Educ, 44(1), p.68. Available from: doi: 10.5465/amp.2009.37008011

Candello, H., Pinhanez, C. and Figueiredo, F., 2017, May. Typefaces and the perception of humanness in natural language chatbots. In Proceedings of the 2017 chi conference on human factors in computing systems (pp. 3476-3487).

Hayes, P. and Ford, K., 1995, August. Turing test considered harmful. In *IJCAI (1)* (pp. 972-977).

Hernández-Orallo, J. (2020). Twenty Years Beyond the Turing Test: Moving Beyond the Human Judges Too. Minds and Machines. 30(4):533-562. Available from: doi: 10.1007/s11023- 020-09549-0

Ippolito, D., Kriz, R., Kustikova, M., Sedoc, J., & Callison-Burch, C. (2019). Comparison of diverse decoding methods from conditional language models. arXiv preprint arXiv:1906.06362.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. *Advances in Neural Information Processing Systems*, *35*, 22199-22213.

OpenAI Platform. *GPT Best Practices. Six strategies for getting better results.* Available from: https://platform.openai.com/docs/guides/gpt-best-practices/six-strategies-for-getting-better-results [Accessed 14th July 2023]

Radford, A., Narasimhan, K., Salimans, T. and Sutskever, I., 2018. Improving language understanding by generative pre-training.

Turing, A.M., 2009. *Computing machinery and intelligence* (pp. 23-65). Springer Netherlands.

Von Ahn, L., Blum, M., Hopper, N.J. and Langford, J., 2003. CAPTCHA: Using hard AI problems for security. In *Advances in Cryptology—EUROCRYPT 2003: International Conference on the Theory and Applications of Cryptographic Techniques, Warsaw, Poland, May 4–8, 2003 Proceedings 22* (pp. 294-311). Springer Berlin Heidelberg.

Warwick, K. and Shah, H., 2016. The importance of a human viewpoint on computer natural language capabilities: a Turing test perspective. *AI & society*, *31*, pp.207-221. Available from: doi: 10.1007/s00146-015-0588-5

Watt, S., 1996. Naive psychology and the inverted Turing test. *Psycoloquy*, *7*(14), pp.463-518.