Section 1 (Introduction)

This study explores whether humans can distinguish between real people and AI systems during short, interactive conversations using a Turing test format. With the rapid advancement of large language models (LLMs) like GPT-4, which perform humanlike communication, the risks of AI impersonation have increased. These risks include fraud, misinformation, manipulation, and loss of trust in digital interactions. Interactive dialogues, where participants actively try to detect AI impersonation, provide a high-stakes environment to evaluate these risks. This research aims to assess the success of AI systems in passing the Turing test and to understand users' detection strategies, perceptions of human uniqueness, and potential vulnerabilities.

Historically, many attempts have been made to run Turing tests, but no machine has consistently passed in controlled experiments. Previous studies using gamified online settings showed models like GPT-4 could achieve a 50% pass rate. However, these studies lacked controls, did not publish data, and used convenience samples. To address these limitations, the authors conducted a randomized, controlled two-party Turing test. The study pre-registered its hypotheses and used a defined, compensated participant pool. Initial experiments with Prolific workers showed GPT-4 reached a 50% pass rate, indicating performance at chance level. A replication study with undergraduates showed an even higher pass rate of 77%, slightly surpassing the human average of 71%.

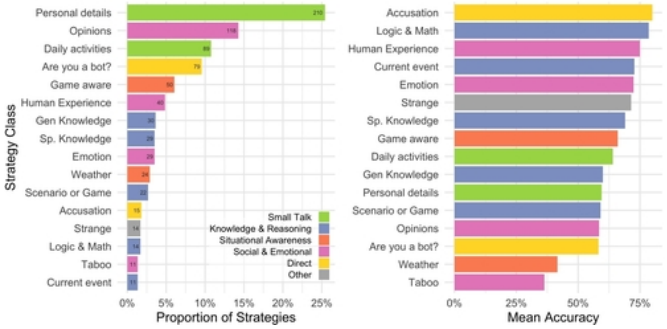
Section 2 (Methods)

The study evaluated 3 different types of AI Witness: gpt-4-0613, gpt-3.5-turbo-0125, and ELIZA on their ability to convince testing candidates they are human. Before the study commenced a Prompt was devised through testing to ensure the model returned the desired behaviour for the study. The prompt instructed the model to adopt the persona of a young person who was very concise, didn’t take the game very seriously, used slang, and made occasional spelling errors. At the start of each game, several additional pieces of information were appended to the prompt, including a suggested name, location, languages spoken, and timestamps for the model.

To make the models more convincing, messages were sent with a randomized delay to imitate typing speed and thinking time. The central challenge for interpreting the Turing test is the ELIZA effect: a phenomenon in which participants attribute human characteristics to even very simple systems. Thus, ELIZA is included as a baseline in the experiment, both to empirically quantify the ELIZA effect, and to serve as a manipulation check in the experiment. The game interface was built like a messaging app and both the testers and AI were given a 300 character limit. The study aimed for a sample of 100 games for each of our 4 conditions (Human, GPT-4, GPT-3.5, ELIZA). 32 participants were excluded. Of the 402 interrogators, 168 were female, 212 male, 9 nonbinary, and 13 preferred not to say. Participants’ mean age was 35.

Section 3 (Results)

This section reports results from a Turing test–style study comparing GPT-4, GPT-3.5, ELIZA, and human participants. Pass rate, how often interrogators judged a witness as human, was highest for humans (67%), followed by GPT-4 (54%), GPT-3.5 (50%), and ELIZA (22%). Statistical analyses confirmed GPT-4's performance was significantly above ELIZA and below humans, but not significantly different from GPT-3.5. Participant strategies often involved small talk, social or emotional questions, and knowledge or reasoning tasks, with logic- and experience-based questions yielding higher accuracy.



Reasons for AI verdicts frequently cited unnatural tone, forced persona, or lack of personality, while human verdicts emphasized plausibility and informal tone. A conceptual replication using GPT-4o expanded the study with new prompts and repeated rounds to assess robustness and demographic generalizability, with full anonymized data made available for future research.

Section 4 (Replication)

The study was replicated due to the original experiment's inconsistency and lack of transparency, particularly regarding the availability of conversation transcripts. Additionally, the researchers sought to evaluate the responses of a different population group. Rather than building on the flawed structure of the original study, the new experiment was designed from the ground up. Notably, it employed GPT-4o instead of GPT-4 and introduced multiple rounds of interaction to assess whether participants might behave in ways that could raise suspicion over time. Key changes were made to improve upon the original methodology. Instead of recruiting participants from Prolific.com, the new study involved undergraduate students. The AI witnesses were diversified: rather than relying on a single prompt, the study used nine distinct prompts, each instructing the AI to roleplay as an undergraduate student. This variety was intended to prevent participants from recognizing the AI based on repeated patterns.

Each of the 111 participants took part in eight rounds: four rounds as an interrogator speaking with AI, two as an interrogator speaking with another human, and two rounds as a witness responding to a human interrogator. Participants were never paired with the same human or AI partner more than once. AI interrogators were drawn from a pool of nine models, with the four primary prompts being used first, followed by the rest. The four main AI prompts tested were: TIKTOK (used in the previous study), Minimal (a stripped-down prompt simply instructing the AI to act human), Sardonic (instructing the AI to adopt a sardonic

tone), and ELIZA (serving as the non-AI baseline). Three hypotheses were tested, all concerning the TIKTOK prompt.

First, it was hypothesized that the TIKTOK model would pass at a rate not significantly below 50%; in reality, it passed at approximately 76%, which is well above the threshold for success by chance. Second, TIKTOK was expected to outperform ELIZA significantly, this was confirmed. The third hypothesis posited that TIKTOK would significantly outperform humans; however, with humans achieving a 71% pass rate, this was not conclusively true, as the overlap in confidence intervals meant human performance may have matched or exceeded TIKTOK's. Other findings from the experiment included that the Minimal prompt failed badly, with results significantly worse than random chance, much like ELIZA. Sardonic achieved an average pass rate, neither excelling nor failing. Interestingly, one lower-priority prompt, QINN, attained an impressive 86% pass rate; however, it was only tested in 15 instances and thus requires further validation before any firm conclusions can be drawn.

Section 5 (Discussion)

The discussion section talks about the results of the Turing tests, particularly whether GPT-4 (and later GPT-4o) can be considered to have “passed” it. The authors argue that while Turing’s original 70% chance of making the right identification after five minutes is an arbitrary benchmark, a 50% baseline indicating performance at chance level is more justified for determining if interrogators can reliably identify machines. Under this 50% criteria, and considering ELIZA as a baseline manipulation check (where its pass rate was significantly below chance), the original study’s GPT-4 should be considered pass. The replication of the study done with GPT-4o further proved this with a pass rate of 76%, which even surpassed the original human pass rate of 71%, indicating a “parity” with human performance in this specific setup. The discussion also touches upon the surprising finding that humans were only identified as human 67% of the time in the main study, suggesting a shift in assumptions about AI capabilities might lead the interrogators to misidentify humans as AI.

The paper then leads the discussion into the troubling implications of AI’s ability to deceive. Being mistaken for an AI can be “jarring or even dehumanising” for individuals, and on a broader societal level, it can even threaten mutual trust, transparency and accountability in online interactions like social media, wikipedia and others. The authors also re-evaluate what the Turing test measures, suggesting that current AI systems success hinges more on imitating “linguistic style and socio emotional factors” rather than traditional notions of intelligence. The “ELIZA effect” attributing human qualities to simple programs is examined, with the study’s results indicating that LLMs significantly outperform ELIZA, but also that ELIZA perceived “uncooperativeness” can be misinterpreted as human. Finally the discussion proposes mitigation strategies against AI deception, such as statistical watermarks, digital identity authorisation, and disclosure of data provenance, to prevent a future where users are forced to doubt the authenticity of real people online.

I think that the Turing test does require a re-evaluation to fit into the modern environment of what intelligence is and does LLM’s count as intelligence or just really good at sounding like humans. So I believe the Turing test needs to be updated or a new way of evaluating machines intelligent apart from just can deceive a human into thinking it is a human. One thing that they did not account for is that even more modern LLM’s like GPT-4o are

multimodal meaning they can accept multiple media as inputs like voice, images or recordings. And produce outputs in them as well. So I think the Turing test could include speech input, image analysis and video recording. This can give humans more of a chance of proving they are human and harder for LLM's to replicate at this moment in time. But as shown in the research paper the participants only interacted with the models and humans through text, this could change to include voice and images. And another thing mentioned was, because these models can deceive humans into thinking they are human there are a lot of implications of deepfakes and scams which can occur through online platforms or text messages.

One issue drawn with the methodology is the uneven distribution of interrogation time between AI (4) and Human (2). Thus, due to humans be under represented in testing any given interrogation is more likely to be an AI, this might either cause AI to be over valued as human because it might be assumed that the ratiois 50% AI and 50% humans thus also undervaluing humans as it does not take into account the participants recollections of previous interrogations. However if the participants are informed of the ratio they are more likely to be more critical of both humans and AI, assuming due to the higher likelihood of being an AI. The 400 participants were US undergraduate students. This sample is being generalised to People as a whole.

GPT-4 passed as human more than chance and far outperformed ELIZA but despite this it still fell notably short of human participants indicating current LLMs are improving but remain distinguishable in social interaction. There is a lack of significant difference between GPT-4 and 3.5 hinting there may be less progress in humanness of responses compared to other models. Different interrorgator strategies strongly influenced accuracy with questions involving logic and maths and human experiences proving more effective Many interrigators showed bias towards linguistic style and emotional queues rather than reasoning suggesting people tend to rely on those rather than deeper content. The conceptual replication using GPT-4o broadens the dataset, improves generalizability, and allows sharing transcripts, supporting transparency and future research.