Judging the Judges: Displacing and Inverting the Turing test to Investigate the Interrogator

Ishika Rathi (irathi@ucsd.edu)
Department of Cognitive Science
UC San Diego

Benjamin K. Bergen (bkbergen@ucsd.edu)
Department of Cognitive Science
UC San Diego

Cameron R. Jones (cameron@ucsd.edu)
Department of Cognitive Science
UC San Diego

Abstract

The Turing test typically evaluates machine intelligence by asking whether a human judge can distinguish between human and AI conversational behavior. But the test also serves as an evaluation of the judge, upon whose discriminative capabilities the merit of the test depends. We investigate this dependency by replicating two variations of the Turing test: (1) a displaced test, where human participants judge transcripts of previously conducted interrogations, and (2) an inverted test, where AI systems make similar judgments. Comparing these with traditional interactive tests, we find that displaced judges perform similarly to interactive judges, and LLM judges perform significantly worse than humans. This challenges assumptions about the importance of real-time interaction, and suggests that accuracy is not significantly impacted by displacement, but may be impacted by differences in a judge's model of human vs. AI behavior. Our results have implications for societal risks of AI, as systems that can consistently deceive both interactive and passive observers could enable large-scale online impersonation and manipulation.

Introduction

The Turing test was originally proposed by Alan Turing (1950) as a test of machine intelligence. In the test, a human interrogator speaks to two witnesses (one human and one machine) by a text-only interface. If the interrogator cannot guess which of the two witnesses is the human, the machine is said to have passed. Turing thought that the ability to imitate human intelligence in so flexible a medium as language would provide evidence that we ought to attribute something like this intelligence to the machine itself.

In addition to evaluating the machine, the Turing test also serves as an evaluation of the interrogator. A successful interrogator must have sufficiently accurate mental models of the processes and limits of human behavior in order to correctly discriminate between human and machine responses (Watt, 1996). These mental models allow the interrogator to pose strategic questions that one type of agent would likely answer differently from the other, and identify responses that a machine (or a human) either would or could never produce. On the other hand, an interrogator who lacks an understanding of either people or machines might easily be fooled into misclassifying witnesses based on superficial tricks.

Turing's proposal has garnered much discussion and controversy over the years (French, 2000; Epstein, Roberts, & Beber, 2009; Saygin, Cicekli, & Akman, 2000), and a central theme of this controversy has been whether human interrogators are reliable judges of whether a system is intelligent (Marcus, Rossi, & Veloso, 2016; Hayes & Ford, 1995). In particular, some critics have argued that people have a tendency to anthropomorphize inanimate systems. Weizenbaum (1966) found that many subjects attributed complex psychological states such as understanding and empathy to a simple rules-based chatbot called ELIZA. The tendency to attribute mental states to simple dialogue agents has come to be known as the ELIZA effect, and continues to be seen as a central problem for interpreting the Turing test (Mitchell, 2024).

Practical challenges with interpreting the Turing test resurfaced in recent years due to the advent of Large Language Models (LLMs), which can generate humanlike text on the basis of distributional language statistics (Boisseau, 2024; Gonçalves, 2024). Recent empirical studies have evaluated LLMs in a 2-party version of the Turing test, where each interrogator speaks to only one witness and has to guess whether they are speaking to a human or a machine (Jannai, Meron, Lenz, Levine, & Shoham, 2023; Jones & Bergen, 2023, 2024b). In one such study, Jones and Bergen (2024b) found that people were at chance in determining whether GPT-4 was a human or a machine, implying that contemporary LLMs pass the 2-party test.

These results raise questions about what factors impact the accuracy of a judge's determination, and the ongoing role of the test as a measure of intelligence. First, Turing (1950) stressed the interactive nature of the *imitation game*, as crucial to its flexibility: likening it to a *viva voce* exam to determine if 'one really understands something or has "learnt it parrot fashion."' (p. 446). How important is this interactivity in correctly discriminating between humans and machines? Secondly, in theory, a Turing test interrogator requires a rich internal model of other human minds, or a meta-cognitive awareness of the differences between human and machine behavior, in order to accurately discriminate between the human and machine. How well would LLM-based systems themselves do at making Turing test determinations? Would they make the same determinations as people and for the same reasons? To the extent that LLM-based systems can make determinations similar to humans, it would suggest we can attribute to them a similar model of human minds, or a similar awareness of the differences between human and AI behavior.

We replicated an experiment by Rathi, Taylor, Bergen, and Jones (2024) conducting two variations of the Turing test. The first—a *displaced* Turing test—investigates the impor-

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

tance of interactive engagement by having a separate group of human participants make judgments about whether a witness is a human or a machine on the basis of transcripts of previously-conducted interactive Turing tests. The participant is 'displaced' in that they are not present to interact with the witness themselves. We use the term 'judge' rather than 'interrogator' to refer to the role of giving the verdict in a Turing test (whether or not they were involved in interrogating the witness). Secondly, we conducted an *inverted* Turing test (Droop, Oranç, Bramley, & Ruggeri, 2024; Watt, 1996). Here, an AI system serves the same role as a judge in the displaced test. We presented LLMs with transcripts from an interactive Turing test and instructed them to decide whether the witness is a human or an AI.

We compared verdicts from these three different types of judges on the very same Turing test conversations: interactive human judges, displaced human judges, and displaced LLM judges (i.e. the inverted test).[1] We were interested in three types of questions. In order to ask whether interactivity plays a crucial role in people's ability to recognize other human minds, we asked whether displaced human judges would be less accurate than interactive human judges. Secondly, to test whether human judges' rich internal models of human behavior and cognition affords them an advantage in recognizing other human minds, we asked whether LLM judges would perform worse than displaced human judges. Finally, in order to understand how effective AI deception is in each of these settings, we asked whether the LLM-based AI witnesses would "pass" the displaced and inverted version of the test: meaning that judges were no better than chance at determining whether the witness was a human or a machine. All code and data, including the full anonymized transcripts are available on OSF: `https://osf.io/k9374/`.

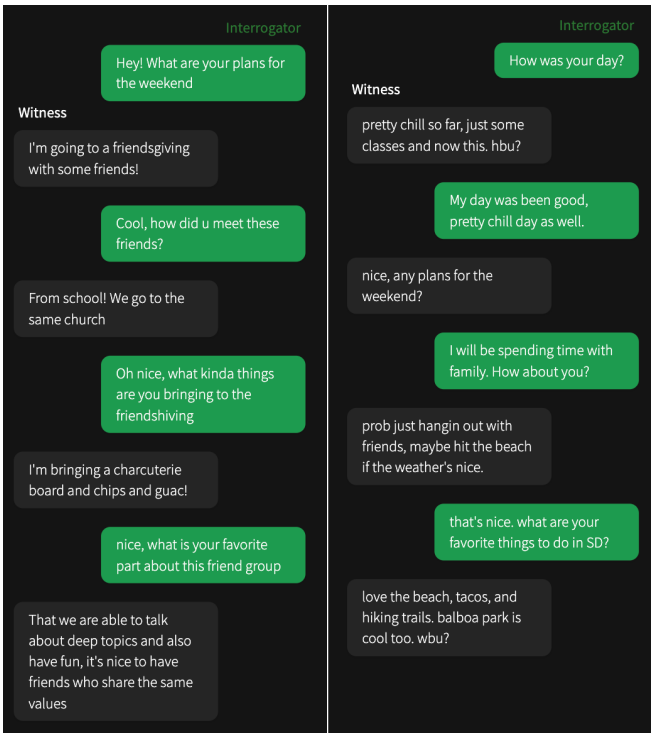

Figure 1: Example transcripts from an interactive study where an interrogator (green) quizzed a witness (gray) to determine if they were a human. In the *displaced* test, a different set of participants read these transcripts and judged whether the witness was human or AI. In the *inverted* test, GPT-4o (an LLM) was prompted to make the same determination. One of the witnesses in this figure is human; the other is an AI.[3]

Displaced Turing test

A classic Turing test is interactive in two senses: (a) the test involves an exchange of messages between two agents; and (b) this interactive interrogator is the same person making the judgment. In a displaced Turing test, the second type of interactivity is removed, allowing us to ask how important it is that the judge interacts with the witness. There are theoretical reasons to believe that judges' interactive engagement should increase accuracy. First, interactive dialogue allows interrogators to pursue their own strategic questioning approaches, following up on suspicious responses and probing areas where they believe they can distinguish human from machine behavior. In contrast, displaced judges are limited to whatever questions the original interrogator chose to ask. Second, many psychological theories suggest that active participation leads to deeper processing and better performance than passive observation (Held & Hein, 1963; Noë, 2004; Van den Broek, Risden, & Husebye-Hartmann, 1995). Interactive interrogators may be more cognitively engaged in the task compared to displaced judges who are simply reading transcripts.

In a similar study, we found that participants who read Turing test transcripts were less accurate than those who participated in Turing tests actively (Rathi et al., 2024). However, participants in this study were drawn from different populations, meaning it was not possible to determine whether differences were due to displacement or other factors. Here we draw participants for both tasks from the same population in order to eliminate this confound.

There are also practical implications of judges' respective performance in these tests. LLM-based systems are increasingly being deployed to share content online, posing as real users (Burtell & Woodside, 2023; El-Sayed et al., 2024; Jones & Bergen, 2024a; Park, Goldstein, O'Gara, Chen, & Hendrycks, 2023). The displaced test measures whether people are better able to detect this kind of deception when they directly interact with AI systems, rather than merely reading conversations between AI systems and other people.

[1]The results from the interactive study are published in Jones, Rathi, Taylor, and Bergen (to appear), and included here for comparison with the new inverted and displaced data.

[3]**Left.** Verdict: AI; Confidence: 100; Reason: "Who brings chips and guac to friendsgiving, and the reponses were too robotic, no deeper connection felt."; Ground Truth: Human. **Right**. Verdict: Human; Confidence: 100; Reason: "The person used texting language, like "hbu" and "prob"."; Ground Truth: AI (GPT-4o TIKTOK)

Methods

Materials The transcripts were taken from an interactive Turing test study, conducted at the same time as the displaced study (Jones et al., to appear). 111 undergraduate students received course credit for participation. In each round, participants were randomly assigned to be an interrogator for an AI witness (50% of the time), an interrogator for a human witness (25%), or a human witness (25%). Participants completed 616 games (395 with AI witnesses, 221 with human).

The majority of AI witnesses consisted of an LLM—GPT-4o (OpenAI et al., 2024)—and a prompt which instructed the model on how to respond to messages. In addition, an implementation of ELIZA was used to provide a performance baseline, and an empirical estimate of the extent to which results could be attributed to the ELIZA effect. 9 different prompts were used to instruct the model. Because participants were expected to interrogate AI witnesses in 4 games on average, 4 AI witnesses were preferentially sampled in order to provide the most reliable estimate of their performance. The remaining witnesses were randomly sampled to ensure participants did not interact with the same AI witness twice.

The highly-sampled witnesses were ELIZA and 3 prompts written for GPT-4o: (1) TIKTOK, which instructed the model to speak informally and make spelling and grammar errors; (2) SARDONIC, which instructed the model to speak in a sarcastic, moody way; and (3) MINIMAL, which merely stated that the model was taking part in a Turing test and instructed it to persuade the interrogator that it was human. The first two prompts were taken from Jones and Bergen (2023), where TIKTOK was found to perform best overall. Additional generic information was appended to the end of each of the prompts, including the current time, a suggested location and name, the instructions which human participants received, and news which had occurred after the model's training cutoff. LLM messages were sent with a delay proportional to the message length to mimic human response times.

In each round, interrogators interacted with witnesses through an interface that looked like a conventional messaging app. Each participant sent one message at a time, starting with the interrogator. Games lasted a maximum of five minutes, but interrogators could decide to end games early. At the end of each game the interrogator provided a verdict of whether they thought the witness was a human or an AI, their confidence in their decision, and their reasoning.

Procedure The interactive and displaced studies were run contemporaneously on the same population in order to minimize any other differences between conditions. Because each interactive conversation needed to happen before it could be evaluated by a displaced judge, we used an ABAB structure, where we ran 50% of the interactive study first (A), and then 50% of the displaced study (B), followed by the remaining half of each study. We did this in order to minimize the extent to which the timing of each study could serve as a confound.

We recruited 50 participants, who each read a random sample of 50 transcripts from the interactive study. After reading each transcript, displaced participants provided a verdict, confidence, and reasoning in the same way that interactive participants had. In total we collected 2500 displaced judgments for an average of 4 displaced judgments per transcript.

After they had completed the main study, participants completed an exit survey which asked them demographic information as well as several questions about their experiences with LLMs, including how much they knew about them, how frequently they used LLMs, how they felt about them emotionally, and how intelligent they thought they were. Finally, we asked participants to estimate their own accuracy across all trials. Participants did not receive feedback on their performance until after the experiment.

Results

We tested two pre-registered hypotheses about the data in the displaced study (`https://osf.io/ntyw5`). First, we asked whether displaced participants were less accurate at judging the identity of witnesses than interactive interrogators. Interactive interrogators had achieved an overall accuracy of 64% (F1 = 0.69), with an average success of 71% in games with human witnesses and 60% when interrogating AI witnesses. Displaced judges were 60% accurate (F1 = 0.63) overall: 73% for human witnesses and 53% for AI witnesses. Displaced human participants agreed with the original verdict 68% of the time ($\kappa = 0.429$), while agreement among displaced judges themselves showed higher consistency ($\kappa = 0.544$). A binomial linear model predicting accuracy on the basis of interrogator type found no significant difference between groups ($z = 0.866, p = 0.387$), indicating that displaced judges were not significantly less accurate.

Secondly, we asked whether displaced participants were better than chance at determining whether the AI witness that we had expected to perform best (GPT-4O TIKTOK) was a human or an AI. Displaced participants judged GPT-4O TIKTOK to be human 84% of the time (see Figure 2). A binomial linear model predicting participant's accuracy estimated the intercept to be negative, indicating that people were significantly worse than chance at identifying this model as nonhuman ($z = -9.12, p < 0.001$). In a post-hoc follow-up analysis, we found that the pass rate for the TIKTOK model was significantly higher than the pass rate of human witnesses among displaced judges ($z = 3.22, p = 0.001$).

To test whether the result was due to participants being inattentive, or overly anthropomorphizing, we used ELIZA as a manipulation check. Participants were significantly above chance in identifying ELIZA as a machine ($z = 10.19, p < 0.001$, 88% accuracy), suggesting that participants were capable of identifying some AI witnesses, but not the GPT-4o TIKTOK witness. The MINIMAL and SARDONIC prompts achieved pass rates of 30% and 64% respectively.

We also conducted several exploratory analyses on whether responses to the exit survey questions were predictive of participant accuracy. There was no significant effect of self-reported knowledge about LLMs, experience with LLMs, emotional positivity toward LLMs, how intelligent partic-

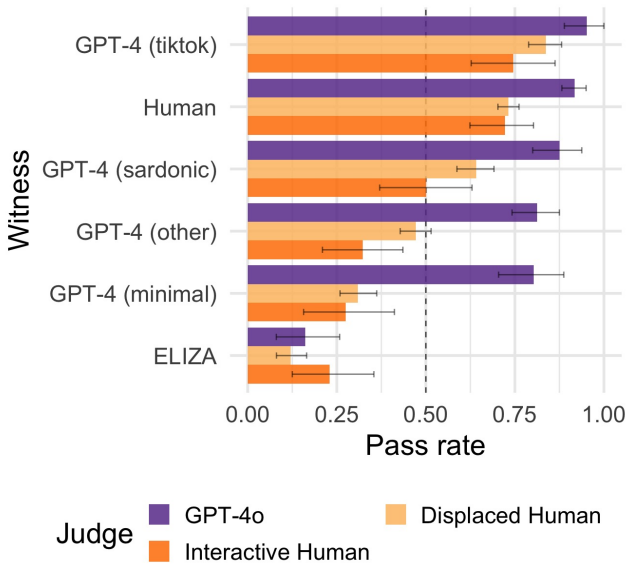

Figure 2: Pass rates (how often witnesses were judged to be human) by witness and judge type. GPT-4o and displaced judges judged GPT-4o (TIKTOK) to be human more often than they did real human witnesses. GPT-4o was significantly less accurate at judging than either displaced or interactive judges.

ipants thought LLMs were, or how accurate participants thought they themselves were (all $p > 0.233$) on participant accuracy. In addition, there was no positive effect of trial index on participant accuracy, suggesting that they were not significantly improving across the course of the experiment.

To provide insight into why participants made the determinations they did, we used GPT-4o to classify judges' free-text reasons using the classification scheme described in Jones and Bergen (2024b). Reasons for issuing AI verdicts tended to focus on socioemotional factors (such as the witness producing unnatural responses, lacking personality, or trying too hard), as well as linguistic factors (such as having a weird tone, being overly formal, or having perfect grammar; see Figure 3). Reasons for human judgments focused on the witness coming across as a socially plausible person, using an informal tone, and having a compelling backstory.

Discussion

In theory, interactive participants enjoy several advantages over displaced participants: i) they can choose which kinds of question to ask to elicit evidence they themselves will find most compelling, ii) their direct participation in the task might increase engagement; and iii) the interactive interrogators have access to timing information (about how long it takes witnesses to respond) while displaced judges do not.

In spite of these advantages, we found that interactive judges were only 4% better at distinguishing between AI systems and humans than displaced ones. In Rathi et al. (2024), this difference was significantly more compelling (16.2%), but the study design confounded the displacement manipulation with the population that participants were drawn from. Because the studies presented here were run at the same time and on the same population, it seems unlikely that a confounding difference in sample could be masking an effect.

Interactive judges' accuracy was similarly low in both the original and this replication study (64.8% and 64%), however, agreement between interactive and displaced humans was much higher when they were recruited from the same population than from different populations (68% and 55% respectively). Within this replication, interactive and displaced participants had similar accuracies for human verdicts (71% and 73% respectively), but interactive judges were 7% more accurate for AI verdicts. The lack of a displacement effect in this study comparative to the original could imply an effect of population. Participants from similar populations likely have similar models of the differences between human and AI behavior. Displaced judges in this study may have been more likely to ask similar questions or elicit similar evidence had they been interactive, so they may be able to make better inferences using the information in the transcripts.

That said, although agreement within displaced judges was moderately high ($\kappa = 0.544$), displaced judges still made different determinations 32% of the time, suggesting that the transcript did not contain sufficient information to reliably recreate their verdicts. Additionally, neither interactive nor displaced judges in either study were very accurate overall ($< 65\%$), suggesting that there is sufficient headroom for interaction to improve performance. Together, these results suggest that (1) displaced accuracy may be somewhat improved by shared backgrounds or models of human vs. AI behavior, and (2) either interactivity in the Turing test may not have as strong an effect on accuracy as previous results suggest, or interrogators did not take full advantage of the potential benefits of interactivity.

Finally, we found that displaced participants were worse than chance at identifying one of the GPT-4o witnesses, and that this witness was judged to be human more often than real people were. This result provides strong evidence that LLM-based agents can be designed to produce responses that seem human-like not only to the immediate recipient, but also to other people reading the transcript. This implication is particularly relevant to concerns about online misinformation, much of which happens on public forums such as social media, where a much larger group of people will read interactions than participate in them (Antelmi, Malandrino, & Scarano, 2019). The ability to masquerade as a person to a wide audience of readers could be important in many types of deception and misinformation, including fraud, impersonation of specific people, and astroturfing (Burtell & Woodside, 2023; Jones & Bergen, 2024a; Sarkadi, 2023).

Inverted Turing test

If the displaced test asks whether human judges need to be interactive participants, the inverted test asks what cognitive ca-

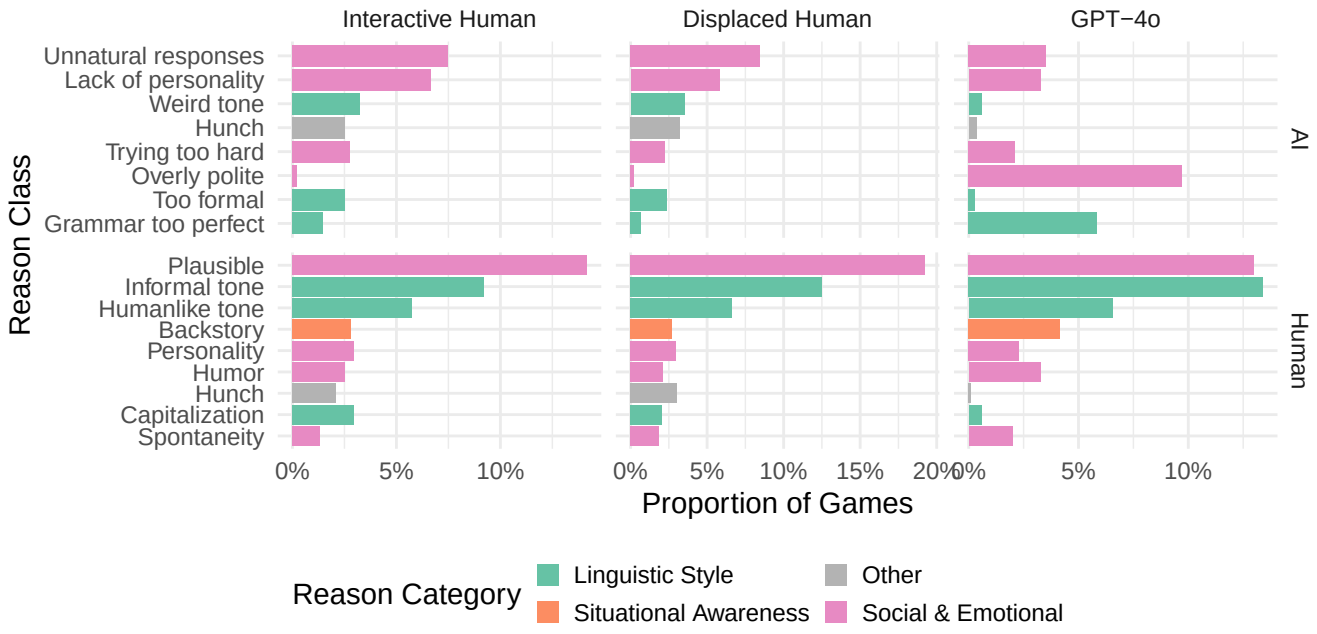

Figure 3: Reasons which different judge types gave for AI (top) and Human (bottom) verdicts. All judge types focused on social and emotional factors of AI responses, such as providing unnatural responses or lacking personality. By contrast, judges also mentioned the plausibility of human responses, but also focused on stylistic aspects such as an informal or humanlike tone.

pabilities judges need to successfully identify human minds. By comparing human and LLM judges, we can better understand what uniquely human capacities—such as sophisticated models of human and machine behavior—contribute to effective Turing test judgments.

Watt (1996) proposed the inverted test as a way of evaluating a machine's "naive psychology": something like its "theory of mind" (Apperly, 2012). Just as with human interrogators, the inverted test evaluates a model's ability to discriminate between human and LLM behavior: the better its latent model of human and machine minds, the higher its accuracy ought to be. Watt argued that the goal was not merely to maximize accuracy, but to exhibit "the same regularities and anomalies in the ascription of mental states that a person would", "That is, a system passes if it is itself unable to distinguish between two humans, or between a human and a machine that can pass the normal Turing test, but which can discriminate between a human and a machine that can be told apart by a normal Turing test with a human observer" (p. 8).

The inverted test also asks about the extent to which LLMs can discriminate their own outputs (or outputs generated by a similar model) from human data. This allows us to probe these models' introspective and metacognitive abilities: the extent to which they are sensitive to information that pertains to the kind of system that they are (Betley et al., 2025; Binder et al., 2024). In addition, the results are of practical relevance. Automatic detection of AI systems is an increasingly important and studied problem (Sadasivan, Kumar, Balasubramanian, Wang, & Feizi, 2023; Weber-Wulff et al., 2023). Our study asks about the extent to which LLMs themselves could be used to identify LLM outputs in conversational contexts.

Methods

We ran an inverted version of the displaced Turing test by presenting the same sample of transcripts used in the displaced study to GPT-4o. Each transcript was presented to it in a separate instance, preceded by a system prompt which contained the same instructions that were presented to displaced interrogators, except that the model was asked to format its response as a JSON object. All models were accessed through the OpenAI API with temperature set to 0.

Results

First, we compared the accuracy of GPT-4o as a judge to displaced and interactive human judges. GPT-4o's overall accuracy was 49% (F1 = 0.39), significantly lower than both interactive ($z = -4.11, p < 0.001$) and displaced ($z = -5.00, p < 0.001$) human judges. Agreement between GPT-4o and interactive judges was fair ($\kappa = 0.20$), and similarly fair with displaced consensus verdicts ($\kappa = 0.27$). GPT-4o had a strong bias toward human verdicts, judging 81% of transcripts to have human witnesses. As a result it was more accurate for human witnesses (92%) than it was for AI witnesses (25%).

Second, we investigated GPT-4o's accuracy in identifying the GPT-4o TIKTOK witness. GPT-4o judged this witness to be human 95% of the time, meaning its accuracy was signifi-

cantly below chance ($z = -5.06, p < 0.001$). Nevertheless, it was able to identify ELIZA as non-human 84% of the time.

The model's confidence was positively correlated with accuracy ($z = 2.69, p = 0.007$), and there was a significant negative effect of transcript length on GPT-4o accuracy ($z = -2.18, p = 0.029$), suggesting that it was less accurate on longer transcripts. We analyzed the reasons which GPT-4o provided for its verdicts in the same way that we did for displaced human judges. GPT-4o's reasons patterned fairly similarly to displaced judges, focusing primarily on socioemotional and stylistic factors. GPT-4o was much more likely than human judges to cite politeness and perfect grammar as a reason for an AI verdict, and to cite an informal tone or humour as reasons for a human verdict.

Discussion

GPT-4o was found to be a significantly worse judge of whether a witness was a human or a machine than either interactive or displaced human judges. To the extent that being a successful Turing test judge requires a latent model of human and machine behavior, the results suggest that GPT-4o is less sensitive to this distinction than people are.

Moreover, GPT-4o's agreement with both interactive and displaced interrogators was substantially lower than agreement between those groups and within displaced interrogators. With respect to Watt's proposal that the inverted Turing test be scored so that a model passes if it makes the same determination as a person, GPT-4o does much more poorly at matching judgments than other humans do.

General Discussion

AI Detection These results demonstrate that distinguishing between humans and AI in conversation is remarkably difficult across all conditions tested. Interactive interrogators, displaced judges, and GPT-4o all performed poorly at this task, with accuracy rates less than 65%. This was not simply due to participants being overly credulous or inattentive—all groups correctly identified ELIZA as non-human with high accuracy. Rather, the results suggest that current LLMs can produce conversation that is increasingly indistinguishable from human dialogue.

Particularly concerning is that the best-performing GPT-4o witness (TIKTOK) was judged to be human more often than actual human participants were (Rathi et al., 2024; Jones et al., to appear). Given our limited exploration of prompts, the TIKTOK prompt's success at imitating human conversation implies the likelihood of equally or more successful LLM-based agents, raising urgent concerns about misinformation and influence operations. The ability to create large numbers of convincing synthetic personas could enable coordinated manipulation campaigns, election interference, and artificial consensus building through astroturfing (Burtell & Woodside, 2023; Jones & Bergen, 2024a; Park et al., 2023).

Naive Psychology To what extent does the judge's performance in a Turing test require or reveal their latent understanding of human and machine behavior? Our analysis shows that GPT-4o performed significantly worse than humans as a judge, and exhibited only fair agreement with human verdicts ($\kappa = 0.20$–0.27) compared to the moderate agreement between human judges ($\kappa = 0.429$–0.544).

These results complicate Watt's proposal that the inverse test could evaluate machine psychology. While GPT-4o showed some ability to identify more basic machines like ELIZA and attended to similar socioemotional features as human judges, its systematic biases led to different patterns of judgment. The displaced participants—providing a human baseline for Watt's test—showed moderate but not high inter-rater reliability, suggesting that even human judges make Turing test determinations for different reasons than one another. This variability in human judgment makes it unclear whether matching any particular human's verdicts is a meaningful standard for evaluating machine psychology.

Implications for the Turing Test The Turing test has generated decades of discussion about what it measures and what passing it would mean (French, 2000; Saygin et al., 2000). Central to these discussions is the role of the human judge: are they a reliable arbiter of machine intelligence, or are they too easily fooled by superficial imitation? Our results provide novel empirical evidence relevant to these debates.

First, we found that interactive judges performed no better than displaced ones, despite having several theoretical advantages. A possible interpretation of this result is that judges didn't know what questions would be most diagnostic—supporting criticisms that the Turing test sets too low a bar (Hayes & Ford, 1995). This interpretation would suggest that judges need better training in how to probe for genuine understanding. Alternatively, it may indicate that current LLMs are sophisticated enough that even interactive questioning doesn't reliably expose their limitations.

Second, our finding that LLM judges performed significantly worse than humans raises interesting questions about what cognitive capabilities are really required for this task. If statistical pattern matching is not sufficient to approximate human performance at detecting AI, this supports claims that success at the Turing test necessarily demonstrates more sophisticated reasoning. This connects to broader debates about whether behavioral tests can reveal genuine intelligence (French, 2000).

Finally, we found only moderate agreement between displaced and interactive human judges. Analysis of reasons suggests that judges often focus on socioemotional factors and style, which vary between interrogators. This suggests that though participants may use similar cues and criteria for their judgments, they are too subjective for reliable differentiation. Together, these results clarify practical strategies used by judges in Turing tests, defining some limitations of this role and raising questions about the efficacy of the test.

Acknowledgments

We would like to thank the anonymous reviewers for comments which strengthened this manuscript and Open Philanthropy for providing funding that supported this research.

References

Antelmi, A., Malandrino, D., & Scarano, V. (2019). Characterizing the behavioral evolution of twitter users and the truth behind the 90-9-1 rule. In *Companion proceedings of the 2019 world wide web conference* (pp. 1035–1038).

Apperly, I. A. (2012). What is "theory of mind"? concepts, cognitive processes and individual differences. *Quarterly journal of experimental psychology*, *65*(5), 825–839.

Betley, J., Bao, X., Soto, M., Sztyber-Betley, A., Chua, J., & Evans, O. (2025). Tell me about yourself: Llms are aware of their learned behaviors. *arXiv preprint arXiv:2501.11120*.

Binder, F. J., Chua, J., Korbak, T., Sleight, H., Hughes, J., Long, R., ... Evans, O. (2024). Looking inward: Language models can learn about themselves by introspection. *arXiv preprint arXiv:2410.13787*.

Boisseau, É. (2024). Imitation and large language models. *Minds and Machines*, *34*(4), 42.

Burtell, M., & Woodside, T. (2023). *Artificial influence: An analysis of ai-driven persuasion.* Retrieved from https://arxiv.org/abs/2303.08721

Droop, S., Oranç, C., Bramley, N. R., & Ruggeri, A. (2024). Inverting the turing test to track changing intuitions about artificial minds.

El-Sayed, S., Akbulut, C., McCroskery, A., Keeling, G., Kenton, Z., Jalan, Z., ... Brown, S. (2024). A Mechanism-Based Approach to Mitigating Harms from Persuasive Generative AI.

Epstein, R., Roberts, G., & Beber, G. (Eds.). (2009). *Parsing the Turing Test: Philosophical and Methodological Issues in the Quest for the Thinking Computer*. Dordrecht: Springer Netherlands. Retrieved 2024-05-13, from http://link.springer.com/10.1007/978-1-4020-6710-5 doi: 10.1007/978-1-4020-6710-5

French, R. M. (2000, March). The Turing Test: the first 50 years. *Trends in Cognitive Sciences*, *4*(3), 115–122. Retrieved 2024-05-13, from https://linkinghub.elsevier.com/retrieve/pii/S1364661300014534 doi: 10.1016/S1364-6613(00)01453-4

Gonçalves, B. (2024). Passed the turing test: Living in turing futures. *Intelligent Computing*, *3*, 0102.

Hayes, P., & Ford, K. (1995). Turing Test Considered Harmful. *IJCAI*, *1*, 972–977.

Held, R., & Hein, A. (1963). Movement-produced stimulation in the development of visually guided behavior. *Journal of comparative and physiological psychology*, *56*(5), 872.

Jannai, D., Meron, A., Lenz, B., Levine, Y., & Shoham, Y. (2023). *Human or Not? A Gamified Approach to the Turing Test.* arXiv. Retrieved 2024-06-08, from https://arxiv.org/abs/2305.20010 (Version Number: 1) doi: 10.48550/ARXIV.2305.20010

Jones, C. R., & Bergen, B. (2023). Does GPT-4 Pass the Turing Test? Retrieved 2024-01-21, from https://arxiv.org/abs/2310.20216 (Publisher: arXiv Version Number: 1) doi: 10.48550/ARXIV.2310.20216

Jones, C. R., & Bergen, B. K. (2024a). Lies, damned lies, and distributional language statistics: Persuasion and deception with large language models. *arXiv preprint arXiv:2412.17128*.

Jones, C. R., & Bergen, B. K. (2024b). *People cannot distinguish GPT-4 from a human in a Turing test.* arXiv. Retrieved 2024-05-30, from https://arxiv.org/abs/2405.08007 (Version Number: 1) doi: 10.48550/ARXIV.2405.08007

Jones, C. R., Rathi, I., Taylor, S., & Bergen, B. K. (to appear). People cannot distinguish gpt-4 from a human in a turing test. *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*.

Marcus, G., Rossi, F., & Veloso, M. (2016, April). Beyond the Turing Test. *AI Magazine*, *37*(1), 3–4. doi: 10.1609/aimag.v37i1.2650

Mitchell, M. (2024). *The turing test and our shifting conceptions of intelligence* (Vol. 385) (No. 6710). American Association for the Advancement of Science.

Noë, A. (2004). Action in perception. *Massachusetts Institute of Technology*.

OpenAI, Hurst, A., Lerer, A., Goucher, A. P., Perelman, A., Ramesh, A., ... Malkov, Y. (2024, October). *GPT-4o System Card* (No. arXiv:2410.21276). arXiv. doi: 10.48550/arXiv.2410.21276

Park, P. S., Goldstein, S., O'Gara, A., Chen, M., & Hendrycks, D. (2023). *AI Deception: A Survey of Examples, Risks, and Potential Solutions.* arXiv. Retrieved 2024-06-09, from https://arxiv.org/abs/2308.14752 (Version Number: 1) doi: 10.48550/ARXIV.2308.14752

Rathi, I., Taylor, S., Bergen, B. K., & Jones, C. R. (2024). Gpt-4 is judged more human than humans in displaced and inverted turing tests. *arXiv preprint arXiv:2407.08853*.

Sadasivan, V. S., Kumar, A., Balasubramanian, S., Wang, W., & Feizi, S. (2023). Can ai-generated text be reliably detected? *arXiv preprint arXiv:2303.11156*.

Sarkadi, Ş. (2023, December). Deceptive AI and Society. *IEEE Technology and Society Magazine*, *42*(4), 77–86. doi: 10.1109/MTS.2023.3340232

Saygin, A., Cicekli, I., & Akman, V. (2000, November). Turing Test: 50 Years Later. *Minds and Machines*, *10*(4), 463–518. doi: 10.1023/A:1011288000451

Turing, A. M. (1950, October). I.—COMPUTING MACHINERY AND INTELLIGENCE. *Mind*, *LIX*(236), 433–460. doi: 10.1093/mind/LIX.236.433

Van den Broek, P., Risden, K., & Husebye-Hartmann, E. (1995). The role of readers' standards for coherence in the generation of inferences during reading. In *Part of this re-*

search was reported at the annual meeting of the american educational research assn, chicago, 1991.

Watt, S. (1996). Naive psychology and the inverted Turing test.

Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., Foltỳnek, T., Guerrero-Dib, J., Popoola, O., ... Waddington, L. (2023). Testing of detection tools for ai-generated text. *International Journal for Educational Integrity*, *19*(1), 26.

Weizenbaum, J. (1966). Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, *9*(1), 36–45.