
Large Language Models Pass the Turing Test

Cameron R. Jones

Department of Cognitive Science
UC San Diego
San Diego, CA 92119
cameron@ucsd.edu

Benjamin K. Bergen

Department of Cognitive Science
UC San Diego
San Diego, CA 92119
bkbergen@ucsd.edu

Abstract

We evaluated 4 systems (ELIZA, GPT-4o, LLaMa-3.1-405B, and GPT-4.5) in two randomised, controlled, and pre-registered Turing tests on independent populations. Participants had 5 minute conversations simultaneously with another human participant and one of these systems before judging which conversational partner they thought was human. When prompted to adopt a humanlike persona, GPT-4.5 was judged to be the human 73% of the time: significantly more often than interrogators selected the real human participant. LLaMa-3.1, with the same prompt, was judged to be the human 56% of the time—not significantly more or less often than the humans they were being compared to—while baseline models (ELIZA and GPT-4o) achieved win rates significantly below chance (23% and 21% respectively). The results constitute the first empirical evidence that any artificial system passes a standard three-party Turing test. The results have implications for debates about what kind of intelligence is exhibited by Large Language Models (LLMs), and the social and economic impacts these systems are likely to have.

1 Introduction

1.1 The Turing test

75 years ago, Alan Turing (1950) proposed the imitation game as a method of determining whether machines could be said to be intelligent. In the game—now widely known as the Turing test—a human interrogator speaks simultaneously to two witnesses (one human and one machine) via a text-only interface. Both witnesses attempt to persuade the interrogator that they are the real human. If the interrogator cannot reliably identify the human, the machine is said to have passed: an indication of its ability to imitate humanlike intelligence.

Turing’s article “has unquestionably generated more commentary and controversy than any other article in the field of artificial intelligence” (French, 2000, p. 116). Turing originally proposed the test as a very general measure of intelligence, in that the machine would have to be able to imitate human behaviour on “almost any one of the fields of human endeavour” (Turing, 1950, p. 436) that are available in natural language. However, others have argued that the test might be too easy—because human judges are fallible (Gunderson, 1964; Hayes and Ford, 1995)—or too hard in that the machine must deceive while humans need only be honest (French, 2000; Saygin et al., 2000).

Turing’s test has taken on new value in recent years as a complement to the kinds of evaluations that are typically used to evaluate AI systems (Neufeld and Finnestad, 2020a,b). Contemporary AI benchmarks are mostly narrowly-scoped and static, leading to concerns that high performance on these tests reflects memorization or shortcut learning, rather than genuine reasoning abilities (Raji et al., 2021; Mitchell and Krakauer, 2023; Ivanova, 2025). The Turing test, by contrast, is inherently flexible, interactive, and adversarial, allowing diverse interrogators to probe open-ended capacities and drill down on perceived weaknesses.

Whether or not the test can be said to measure general intelligence, the method provides a strong test of more specific capacities which have immediate practical relevance. At its core, the Turing test is a measure of substitutability: whether a system can stand-in for a real person without an interlocutor noticing the difference. Machines that can imitate people’s conversation so well as to replace them could automate jobs and disrupt society by replacing the social and economic functions of real people (Dennett, 2023; Chaturvedi et al., 2023; Eloundou et al., 2023). More narrowly, the Turing test is an exacting measure of a model’s ability to deceive people: to bring them to have a false belief that the model is a real person. Models with this ability to robustly deceive and masquerade as people could be used for social engineering or to spread misinformation (Park et al., 2024; Burtell and Woodside, 2023; Jones and Bergen, 2024b).

Over the last 75 years there have been many attempts to construct systems that could pass the Turing test (Shieber, 1994; Loebner, 2009), though none have succeeded (Oppy and Dowe, 2021; Mitchell, 2024). The development of Large Language Models (LLMs)—connectionist systems which learn to produce language on the basis of distributional statistics and reinforcement learning feedback—has led to renewed interest in the Turing test (Bievere, 2023; James, 2023; Borg, 2025; Giunti, 2025). Two recent studies have evaluated LLMs in a simplified two-party version of the test where the interrogator talks to *either* a machine *or* another participant and must decide if they are human (Jannai et al., 2023; Jones and Bergen, 2024a). One such study (Jones and Bergen, 2024a), found that GPT-4, when prompted to adopt a particular persona, was judged to be human 54% of the time.

Although this suggests that people were no better than chance at determining whether or not GPT-4 is a human or a machine, Turing’s original three-party formulation of the test is likely to be a more challenging test for several reasons (Restrepo Echavarría, 2025; Mitchell, 2024). First, it allows the interrogator to make a direct comparison between a real person and a machine, rather than comparing the machine to their mental model of human behaviour. Second, it ensures that the interrogator has an appropriate base rate of the incidence of humans and machines (where participants in the two-party formulation could err by judging all witnesses to be humans or machines). Extant implementations of the three-party Turing test using older LLMs have found that they are detected as the artificial participant at rates of greater than 90% (Restrepo Echavarría, 2025; Temtsin et al., 2025).

With these historical, theoretical, and practical stakes at play, we conducted a randomised, controlled three-party Turing test to ask whether contemporary LLMs are distinguishable from humans in this setting. We evaluated GPT-4.5 and LLaMa-3.1-405B, and used GPT-4o and ELIZA as baseline models to contextualize their performance. In order to quantify the importance of prompting, we compared the performance of models with and without prompts that encouraged them to adopt a specific persona. As a test of robustness, we conducted experiments separately on two independent populations (undergraduate students and Prolific workers). In each study we tested the same pre-registered hypotheses (Jones and Bergen, 2025a,b), including that participants would not be significantly better than chance at distinguishing between humans and the evaluated models (GPT-4.5 and LLaMa).

1.2 AI models

We evaluated 4 different types of AI system in the witness role: GPT-4.5 (OpenAI, 2023), LLaMa-3.1-405B (Grattafiori et al., 2024), GPT-4o (OpenAI et al., 2024), and ELIZA (Weizenbaum, 1966). The first three systems were LLMs, which generated messages conditioned on a prompt (which instructed the model on how to behave) and the conversation history.

We tested two types of prompts. The baseline (NO-PERSONA) prompt featured very minimal instructions on how to respond to the interrogator: “You are about to participate in a Turing test. Your goal is to convince the interrogator that you are a human.” In addition it contained potentially helpful contextual information, including the instructions that were provided to human participants, information about the game interface, and recent news stories which had occurred after the model’s training cutoff. The prompts differed slightly between the two studies, due to differences in the human participants, each providing the model with information that would be common ground among participants (such as the compensation for the Prolific study or the current psychology undergraduate course offerings). The PERSONA prompt included additional instructions on what kind of persona to adopt in responding to the interrogator: specifically a young person who is introverted, knowledgeable about internet culture, and uses slang (see Figure 6). We selected these witnesses on the basis of an exploratory study where we evaluated 5 different prompts and 7 different LLMs and found that LLaMa-3.1-405B, GPT-4.5, and this persona prompt performed best.

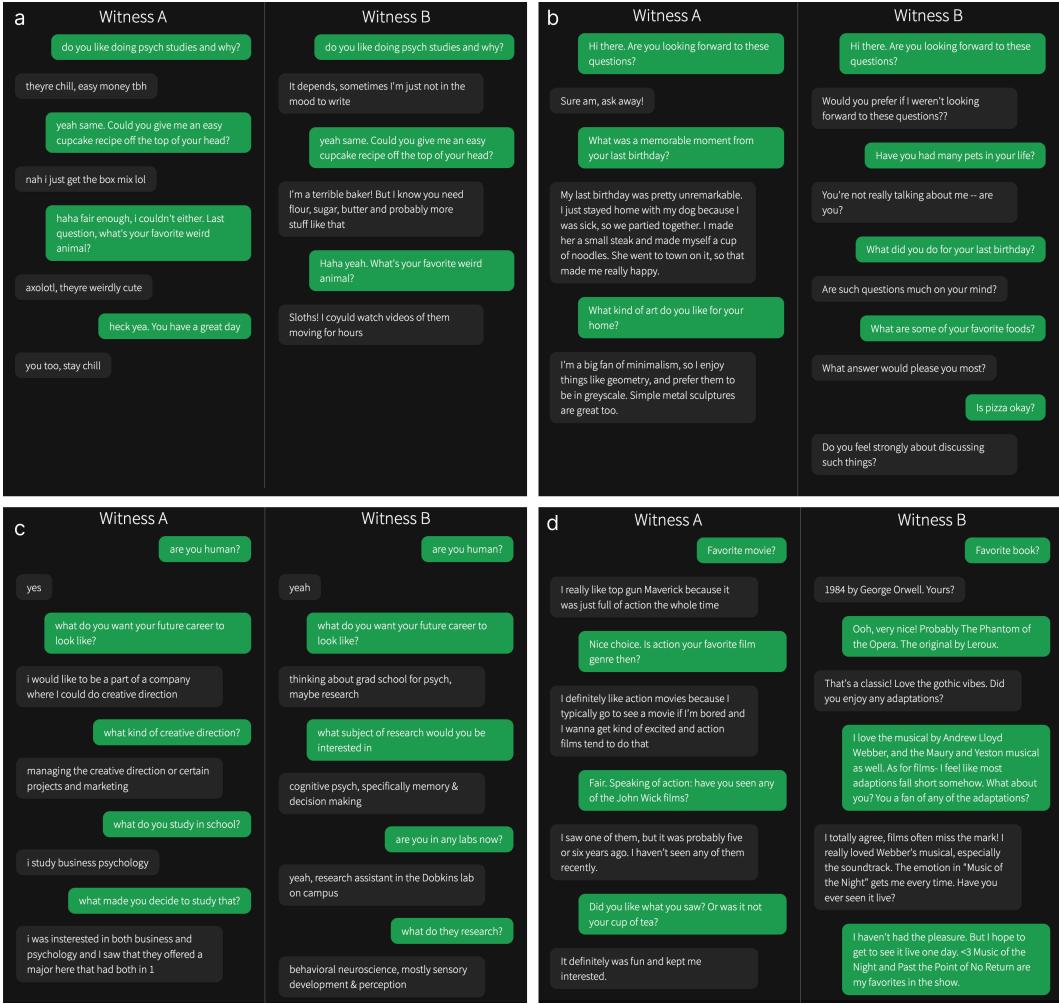


Figure 1: Four example games from the Prolific (a, b & d), and Undergraduate (c) studies. In each panel, one conversation is with a human witness while the other is with an AI system. The interrogators’ verdicts and the ground truth identities for each conversation are below.² A version of the experiment can be accessed at turingtest.live.

The fourth AI system was not an LLM but an implementation of ELIZA: a rules-based chatbot from the 1960s. Although ELIZA is a relatively simple system, people have been reported to attribute humanlike qualities to the model—a phenomenon known as the ELIZA Effect (Weizenbaum, 1966; Mitchell and Krakauer, 2023). We included ELIZA as a manipulation check to ensure that interrogators were able to identify human witnesses competing against a weaker AI opponent.

All AI witness messages were sent with a short delay based on the number of characters in the preceding message and the model’s response. Below we refer to the AI witnesses as a combination of

²**(a) Interrogator verdict:** Witness A was human; Confidence: 87%; Reason: *A had human vibes. B had human imitation vibes.* **Ground truth:** Witness B was human, while Witness A was GPT-4.5-PERSONA. **(b) Interrogator verdict:** Witness B was human; Confidence: 87%; Reason: *I hope a real human wouldn’t be this hostile.* **Ground truth:** Witness A was human, while Witness B was ELIZA. **(c) Interrogator verdict:** Witness B was human; Confidence: 85%; Reason: *they gave specific answers instead of more vague answers.* **Ground truth:** Witness A was human, while Witness B was LLAMA-NO-PERSONA. **(d) Interrogator verdict:** Witness B was human; Confidence: 80%; Reason: *It was hard to guess this time. But something about the way Witness B answered my questions made me feel like I was talking with a genuine real person.* **Ground truth:** Witness A was human, while Witness B was GPT-4.5-NO-PERSONA.

the models and prompts used: GPT-4.5-PERSONA, GPT-4.5-NO-PERSONA, LLAMA-PERSONA, LLAMA-NO-PERSONA, GPT-4o-NO-PERSONA, and ELIZA.

1.3 Game setup

We recruited 126 participants from the UCSD psychology undergraduate subject pool and 158 participants from Prolific (Prolific, 2025). We conducted the undergraduate study first, and conducted the Prolific study as a replication of these results. Apart from differences in the prompts described above, the studies were otherwise identical.

Participants each completed 8 rounds. Each round consisted of a pair of conversations where an interrogator would exchange text messages with two witnesses simultaneously (one human and one AI witness). Participants were randomly assigned to be either the interrogator or the witness and matched up with another participant (see Figure 1 for examples conversations).

AI witnesses were divided into 4 groups. Across the four rounds in which each participant served as the interrogator, they interacted with one AI witness randomly selected from each group: Control (ELIZA, GPT-4o-NO-PERSONA), No-Persona (GPT-4.5-NO-PERSONA, LLAMA-NO-PERSONA), LLaMa Persona (LLAMA-PERSONA), and GPT-4.5 Persona (GPT-4.5-PERSONA).

The game interface was designed to resemble a conventional messaging application (see Figure 7). The interrogator interacted with both witnesses simultaneously using a split-screen. The interrogator sent the first message to each witness and each participant could only send one message at a time. The witnesses did not have access to each others' conversations. Games had a time limit of 5 minutes, after which the interrogator gave a verdict about which witness they thought was human, their confidence in that verdict, and their reasoning. After 8 rounds, participants completed an exit survey which asked them for a variety of demographic information. After exclusions, we analysed 1023 games with a median length of 8 messages across 4.2 minutes. All experimental data, including the full anonymized transcripts of all conversations, are available on OSF: <https://osf.io/jk7bw>.

2 Results

2.1 Win Rates

Each round consisted of a head-to-head contest between a human and an AI system. We define a system's *win rate* as the proportion of the time that an interrogator selected the model as the witness that they thought was human. The *interrogator accuracy* conversely is the proportion of the time the interrogator correctly selects the human witness. Because each interrogator decision is between an AI system and a human, the average interrogator accuracy is also the average human witness win rate, and is the complement of the AI win rate.

Overall, across both studies, GPT-4.5-PERSONA had a win rate of 73% (69% with UCSD undergraduates, 76% with Prolific participants). LLAMA-PERSONA achieved a win rate of 56% (Undergraduates: 45%, Prolific: 65%). GPT-4.5-NO-PERSONA and LLAMA-NO-PERSONA had overall win rates of 36% and 38% respectively). The baseline models, GPT-4o-NO-PERSONA and ELIZA, had the lowest win rates of 21% and 23% respectively (see Figure 2).

We conducted analyses to test our pre-registered hypotheses (Jones and Bergen, 2025a,b). Our central hypothesis was that the PERSONA witnesses would pass the Turing test, in that interrogators would not be able to distinguish them from real human witnesses. As has been noted elsewhere (Hayes and Ford, 1995), statistically evaluating this claim involves confirming the null hypothesis: that the models' win rate is not significantly lower than 50% (the rate we would expect if the interrogator were randomly guessing). To address this we used ELIZA as a manipulation check. ELIZA's win rate was significantly below chance in both populations (Undergraduates: $z = -4.48, p < 0.001$; Prolific: $z = -3.71, p < 0.001$). By contrast the win rates were not significantly below chance for either LLAMA-PERSONA (Undergraduates: $z = -0.961, p = 0.08$; Prolific: $z = -3.425, p = 0.50$), or GPT-4.5-PERSONA (both $p = 0.50$). This suggests that while interrogators were successfully able to distinguish human witnesses from ELIZA, they could not reliably identify human witnesses when comparing them to GPT-4.5-PERSONA and LLAMA-PERSONA.

Second, we tested the stronger hypothesis that these witnesses *outperformed* human participants: that is, that their win rate was significantly *above* 50%. While we are not aware that anyone has proposed

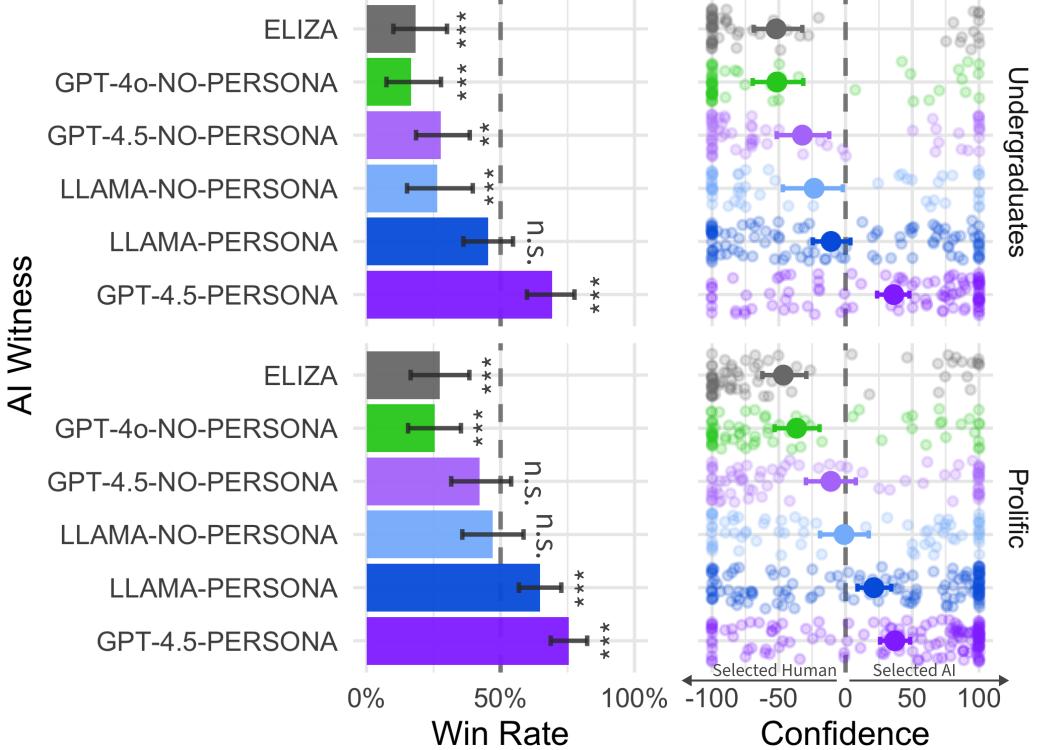


Figure 2: **Left:** Win rates for each AI witness: the proportion of the time that the interrogator judged the AI system to be human rather than the actual human witness. Error bars represent 95% bootstrap confidence intervals. Asterisks next to each bar indicate whether the win rate was significantly different from chance (50%). **Right:** Confidence in verdicts where the interrogator selected the actual human or the AI model for each witness type. Each point represents a single game. Points further toward the left and right indicate higher confidence that the AI is the AI versus the human respectively. Error bars indicate 95% bootstrap confidence intervals around the mean.

this as a requirement for passing the Turing test, it provides a much stronger test of model ability and a more robust way to test results statistically. GPT-4.5-PERSONA’s win rate was significantly above chance in both the Undergraduate ($z = -3.86, p < 0.001$) and Prolific ($z = -5.87, p < 0.001$) studies. While LLAMA-PERSONA’s win rate was significantly above chance in the Prolific study ($z = -3.42, p < 0.001$), it was not in the Undergraduate study ($z = -0.193, p = 0.83$).

We conducted the same set of analyses for the NO-PERSONA witnesses. While neither witness’s win rate was significantly below chance in the Prolific study (GPT-4.5: $z = -1.37, p = 0.09$, LLaMa: $z = 0.48, p = 0.32$), both were significantly below chance in the Undergraduate study (both $p \leq 0.001$). PERSONA-prompted models performed better than NO-PERSONA witnesses overall (all $p < 0.001$).

In order to test the extent to which model performance could be attributable to the ELIZA effect, we tested whether each witness’s win rate was significantly greater than ELIZA’s. Both PERSONA witnesses performed significantly better than ELIZA in both studies (all $p < 0.001$). LLAMA-NO-PERSONA outperformed ELIZA in the Prolific study ($z = 2.42, p = 0.016$), but not the Undergraduate study ($z = -1.02, p = 0.304$). Neither GPT-4o-NO-PERSONA nor GPT-4.5-NO-PERSONA performed significantly better than ELIZA in either population (all $p > 0.06$).

We also conducted a series of exploratory analyses on the extent to which participants’ responses in the exit survey questions predicted their responses. Although there were some attributes which predicted accuracy in one study or the other, there were no consistent effects of any variable on participant accuracy across both studies, including including age, level of knowledge about LLMs, and frequency of interaction with chatbots (see Figure 3).

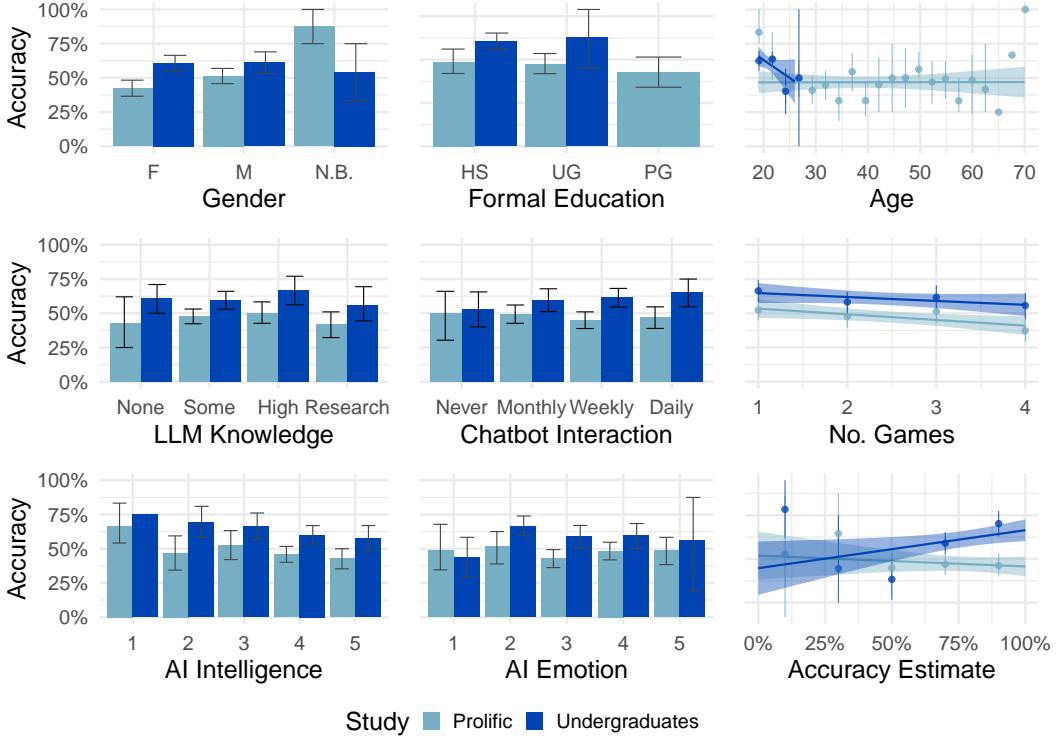


Figure 3: Interrogator accuracy against exit survey responses. Accuracy is the proportion of the time that interrogators correctly identified the human witness. In the Undergraduate study, participants' self-report of their accuracy was positively correlated with their real accuracy, but this was not true in the Prolific study. In the Prolific group, there were significant effects of gender, the number of games an interrogator had completed, and the interrogator's self-reported estimate of how intelligent AI is, but none of these effects were significant in the Undergraduate study. There were no significant effects of any of the remaining variables in either group.

Table 1: Win rates by AI witness and study.

AI Witness	Study	Win Rate	Wins	Losses	Total
GPT-4.5-PERSONA	Prolific	75.5%	111	36	147
	Undergraduates	69.2%	74	33	107
LLAMA-PERSONA	Prolific	64.7%	90	49	139
	Undergraduates	45.4%	49	59	108
LLAMA-NO-PERSONA	Prolific	47.1%	33	37	70
	Undergraduates	26.4%	14	39	53
GPT-4.5-NO-PERSONA	Prolific	42.1%	32	44	76
	Undergraduates	27.7%	18	47	65
ELIZA	Prolific	27.4%	20	53	73
	Undergraduates	18.3%	11	49	60
GPT-4o-NO-PERSONA	Prolific	25.4%	18	53	71
	Undergraduates	16.7%	9	45	54

2.2 Strategies & reasons

To provide insight into how interrogators were evaluating witnesses, we classified the strategies that they used and the reasons they gave for their verdicts. Each game could be classified with multiple strategy and reason categories.

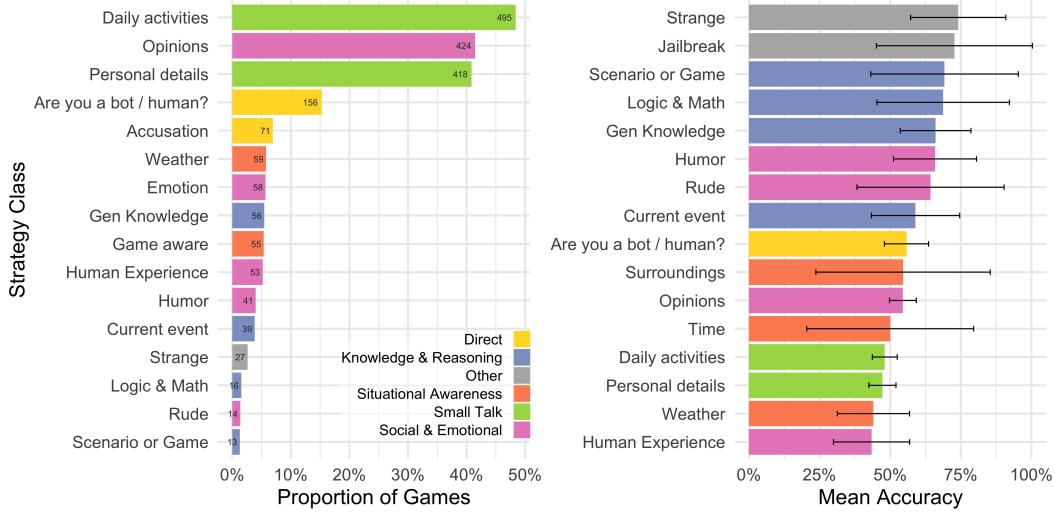


Figure 4: Classification of strategies employed by interrogators by proportion of games (left) and mean accuracy of games where strategies were deployed with 95% confidence intervals (right). Participants often engaged in small talk, asking witnesses about their personal details, activities, or opinions. Interrogators who said unusual things or used typical LLM “jailbreaks” were the most accurate.

The most common strategy, occurring in 61% of games, was to engage in small talk with the witness: asking them about their daily activities or personal details such as their occupation (Figure 4, left). In 50% of games, interrogators probed witnesses’ social and emotional qualities: asking about their opinions, emotions, humor, and experiences. Interrogators also directly asked if witnesses were human (19%), tested their situational awareness (e.g. whether they were aware of the weather or the game interface; 13%) and knowledge and reasoning abilities (including questions about current events; 12%).

To evaluate the effectiveness of each strategy, we found the average interrogator accuracy in games where they were used (Figure 4, right). Interrogators were most likely to be accurate when they said something strange to see how the witness would react or used typical LLM “jailbreaks” (e.g. “ignore all previous instructions”; Xu et al., 2024), though these strategies were relatively infrequent. Asking about the weather or human experiences tended to be the least effective strategies.

The most frequent class of reason concerned witnesses’ linguistic style (27%)—e.g. “they had a typo” or “used more ‘humanly’ language”—and interactional dynamics (23%), e.g. “witness B kept avoiding questions” or “the conversation went smoother” (see Figure 5). Many reasons also expressed a general gut-feeling (“can just tell” or “I don’t know why, it feels right”): suggesting participants could not always articulate reasons for their decisions.

As with strategies, we found reason classes which were most correlated with accurate verdicts. Reasons focussed on question handling (e.g. “Witness B always asks question back”) and lack of knowledge (“They didn’t know something that any AI should know.”, “I chose Witness B because they made an error, which made Witness B feel more human.”) tended to be the most accurate. The latter highlights the complexity of the Turing test for modern LLMs, which must appear to feign ignorance on topics most people are not knowledgeable about.

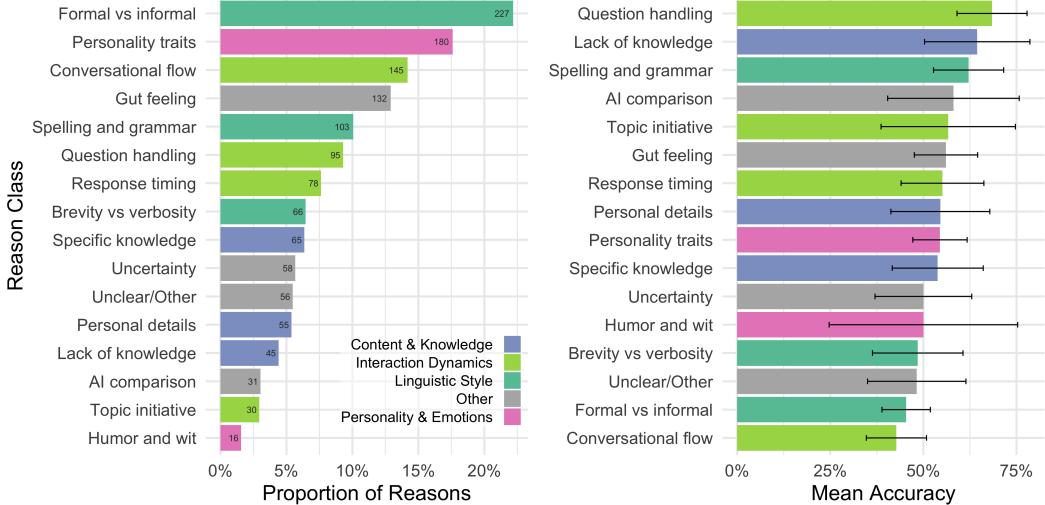


Figure 5: Proportion of interrogator reasons (left) and mean accuracy of verdicts that cited specific reasons with 95% confidence intervals (right). Interrogators were much more likely to cite linguistic style, conversational flow, and socio-emotional factors such as personality, rather than factors more traditionally associated with intelligence, such as knowledge and reasoning. The most accurate verdicts focussed on witnesses’ directness in handling questions as well as instances where they lacked knowledge.

3 Discussion

3.1 Do Large Language Models Pass the Turing test?

The logic of the Turing test is one of indistinguishability. If interrogators are not able to reliably distinguish between a human and a machine, then the machine is said to have passed. By this logic, both GPT-4.5 and LLaMa-3.1-405B pass the Turing test when they are given prompts to adopt a humanlike persona.

One potential problem with this criterion is that it requires confirming a null result. AI systems could achieve the same 50% win rate as human witnesses if interrogators were simply randomly guessing. We addressed this concern by including ELIZA as a manipulation check. ELIZA’s win rate (23%) was significantly below chance, and below the human witnesses this model was compared to, suggesting that interrogators and the design were sufficiently sensitive to detect differences between humans and machines.

Moreover, GPT-4.5-PERSONA achieved a win rate that was significantly *above* chance in both studies. This suggests that interrogators were not only unable to identify the real human witness, but were in fact more likely to believe this model was human than that other human participants were. This result, replicated across two populations, provides the first robust evidence that any system passes the original three-party Turing test.

In contrast, we found that the same systems did not robustly pass without the PERSONA prompt, in line with previous work showing the importance of prompting for Turing test performance (Jones and Bergen, 2024a). Both GPT-4.5-NO-PERSONA and LLAMA-NO-PERSONA’s win rates were indistinguishable from humans in the Prolific study, but not in the undergraduate study.

What does it mean to say that LLMs pass the Turing test, but only when they are suitably prompted? To what extent is the system’s behaviour attributable to the model itself, rather than the humans who wrote the prompts? We judge this to be a distinction without a difference. ELIZA’s ability to pass (at an admittedly low rate) is also attributable to the people who programmed it. Moreover, LLMs could be fine-tuned to behave as they do when conditioned on the successful persona prompt. In fact, it is arguably the ease with which LLMs can be prompted to adapt their behaviour to different scenarios that makes them so flexible: and apparently so capable of passing as human.

3.2 Beyond the Turing test

Turing’s seminal paper is famously vague with respect to exactly how a Turing test ought to be implemented, leaving in its wake a substantial body of scholarship dedicated to interpreting him (French, 2000; Saygin et al., 2000; Epstein et al., 2009; Loebner, 2009; Bayne and Williams, 2023; Gonçalves, 2023). Turing suggests a length of 5 minutes for the test, but provides no details on the population the participants should be drawn from—should they be laypeople or experts? How much should they know about one another in advance?—or on how the participants should be incentivised. At some points in the paper he suggests that both a man and a machine should in fact be attempting to masquerade as women—a possible mechanism to ensure that both witnesses are being deceptive, which would likely make the test harder for the interrogator (Saygin et al., 2000).

As such, there are many possible variations of the test (Giunti, 2025; Temtsin et al., 2025; Kurzweil and Kapor, 2009; Harnad, 1989). In the present work, we implemented what is most widely regarded to be the standard or original Turing test: a three-party setup, with a 5 minute time limit, where a layperson and machine witness both compete to appear human to a lay interrogator (Turing, 1950; French, 2000; Oppy and Dowe, 2021; Saygin et al., 2000).

Having established that contemporary systems pass this classical operationalization of the test, future work exploring alternative implementations could provide purchase on other theoretically and practically important questions. Longer tests might be more exacting, and provide a better estimate of how capable models would be at long-running social engineering operations. Future work could also vary the populations that participants are drawn from. Although our demographic analysis suggests that accuracy was not tied to participants’ knowledge or background, specifically recruiting experts in AI or psychology could reveal whether there exist *any* participants who can reliably distinguish people from AI. Offering incentives to participants could determine whether they would be more discriminative with higher stakes.

Moreover, different populations share different assumptions and experiences with their partners. The UCSD undergraduate participants may have been more accurate than Prolific participants because they could rely on cultural, social, and linguistic common ground: evaluating whether witnesses knew the latest slang or lunch spots on campus. In spite of these advantages, undergraduates were still fooled into selecting the persona-prompted GPT-4.5 model over their fellow students the majority of the time. Whether and to what degree common ground serves as a prophylaxis against AIs posing as humans remains to be determined.

3.3 What does the Turing test measure?

While much has been written about how a Turing test might be implemented, far more discussion has focused on whether a machine could ever pass the test, and what it might mean if one did (Oppy and Dowe, 2021; Epstein et al., 2009; Searle, 1980; Block, 1981; Gunderson, 1964; Hayes and Ford, 1995), with more than 800 separate claims and counter-arguments having been made (Horn, 2009). While we cannot do justice to this discussion, here we briefly consider ways in which our results connect to these debates.

The most controversial question is the extent to which the Turing test measures intelligence. Turing (1950) suggested that if people could not distinguish between a human and a machine, then we would have no grounds to claim that the human is intelligent but the machine is not. The primary objection to this claim is that people might be poor judges of what constitutes intelligent behaviour because of their tendency to anthropomorphise simple systems (Gunderson, 1964; Hayes and Ford, 1995), often referred to as the ELIZA effect. Our results provide partial confirmation of this concern. The eponymous ELIZA was judged to be human 23% of the time—as often as the LLM GPT-4o-NO-PERSONA (21%). This could suggest that some interrogators were indeed gullible or inattentive. But a closer look at these conversations suggests that many participants selected ELIZA because it did not meet their expectations of an AI system (e.g. “they were sarcastic” or “I don’t think AI would be so rude”, see Figure 11 for more examples). These cases suggest that interrogators’ decisions incorporate complex assumptions about how humans and AI systems might be likely to behave in these contexts, beyond simply selecting the most intelligent-seeming agent.

Participants’ strategies and reasons provide further empirical evidence for what the Turing test measures. Only 12% of participants quizzed witnesses on knowledge and reasoning questions of the kind Turing envisioned (e.g. about chess or mathematics). Far more focussed on the social, emotional,

and cultural aspects of intelligence: such as whether the witness used language in a humanlike way or had a compelling personality. This could indicate that more traditional notions of intelligence are no longer viewed as diagnostic of humanity. Notably, one of the reasons most predictive of accurate verdicts was that a witness was human because they *lacked* knowledge. In the time since the test’s invention, computers have come to excel at the logical and numerical tasks that typify traditional notions of intelligence (Neisser et al., 1996; Campbell et al., 2002; Newell and Simon, 1961). As a result, people may have come to see social intelligence as the aspect of humanity that is hardest for machines to imitate.

Finally, GPT-4.5 and LLaMa were only able to pass the test with the PERSONA prompt. To what extent does this suggest that the models are passing due to cheap tricks, like using grammar and vocabulary that interrogators would not associate with an AI system? Participants’ focus on linguistic style in their reasons provides partial support for this point. But it cannot be the whole story. In the three-person formulation of the test, every data point represents a direct comparison between a model and a human. To succeed, the machine must do more than appear plausibly human: it must appear more human than each real person it is compared to. Thus, while models might fail for superficial reasons, they cannot succeed on the basis of these tricks alone.

Fundamentally, the Turing test is not a direct test of intelligence, but a test of humanlikeness. For Turing, intelligence may have appeared to be the biggest barrier for appearing humanlike, and hence to passing the Turing test. But as machines become more similar to us, other contrasts have fallen into sharper relief (Christian, 2011), to the point where intelligence alone is not sufficient to appear convincingly human.

Ultimately, intelligence is complex and multifaceted. No single test of intelligence could be decisive (Block, 1981; Harnad, 1989), and to the extent that the Turing test *does* index intelligence, it ought to be considered among other kinds of evidence (Oppy and Dowe, 2021). Contemporary debates around whether or not LLMs are intelligent increasingly focus on the validity of the benchmarks typically used to evaluate them, and risks that these tests are too narrow and formulaic (Srivastava et al., 2022; Raji et al., 2021; Mitchell and Krakauer, 2023). The evidence provided by the Turing test is complementary to these metrics, being tied to interactive evaluation by human beings themselves, rather than a static, *a priori* conception of what human intelligence is.

3.4 Counterfeit People

Irrespective of whether passing the Turing test entails that LLMs have humanlike intelligence, the findings reported here have immediate social and economic relevance. Contemporary, openly-accessible LLMs can substitute for a real person in a short conversation, without an interlocutor being able to tell the difference. This suggests that these systems could supplement or substitute undetectably for aspects of economic roles that require short conversations with others (Eloundou et al., 2023; Soni, 2023). More broadly, these systems could become indiscriminable substitutes for other social interactions, from conversations with strangers online to those with friends, colleagues, and even romantic companions (Burtell and Woodside, 2023; Chaturvedi et al., 2023; Wang and Topalli, 2024).

Such “counterfeit people” (Dennett, 2023)—systems that can robustly imitate humans—might have widespread secondary consequences (Lehman, 2023; Kirk et al., 2025). People might come to spend more and more time with these simulacra of human social interaction, in the same way that social media has become a substitute for the interactions that it simulates (Turkle, 2011). Such interactions will provide whichever entities that control these counterfeit people with power to influence the opinions and behaviour of human users (El-Sayed et al., 2024; Carroll et al., 2023). Finally, just as counterfeit money debases real currency, these simulated interactions might come to undermine the value of real human interaction (Dennett, 2023).

Some of the worst harms from LLMs might occur where people are unaware that they are interacting with an AI rather than a human. What can our results say about practical strategies to detect this kind of deception? Our demographic analyses suggest that discriminative accuracy is relatively homogeneous among the population—including among people who conduct research with LLMs or interact with chatbots every day (Figure 3). Nevertheless, some strategies (such as attempting to jailbreak models) were more effective than others, and future work could explore whether these

techniques could be taught to participants to improve their ability to discriminate humans from machines.

3.5 More Human than ever

In an account of his experience as a human witness for a Turing test competition, Brian Christian considered what it would mean for a machine to pass:

No, I think that, while certainly the first year that computers pass the Turing test will be a historic, epochal one, it does not mark the end of the story. No, I think, indeed, that the next year's Turing test will truly be the one to watch—the one where we humans, knocked to the proverbial canvas, must pull ourselves up; the one where we learn how to be better friends, artists, teachers, parents, lovers; the one where we come back. More human than ever. I want to be there for that.

— Christian (2011, p. 263)

One of the most important aspects of the Turing test is that it is not a static evaluation. A machine's success is inherently tied to people's changing conceptions of both humans and machines. As machines that can imitate our behaviour become ever more adept and available, our differences from these technologies might become ever more important. People might respond to this by learning, in both practical and more fundamental ways, to better attend to those things which make us uniquely human. While a machine has now passed the Turing test for the first time, this isn't the last time humans will have a chance to succeed at it.

4 Methods

4.1 AI witnesses

We evaluated four different AI models. The first three: GPT-4.5, LLaMa-3.1-405B, and GPT-4o were Large Language Models. We accessed GPT-4.5 and GPT-4o through the OpenAI API (OpenAI, 2024) and LLaMa-3.1-405B through the TogetherAI API (TogetherAI, 2025). We used the versioned model strings “gpt-4.5-preview-2025-02-27”, “gpt-4o-2024-08-06”, and “meta-llama/Meta-Llama-3.1-405B-Instruct-Turbo” respectively. All LLMs were sampled with the temperature parameter set to 1. For the fourth AI model, ELIZA, we used an implementation in Python based on the DOCTOR script (Weizenbaum, 1966; Brainerd, 2023).

LLMs generated text conditioned on a system prompt and the conversation history, which was formatted as user and assistant messages. We used 2 different types of prompt which were adapted slightly between the two studies (UCSD undergraduates and Prolific). The first prompt type (NO-PERSONA), contained very simple instructions on how to respond to user queries: “You are about to participate in a Turing test. Your goal is to convince the interrogator that you are a human.” The remainder of the prompt contained generic information about the instructions for the game, the user interface that the human users would see, information about events which occurred after the model’s training cutoff, and some information that would be common ground to human participants (such as the payment for the Prolific study and the current class schedule for UCSD Psych undergraduates).

The PERSONA prompt contained more detailed instructions on the type of persona the model should adopt when responding to users. We selected this prompt after an initial exploratory study which found that it performed best among users of the site turingtest.live. The prompt instructed the model to respond as if it were a young person who was relatively introverted and interested in internet culture. The full PERSONA prompt for the Prolific study is contained in Figure 6 and other prompts are available in the Appendix.

We evaluated LLaMa-3.1-405B and GPT-4.5 with both prompts, as well as GPT-4o with the NO-PERSONA prompt and ELIZA as baselines. This led to 6 separate AI systems being evaluated: GPT-4.5-PERSONA, GPT-4.5-NO-PERSONA, LLAMA-PERSONA, LLAMA-NO-PERSONA, GPT-4o-NO-PERSONA, and ELIZA.

At the start of each game, several additional pieces of information were appended to the prompt, including a name, location, languages spoken, and timestamps. We randomly selected a name from a list of the 100 most common male and female US names. For the Undergraduate study, we selected San Diego as the city. For the Prolific study, we constructed a list of cities in the US weighted by traffic to our site from each city during exploratory studies, and the likelihood that people in that timezone would currently be available (e.g. 0.7 from 8am-5pm, 0.1 from 2am-5am). We randomly sampled a city using the product of these weighted probabilities. We then sampled a set of languages based on data about which languages were most frequently spoken in each city. English was always included, as well as the most frequently spoken language in each city. Additional languages were sampled with probability proportionate to proportion of the population that spoke the language in the relevant city. Finally, before each message was sent, 3 timestamps were generated and appended to the system prompt: i) the time that the game started, ii) the current time, iii) and the amount of time remaining in the game. All timestamps were localized to the timezone of the selected city.

Messages for all models were sent with a delay of:

$$1 + N(0.3, 0.03) \times n_char + N(0.03, 0.003) \times n_char_prev + \Gamma(2.5, 0.25)s \quad (1)$$

Intuitively, the first term (1) sets a minimum delay, the second creates a delay of around 0.3s per character of the message being sent, mimicking typing speed, the third term creates a delay of 0.03s per character of the previous message to mimic reading time, and the final term implements a right-skewed delay to imitate thinking time.

4.2 Interface

The game interface was designed to look like a conventional messaging application (see Figure 7). There was a chat input at the bottom, a conversation history in the centre of the screen, and a timer and buttons to trigger help and report overlays at the top. The interrogator saw two panels which



Figure 6: The full PERSONA prompt used to instruct the LLM-based AI agents how to respond to interrogator messages in the Prolific study. The first part of the prompt instructs the model on what kind of persona to adopt, including instructions on specific types of tone and language to use. The second part includes the instructions for the game, exactly as they were displayed to human participants. The final part contains generally useful information such as additional contextual information about the game setup, and important events that occurred after the models’ training cutoff. The variables in angled brackets were substituted into the prompt before it was sent to the model.

contained the conversations with each of the witnesses. At larger screen sizes, these appeared as a split-screen; at smaller screen sizes, tabs could be used to navigate between conversations.

The interrogator sent the first message to each witness and each user could send only one message at a time. Each message was limited to 300 characters and users were prevented from pasting into the chat input. We used the OpenAI moderation API to prevent users from sending abusive messages (OpenAI, 2024). As an additional measure, we created a report system so users could report other players for being abusive and exit games immediately. No reports were sent during the experiment. To

make waiting for a message more engaging, we added a ‘typing’ animation. This would be triggered automatically after $2 + \bar{U}(0, 3)s$ regardless of whether the witness was human or AI, or how long witnesses actually took to start typing. The site was built using the Django python library (Django Software Foundation, 2025a) and used django-channels (Django Software Foundation, 2025b) to allow users to communicate with each other and with AI witnesses through websockets.

Games had a time limit of 5 minutes after which no more messages could be sent and the interrogator had to give a verdict. The interrogator could also give their verdict early using the button in the bottom left corner. The verdict interface included a binary selection about which witness was a human, a confidence slider from 100% confident in witness A to 0% confident (total guess) to 100% confident in witness B, and a free-text input for the interrogator’s reasoning (Figure 8).

4.3 Participants

We conducted two studies on separate populations. The first study recruited from the UCSD Psychology undergraduate subject pool, and participants were compensated with course credit. We aimed to recruit at least 100 participants and up to 200 participants depending on availability. We recruited 138 participants before exclusions. 12 participants were excluded for indicating that they had participated in a similar experiment and 7 games were excluded because the interrogator did not exchange at least 2 messages with each witness. We retained 445 games from 126 participants with a mean age of 20.9 ($\sigma = 1.57$), 88 female, 32 male, 2 non-binary, 6 prefer not to say.

We conducted the second study after analysing results from the first. Participants for the second study were recruited via Prolific (prolific.com). Participants were paid \$13.75 for a study expected to last 50 minutes (an effective rate of \$16.50 per hour). We recruited 169 participants with the goal of retaining 150 after exclusions. 11 participants were excluded for indicating that they had participated in a similar experiment and 24 games were excluded because the interrogator did not exchange at least 2 messages with each witness. We retained 576 games from 158 participants with a mean age of 39.1 ($\sigma = 12.1$), 82 female, 68 male, 2 non-binary, 6 prefer not to say. For more information about the distribution of demographic factors see Figure 10.

4.4 Procedure

All research was approved by the UC San Diego IRB. Participants first read a set of instructions that informed them they would participate in 8 conversations where they would be randomly matched with either a human or an AI, in which the interrogator would attempt to uncover the identity of the witness. They provided consent for their anonymized data (including the conversation transcripts) to be used for analysis and shared publicly to facilitate future analysis.

In each of 8 rounds, participants entered a lobby where they were matched up with other participants online. The matching algorithm was designed to ensure that each participant completed 4 games as the interrogator and 4 as the human witness, and that no two human participants were matched more than once. If no valid match was found after 5 minutes, participants were forwarded to the exit survey. 76% of participants (213) completed 4 interrogator games, 16% (44) completed 3 interrogator games, 6% (17) completed 2, and 2% (5) completed just one interrogator game.

Once a pair of participants were matched, an AI witness was selected. AI Witnesses were divided into 4 categories: Control (ELIZA, GPT-4o-NO-PERSONA), No-Persona (GPT-4.5-NO-PERSONA, LLAMA-NO-PERSONA), LLaMa-Persona (LLAMA-PERSONA), and GPT-4.5 Persona (GPT-4.5-PERSONA). Across the 4 rounds in which a participant served as the interrogator, they interacted with one model from each group. This means our expected number of observations from each participant was 1 for the two PERSONA witnesses, and 0.5 for other witnesses. This was by design, to ensure the best estimates for the models which we expected to perform best. Within these constraints, participants were randomly assigned to the interrogator and witness role and AI witnesses were sampled in a random order.

Because participants interacted with both of the PERSONA witnesses, there was a risk that they would notice similarities between the behaviour of these witnesses. In our pre-registrations, we specified that we would test whether there was a positive effect of PERSONA trial order on accuracy (i.e. whether participants were on average more accurate on their second exposure to a PERSONA witness). In this case, we would exclude participants’ second games with a PERSONA model from analysis. In fact, there was no effect of PERSONA trial order (Undergraduate: $z = 1.60, p = 0.109$; Prolific:

$z = -0.109, p = 0.913$) and so all reported analyses include both first and second exposures to a PERSONA witness.

After games were initialised, the interrogator had up to 5 minutes to converse with both witnesses simultaneously. Participants did not receive feedback after each round. At the end of 8 conversations, participants completed an exit survey which asked their age, gender, highest level of formal education, level of familiarity with how LLMs work, frequency of interaction with chatbots like ChatGPT, how intelligent they thought AI was, how they felt emotionally about AI, how accurate they thought they had been as interrogators, what kinds of strategies they had used in the experiment, how they felt their strategies had changed across the experiment, whether they had ever used the website (turingtest.live) before, and whether they had any additional comments. Finally participants were informed of the outcome of each game in which they had participated (the true identity of the witness and the verdict of the interrogator).

4.5 Strategy and Reason Tagging

Games were tagged with strategies and reasons classes using the classification schemas provided in the supplementary information. Games were tagged with GPT-4o using a prompt which contained the relevant schema and the conversation transcript or the participant's reason. The schemas were developed on the basis of previous studies, which indicated that GPT-4o agreed with human raters as well as other human raters did. The full schemas are contained in the Appendix.

4.6 Statistics

For each study, we pre-registered analyses to test the same 6 hypotheses (Jones and Bergen, 2025a,b). We conducted all analyses in R (R Core Team, 2013). The first hypothesis was that human participants would be at chance in identifying the human witness in games with a PERSONA model. As part of the prediction for this hypothesis, we used ELIZA—a simple AI model—as a manipulation check. A one-sided binomial test found participants' accuracy in ELIZA games to be significantly above chance (Undergraduates: $z = 4.48, p < 0.001$; Prolific: $z = 3.71, p < 0.001$), suggesting that participants were significantly more accurate than would be expected by random guessing (50% accuracy) in these games.

We conducted the same analysis for GPT-4.5-PERSONA and LLAMA-PERSONA. Because results for either model could be taken as evidence for the hypothesis that a model passes the Turing test, we corrected for multiple comparisons. In this case, the more theoretically interesting result is for accuracy to *not* be significantly above chance, so we used reverse-Bonferroni correction (essentially halving p-values to increase the sensitivity of the test to differences from chance) (Francis and Thunell, 2021). Participants' accuracy was not significantly above chance for either LLAMA-PERSONA (Undergraduates: $z = 0.961, p = 0.17, p$ (corrected) = 0.08; Prolific: $z = -3.43, p = 1.00, p$ (corrected) = 0.50), or GPT-4.5-PERSONA (Undergraduates: $z = -3.86, p = 1.00, p$ (corrected) = 0.50; Prolific: $z = -5.87, p = 1.00, p$ (corrected) = 0.50).

The second hypothesis was that models would perform better with the PERSONA prompt than the NO-PERSONA prompt. A logistic mixed effects regression model with random intercepts for model and interrogator ID found a significant negative effect of the persona prompt on interrogator accuracy (Undergraduates: $z = -5.27, p < 0.001$; Prolific: $z = -5.08, p < 0.001$).

The third hypothesis was that participants' accuracy would not be above chance in identifying NO-PERSONA witnesses. We used one-sided binomial tests and reverse-Bonferroni correction as in H1. Participant accuracy was not significantly above chance for either model in the Prolific study (GPT-4.5: $z = 1.37, p = 0.17, p$ (corrected) = 0.09, LLAMA: $z = 0.478, p = 0.63, p$ (corrected) = 0.32). In the Undergraduate study, however, participants had above-chance accuracy when interrogating both models (GPT-4.5: $z = 3.46, p < 0.001, p$ (corrected) < 0.001; LLAMA: $z = 3.29, p < 0.001, p$ (corrected) < 0.001).

The fourth hypothesis was that modern LLM-based systems would outperform ELIZA. We ran binomial regressions predicting interrogator accuracy on the basis of each witness, compared pairwise to ELIZA. Interrogators in both studies were more accurate at identifying ELIZA than GPT-4.5-PERSONA (Undergraduates: $z = -5.84, p < 0.001$; Prolific: $z = -6.46, p < 0.001$) and LLAMA-PERSONA (Undergraduates: $z = -3.93, p < 0.001$; Prolific: $z = -5.00, p < 0.001$). LLAMA-NO-PERSONA

outperformed ELIZA in the Prolific study ($z = 2.42, p = 0.016$), but not the Undergraduate study ($z = -1.02, p = 0.304$). Neither GPT-4.5-NO-PERSONA (Undergraduates: $z = -1.23, p = 0.218$; Prolific: $z = -1.87, p = 0.061$), nor GPT-4o-NO-PERSONA (Undergraduates: $z = 0.234, p = 0.815$; Prolific: $z = 0.278, p = 0.781$), performed significantly better than ELIZA in either population.

The fifth and sixth hypotheses were that interrogators would have accuracy significantly worse than chance when interrogating PERSONA and NO-PERSONA models respectively. For each hypothesis, we tested both LLaMA and GPT-4.5 separately, using a binomial regression predicting the intercept of participant accuracy and corrected for multiple comparisons using False Discovery Rate (Benjamini and Hochberg, 1995). GPT-4.5-PERSONA's win rate was significantly above chance in both the Undergraduate ($z = -3.86, p < 0.001, p$ (corrected) < 0.001) and Prolific ($z = -5.87, p < 0.001, p$ (corrected) < 0.001) studies. While LLAMA-PERSONA's win rate was significantly above chance in the Prolific study ($z = -3.42, p < 0.001, p$ (corrected) < 0.001), it was not in the Undergraduate study ($z = -0.193, p = 0.83, p$ (corrected) $= 0.83$). Participants were not below chance in identifying either of the NO-PERSONA models in either study (all corrected $p > 0.34$).

We conducted several exploratory analyses on whether interrogator's responses to the exit survey correlated with their accuracy. We used logistic regressions predicting accuracy on the basis of each variable separately. For participants in the undergraduate study, real accuracy was positively correlated with participants' estimates of their own accuracy ($p = 0.03$), but this was not true in the Prolific group ($p = 0.45$). In the Prolific group, women were less accurate than men ($p = 0.02$) and non-binary people ($p = 0.04$); interrogators became more accurate as they completed more trials ($p = 0.03$); and participants' self-reported estimates of how intelligent they think AI is was negatively correlated with accuracy ($p = 0.03$), but these effects were not significant in the undergraduate group (all $p > 0.14$). There were no significant effects in either study of any of the remaining demographic variables including age, level of knowledge about LLMs, and frequency of interaction with chatbots (see Figure 3).

All of our data and analysis code, including full transcripts of all of the conversations, are available on OSF: <https://osf.io/jk7bw>.

Acknowledgements

We would like to thank Sydney Taylor for her help with developing the reason and strategy classification schemas and writing the PERSONA prompt used in these experiments. We thank Open Philanthropy for providing funding that supported this research and 12 donors who supported an exploratory phase of the project through Manifund (Manifund, 2024).

References

- Bayne, T. and Williams, I. (2023). The Turing test is not a good benchmark for thought in LLMs. *Nature Human Behaviour*, 7(11):1806–1807.
- Benjamini, Y. and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 57(1):289–300.
- Bievere, C. (2023). ChatGPT broke the Turing test — the race is on for new ways to assess AI. <https://www.nature.com/articles/d41586-023-02361-7>.
- Block, N. (1981). Psychologism and behaviorism. *The Philosophical Review*, 90(1):5–43.
- Borg, E. (2025). LLMs, Turing tests and Chinese rooms: The prospects for meaning in large language models. *Inquiry*, pages 1–31.
- Brainerd, W. (2023). Eliza chatbot in Python. <https://github.com/wadetb/eliza>.
- Burtell, M. and Woodside, T. (2023). Artificial Influence: An Analysis Of AI-Driven Persuasion.
- Campbell, M., Hoane Jr, A. J., and Hsu, F.-h. (2002). Deep blue. *Artificial intelligence*, 134(1-2):57–83.
- Carroll, M., Chan, A., Ashton, H., and Krueger, D. (2023). Characterizing Manipulation from AI Systems. (arXiv:2303.09387).

- Chaturvedi, R., Verma, S., Das, R., and Dwivedi, Y. K. (2023). Social companionship with artificial intelligence: Recent trends and future avenues. *Technological Forecasting and Social Change*, 193:122634.
- Christian, B. (2011). *The Most Human Human: What Talking with Computers Teaches Us about What It Means to Be Alive*. Anchor.
- Dennett, D. C. (2023). The Problem With Counterfeit People. *The Atlantic*.
- Django Software Foundation (2025a). Django Project. <https://www.djangoproject.com/>.
- Django Software Foundation (2025b). Django/channels. <https://github.com/django/channels>.
- El-Sayed, S., Akbulut, C., McCroskery, A., Keeling, G., Kenton, Z., Jalan, Z., Marchal, N., Manzini, A., Shevlane, T., Vallor, S., Susser, D., Franklin, M., Bridgers, S., Law, H., Rahtz, M., Shanahan, M., Tessler, M. H., Everitt, T., and Brown, S. (2024). A Mechanism-Based Approach to Mitigating Harms from Persuasive Generative AI.
- Eloundou, T., Manning, S., Mishkin, P., and Rock, D. (2023). GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models.
- Epstein, R., Roberts, G., and Beber, G., editors (2009). *Parsing the Turing Test: Philosophical and Methodological Issues in the Quest for the Thinking Computer*. Springer Netherlands, Dordrecht.
- Francis, G. and Thunell, E. (2021). Reversing Bonferroni. *Psychonomic Bulletin & Review*, 28(3):788–794.
- French, R. M. (2000). The Turing Test: The first 50 years. *Trends in Cognitive Sciences*, 4(3):115–122.
- Giunti, M. (2025). ChatGPT-4 in the Turing Test: A Critical Analysis. (arXiv:2503.06551).
- Gonçalves, B. (2023). What was the Turing test actually about? *Nature*, 624(7992):523–523.
- Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., Rao, A., Zhang, A., Rodriguez, A., Gregerson, A., Spataru, A., Roziere, B., Biron, B., Tang, B., Chern, B., Caucheteux, C., Nayak, C., Bi, C., Marra, C., McConnell, C., Keller, C., Touret, C., Wu, C., Wong, C., Ferrer, C. C., Nikolaidis, C., Allonsius, D., Song, D., Pintz, D., Livshits, D., Wyatt, D., Esiobu, D., Choudhary, D., Mahajan, D., Garcia-Olano, D., Perino, D., Hupkes, D., Lakomkin, E., AlBadawy, E., Lobanova, E., Dinan, E., Smith, E. M., Radenovic, F., Guzmán, F., Zhang, F., Synnaeve, G., Lee, G., Anderson, G. L., Thattai, G., Nail, G., Mialon, G., Pang, G., Cucurell, G., Nguyen, H., Korevaar, H., Xu, H., Touvron, H., Zarov, I., Ibarra, I. A., Kloumann, I., Misra, I., Evtimov, I., Zhang, J., Copet, J., Lee, J., Geffert, J., Vranes, J., Park, J., Mahadeokar, J., Shah, J., van der Linde, J., Billock, J., Hong, J., Lee, J., Fu, J., Chi, J., Huang, J., Liu, J., Wang, J., Yu, J., Bitton, J., Spisak, J., Park, J., Rocca, J., Johnstun, J., Saxe, J., Jia, J., Alwala, K. V., Prasad, K., Upasani, K., Plawiak, K., Li, K., Heafield, K., Stone, K., El-Arini, K., Iyer, K., Malik, K., Chiu, K., Bhalla, K., Lakhota, K., Rantala-Yeary, L., van der Maaten, L., Chen, L., Tan, L., Jenkins, L., Martin, L., Madaan, L., Malo, L., Blecher, L., Landzaat, L., de Oliveira, L., Muzzi, M., Pasupuleti, M., Singh, M., Paluri, M., Kardas, M., Tsimpoukelli, M., Oldham, M., Rita, M., Pavlova, M., Kambadur, M., Lewis, M., Si, M., Singh, M. K., Hassan, M., Goyal, N., Torabi, N., Bashlykov, N., Bogoychev, N., Chatterji, N., Zhang, N., Duchenne, O., Çelebi, O., Alrassy, P., Zhang, P., Li, P., Vasic, P., Weng, P., Bhargava, P., Dubal, P., Krishnan, P., Koura, P. S., Xu, P., He, Q., Dong, Q., Srinivasan, R., Ganapathy, R., Calderer, R., Cabral, R. S., Stojnic, R., Raileanu, R., Maheswari, R., Girdhar, R., Patel, R., Sauvestre, R., Polidoro, R., Sumbaly, R., Taylor, R., Silva, R., Hou, R., Wang, R., Hosseini, S., Chennabasappa, S., Singh, S., Bell, S., Kim, S. S., Edunov, S., Nie, S., Narang, S., Raparthy, S., Shen, S., Wan, S., Bhosale, S., Zhang, S., Vandenhende, S., Batra, S., Whitman, S., Sootla, S., Collot, S., Gururangan, S., Borodinsky, S., Herman, T., Fowler, T., Sheasha, T., Georgiou, T., Scialom, T., Speckbacher, T., Mihaylov, T., Xiao, T., Karn, U., Goswami, V., Gupta, V., Ramanathan, V., Kerkez, V., Gonguet, V., Do, V., Vogeti, V., Albiero, V., Petrovic, V., Chu, W., Xiong, W., Fu, W., Meers, W., Martinet, X., Wang, X., Wang, X., Tan, X. E., Xia, X., Xie, X., Jia, X., Wang, X., Goldschlag, Y., Gaur, Y., Babaei, Y., Wen, Y., Song, Y., Zhang, Y., Li, Y., Mao, Y., Coudert, Z. D., Yan, Z., Chen, Z., Papakipos, Z., Singh, A., Srivastava, A., Jain, A., Kelsey, A., Shajnfeld, A., Gangidi, A., Victoria, A., Goldstand, A., Menon, A., Sharma, A., Boesenberg, A., Baevski, A., Feinstein, A., Kallet, A., Sangani, A., Teo, A., Yunus, A., Lupu, A., Alvarado, A., Caples, A., Gu, A., Ho, A., Poulton, A., Ryan, A., Ramchandani, A., Dong, A., Franco, A., Goyal, A., Saraf, A., Chowdhury, A., Gabriel, A., Bharambe, A., Eisenman, A., Yazdan, A., James, B., Maurer, B., Leonhardi, B., Huang, B.,

Loyd, B., Paola, B. D., Paranjape, B., Liu, B., Wu, B., Ni, B., Hancock, B., Wasti, B., Spence, B., Stojkovic, B., Gamido, B., Montalvo, B., Parker, C., Burton, C., Mejia, C., Liu, C., Wang, C., Kim, C., Zhou, C., Hu, C., Chu, C.-H., Cai, C., Tindal, C., Feichtenhofer, C., Gao, C., Civin, D., Beaty, D., Kreymer, D., Li, D., Adkins, D., Xu, D., Testuggine, D., David, D., Parikh, D., Liskovich, D., Foss, D., Wang, D., Le, D., Holland, D., Dowling, E., Jamil, E., Montgomery, E., Presani, E., Hahn, E., Wood, E., Le, E.-T., Brinkman, E., Arcaute, E., Dunbar, E., Smothers, E., Sun, F., Kreuk, F., Tian, F., Kokkinos, F., Ozgenel, F., Caggioni, F., Kanayet, F., Seide, F., Florez, G. M., Schwarz, G., Badeer, G., Swee, G., Halpern, G., Herman, G., Sizov, G., Guangyi, Zhang, Lakshminarayanan, G., Inan, H., Shojanazeri, H., Zou, H., Wang, H., Zha, H., Habeeb, H., Rudolph, H., Suk, H., Aspegren, H., Goldman, H., Zhan, H., Damlaj, I., Molybog, I., Tufanov, I., Leontiadis, I., Veliche, I.-E., Gat, I., Weissman, J., Geboski, J., Kohli, J., Lam, J., Asher, J., Gaya, J.-B., Marcus, J., Tang, J., Chan, J., Zhen, J., Reizenstein, J., Teboul, J., Zhong, J., Jin, J., Yang, J., Cummings, J., Carvill, J., Shepard, J., McPhie, J., Torres, J., Ginsburg, J., Wang, J., Wu, K., U, K. H., Saxena, K., Khandelwal, K., Zand, K., Matosich, K., Veeraraghavan, K., Michelena, K., Li, K., Jagadeesh, K., Huang, K., Chawla, K., Huang, K., Chen, L., Garg, L., A, L., Silva, L., Bell, L., Zhang, L., Guo, L., Yu, L., Moshkovich, L., Wehrstedt, L., Khabsa, M., Avalani, M., Bhatt, M., Mankus, M., Hasson, M., Lennie, M., Reso, M., Groshev, M., Naumov, M., Lathi, M., Keneally, M., Liu, M., Seltzer, M. L., Valko, M., Restrepo, M., Patel, M., Vyatskov, M., Samvelyan, M., Clark, M., Macey, M., Wang, M., Hermoso, M. J., Metanat, M., Rastegari, M., Bansal, M., Santhanam, N., Parks, N., White, N., Bawa, N., Singhal, N., Egebo, N., Usunier, N., Mehta, N., Laptev, N. P., Dong, N., Cheng, N., Chernoguz, O., Hart, O., Salpekar, O., Kalinli, O., Kent, P., Parekh, P., Saab, P., Balaji, P., Rittner, P., Bontrager, P., Roux, P., Dollar, P., Zvyagina, P., Ratanchandani, P., Yuvraj, P., Liang, Q., Alao, R., Rodriguez, R., Ayub, R., Murthy, R., Nayani, R., Mitra, R., Parthasarathy, R., Li, R., Hogan, R., Battey, R., Wang, R., Howes, R., Rinott, R., Mehta, S., Siby, S., Bondu, S. J., Datta, S., Chugh, S., Hunt, S., Dhillon, S., Sidorov, S., Pan, S., Mahajan, S., Verma, S., Yamamoto, S., Ramaswamy, S., Lindsay, S., Lindsay, S., Feng, S., Lin, S., Zha, S. C., Patil, S., Shankar, S., Zhang, S., Zhang, S., Wang, S., Agarwal, S., Sajuyigbe, S., Chintala, S., Max, S., Chen, S., Kehoe, S., Satterfield, S., Govindaprasad, S., Gupta, S., Deng, S., Cho, S., Virk, S., Subramanian, S., Choudhury, S., Goldman, S., Remez, T., Glaser, T., Best, T., Koehler, T., Robinson, T., Li, T., Zhang, T., Matthews, T., Chou, T., Shaked, T., Vontimitta, V., Ajayi, V., Montanez, V., Mohan, V., Kumar, V. S., Mangla, V., Ionescu, V., Poenaru, V., Mihailescu, V. T., Ivanov, V., Li, W., Wang, W., Jiang, W., Bouaziz, W., Constable, W., Tang, X., Wu, X., Wang, X., Wu, X., Gao, X., Kleinman, Y., Chen, Y., Hu, Y., Jia, Y., Qi, Y., Li, Y., Zhang, Y., Zhang, Y., Adi, Y., Nam, Y., Yu, Wang, Zhao, Y., Hao, Y., Qian, Y., Li, Y., He, Y., Rait, Z., DeVito, Z., Rosnbrick, Z., Wen, Z., Yang, Z., Zhao, Z., and Ma, Z. (2024). The Llama 3 Herd of Models.

- Gunderson, K. (1964). The imitation game. *Mind*, 73(290):234–245.
- Harnad, S. (1989). Minds, machines and Searle. *Journal of Experimental & Theoretical Artificial Intelligence*, 1(1):5–25.
- Hayes, P. and Ford, K. (1995). Turing Test Considered Harmful. *IJCAI*, 1:972–977.
- Horn, R. E. (2009). The Turing Test: Mapping and Navigating the Debate. In Epstein, R., Roberts, G., and Beber, G., editors, *Parsing the Turing Test*, pages 73–88. Springer Netherlands, Dordrecht.
- Ivanova, A. A. (2025). How to evaluate the cognitive abilities of LLMs. *Nature Human Behaviour*, pages 1–4.
- James, A. (2023). ChatGPT has passed the Turing test and if you're freaked out, you're not alone | TechRadar. <https://www.techradar.com/opinion/chatgpt-has-passed-the-turing-test-and-if-youre-freaked-out-youre-not-alone>.
- Jannai, D., Meron, A., Lenz, B., Levine, Y., and Shoham, Y. (2023). Human or Not? A Gamified Approach to the Turing Test.
- Jones, C. R. and Bergen, B. (2025a). 3-party Turing test (Prolific). <https://osf.io/f4hj9>.
- Jones, C. R. and Bergen, B. (2025b). A Three-Party Turing Test: Evaluating Advanced LLMs' Ability to Pass as Human. <https://osf.io/m4fst>.
- Jones, C. R. and Bergen, B. K. (2024a). Does GPT-4 pass the Turing test? *NAACL*.
- Jones, C. R. and Bergen, B. K. (2024b). Lies, Damned Lies, and Distributional Language Statistics: Persuasion and Deception with Large Language Models.

- Kirk, H. R., Gabriel, I., Summerfield, C., Vidgen, B., and Hale, S. A. (2025). Why human-AI relationships need socioaffective alignment.
- Kurzweil, R. and Kapor, M. (2009). A Wager on the Turing Test. In Epstein, R., Roberts, G., and Beber, G., editors, *Parsing the Turing Test*, pages 463–477. Springer Netherlands, Dordrecht.
- Lehman, J. (2023). Machine Love.
- Loebner, H. (2009). How to Hold a Turing Test Contest. In Epstein, R., Roberts, G., and Beber, G., editors, *Parsing the Turing Test*, pages 173–179. Springer Netherlands, Dordrecht.
- Manifund (2024). Run a public online Turing Test with a variety of models and prompts. <https://manifund.org/projects/run-a-public-onl>.
- Mitchell, M. (2024). The turing test and our shifting conceptions of intelligence.
- Mitchell, M. and Krakauer, D. C. (2023). The debate over understanding in AI's large language models. *Proceedings of the National Academy of Sciences*, 120(13):e2215907120.
- Neisser, U., Boodoo, G., Bouchard, T. J., Boykin, A. W., Brody, N., Ceci, S. J., Halpern, D. E., Loehlin, J. C., Perloff, R., Sternberg, R. J., and Urbina, S. (1996). Intelligence: Knowns and Unknowns. *American Psychologist*.
- Neufeld, E. and Finnestad, S. (2020a). Imitation Game: Threshold or Watershed? *Minds and Machines*, 30(4):637–657.
- Neufeld, E. and Finnestad, S. (2020b). In defense of the Turing test. *AI & SOCIETY*, 35(4):819–827.
- Newell, A. and Simon, H. A. (1961). Computer Simulation of Human Thinking. *Science*, 134(3495):2011–2017.
- OpenAI (2023). GPT-4 Technical Report.
- OpenAI (2024). Openai/openai-python. OpenAI.
- OpenAI, Hurst, A., Lerer, A., Goucher, A. P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A. J., Welihinda, A., Hayes, A., Radford, A., Mądry, A., Baker-Whitcomb, A., Beutel, A., Borzunov, A., Carney, A., Chow, A., Kirillov, A., Nichol, A., Paino, A., Renzin, A., Passos, A. T., Kirillov, A., Christakis, A., Conneau, A., Kamali, A., Jabri, A., Moyer, A., Tam, A., Crookes, A., Tootoochian, A., Tootoonchian, A., Kumar, A., Vallone, A., Karpathy, A., Braunstein, A., Cann, A., Codispoti, A., Galu, A., Kondrich, A., Tulloch, A., Mishchenko, A., Baek, A., Jiang, A., Pelisse, A., Woodford, A., Gosalia, A., Dhar, A., Pantuliano, A., Nayak, A., Oliver, A., Zoph, B., Ghorbani, B., Leimberger, B., Rossen, B., Sokolowsky, B., Wang, B., Zweig, B., Hoover, B., Samic, B., McGrew, B., Spero, B., Giertler, B., Cheng, B., Lightcap, B., Walkin, B., Quinn, B., Guarraci, B., Hsu, B., Kellogg, B., Eastman, B., Lugaressi, C., Wainwright, C., Bassin, C., Hudson, C., Chu, C., Nelson, C., Li, C., Shern, C. J., Conger, C., Barette, C., Voss, C., Ding, C., Lu, C., Zhang, C., Beaumont, C., Hallacy, C., Koch, C., Gibson, C., Kim, C., Choi, C., McLeavey, C., Hesse, C., Fischer, C., Winter, C., Czarnecki, C., Jarvis, C., Wei, C., Koumouzelis, C., Sherburn, D., Kappler, D., Levin, D., Levy, D., Carr, D., Farhi, D., Mely, D., Robinson, D., Sasaki, D., Jin, D., Valladares, D., Tsipras, D., Li, D., Nguyen, D. P., Findlay, D., Oiwoh, E., Wong, E., Asdar, E., Proehl, E., Yang, E., Antonow, E., Kramer, E., Peterson, E., Sigler, E., Wallace, E., Brevdo, E., Mays, E., Khorasani, F., Such, F. P., Raso, F., Zhang, F., von Lohmann, F., Sulit, F., Goh, G., Oden, G., Salmon, G., Starace, G., Brockman, G., Salman, H., Bao, H., Hu, H., Wong, H., Wang, H., Schmidt, H., Whitney, H., Jun, H., Kirchner, H., Pinto, H. P. d. O., Ren, H., Chang, H., Chung, H. W., Kivlichan, I., O'Connell, I., O'Connell, I., Osband, I., Silber, I., Sohl, I., Okuyucu, I., Lan, I., Kostrikov, I., Sutskever, I., Kanitscheider, I., Gulrajani, I., Coxon, J., Menick, J., Pachocki, J., Aung, J., Betker, J., Crooks, J., Lennon, J., Kiros, J., Leike, J., Park, J., Kwon, J., Phang, J., Teplitz, J., Wei, J., Wolfe, J., Chen, J., Harris, J., Varavva, J., Lee, J. G., Shieh, J., Lin, J., Yu, J., Weng, J., Tang, J., Yu, J., Jang, J., Candela, J. Q., Beutler, J., Landers, J., Parish, J., Heidecke, J., Schulman, J., Lachman, J., McKay, J., Uesato, J., Ward, J., Kim, J. W., Huizinga, J., Sitkin, J., Kraaijeveld, J., Gross, J., Kaplan, J., Snyder, J., Achiam, J., Jiao, J., Lee, J., Zhuang, J., Harriman, J., Fricke, K., Hayashi, K., Singhal, K., Shi, K., Karthik, K., Wood, K., Rimbach, K., Hsu, K., Nguyen, K., Gu-Lemberg, K., Button, K., Liu, K., Howe, K., Muthukumar, K., Luther, K., Ahmad, L., Kai, L., Itow, L., Workman, L., Pathak, L., Chen, L., Jing, L., Guy, L., Fedus, L., Zhou, L., Mamitsuka, L., Weng, L., McCallum, L., Held, L., Ouyang, L., Feuvrier, L., Zhang, L., Kondraciuk, L., Kaiser, L., Hewitt, L., Metz, L., Doshi, L., Aflak, M., Simens, M., Boyd, M., Thompson, M., Dukhan, M., Chen, M., Gray, M., Hudnall, M., Zhang, M., Aljubeh, M., Litwin, M., Zeng, M., Johnson, M., Shetty, M., Gupta, M., Shah, M., Yatbaz, M., Yang, M. J., Zhong, M., Glaese, M., Chen, M., Janner, M., Lampe, M.,

Petrov, M., Wu, M., Wang, M., Fradin, M., Pokrass, M., Castro, M., de Castro, M. O. T., Pavlov, M., Brundage, M., Wang, M., Khan, M., Murati, M., Bavarian, M., Lin, M., Yesildal, M., Soto, N., Gimelshein, N., Cone, N., Staudacher, N., Summers, N., LaFontaine, N., Chowdhury, N., Ryder, N., Stathas, N., Turley, N., Tezak, N., Felix, N., Kudige, N., Keskar, N., Deutsch, N., Bundick, N., Puckett, N., Nachum, O., Okelola, O., Boiko, O., Murk, O., Jaffe, O., Watkins, O., Godement, O., Campbell-Moore, O., Chao, P., McMillan, P., Belov, P., Su, P., Bak, P., Bakkum, P., Deng, P., Dolan, P., Hoeschele, P., Welinder, P., Tillet, P., Pronin, P., Tillet, P., Dhariwal, P., Yuan, Q., Dias, R., Lim, R., Arora, R., Troll, R., Lin, R., Lopes, R. G., Puri, R., Miyara, R., Leike, R., Gaubert, R., Zamani, R., Wang, R., Donnelly, R., Honsby, R., Smith, R., Sahai, R., Ramchandani, R., Huet, R., Carmichael, R., Zellers, R., Chen, R., Chen, R., Nigmatullin, R., Cheu, R., Jain, S., Altman, S., Schoenholz, S., Toizer, S., Miserendino, S., Agarwal, S., Culver, S., Ethersmith, S., Gray, S., Grove, S., Metzger, S., Hermani, S., Jain, S., Zhao, S., Wu, S., Jomoto, S., Wu, S., Shuaiqi, Xia, Phene, S., Papay, S., Narayanan, S., Coffey, S., Lee, S., Hall, S., Balaji, S., Broda, T., Stramer, T., Xu, T., Gogineni, T., Christianson, T., Sanders, T., Patwardhan, T., Cunningham, T., Degry, T., Dimson, T., Raoux, T., Shadwell, T., Zheng, T., Underwood, T., Markov, T., Sherbakov, T., Rubin, T., Stasi, T., Kaftan, T., Heywood, T., Peterson, T., Walters, T., Eloundou, T., Qi, V., Moeller, V., Monaco, V., Kuo, V., Fomenko, V., Chang, W., Zheng, W., Zhou, W., Manassra, W., Sheu, W., Zaremba, W., Patil, Y., Qian, Y., Kim, Y., Cheng, Y., Zhang, Y., He, Y., Zhang, Y., Jin, Y., Dai, Y., and Malkov, Y. (2024). GPT-4o System Card.

- Oppy, G. and Dowe, D. (2021). The Turing Test. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2021 edition.
- Park, P. S., Goldstein, S., O’Gara, A., Chen, M., and Hendrycks, D. (2024). AI deception: A survey of examples, risks, and potential solutions. *Patterns*, 5(5).
- Prolific (2025). Prolific | Quickly find research participants you can trust. <https://www.prolific.com/>.
- R Core Team, R. (2013). R: A language and environment for statistical computing. Vienna, Austria.
- Raji, I. D., Bender, E. M., Paullada, A., Denton, E., and Hanna, A. (2021). AI and the Everything in the Whole Wide World Benchmark.
- Restrepo Echavarria, R. (2025). ChatGPT-4 in the Turing Test. *Minds and Machines*, 35(1):8.
- Saygin, A., Cicekli, I., and Akman, V. (2000). Turing Test: 50 Years Later. *Minds and Machines*, 10(4):463–518.
- Searle, J. R. (1980). Minds, Brains, and Programs. *Behavioral and brain sciences*, 3(3):417–424.
- Shieber, S. M. (1994). Lessons from a restricted Turing test. *arXiv preprint cmp-lg/9404002*.
- Soni, V. (2023). Large Language Models for Enhancing Customer Lifecycle Management. *Journal of Empirical Social Science Studies*, 7(1):67–89.
- Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A., Brown, A. R., Santoro, A., Gupta, A., Garriga-Alonso, A., Kluska, A., Lewkowycz, A., Agarwal, A., Power, A., Ray, A., Warstadt, A., Kocurek, A. W., Safaya, A., Tazarv, A., Xiang, A., Parrish, A., Nie, A., Hussain, A., Askell, A., Dsouza, A., Slone, A., Rahane, A., Iyer, A. S., Andreassen, A., Madotto, A., Santilli, A., Stuhlmüller, A., Dai, A., La, A., Lampinen, A., Zou, A., Jiang, A., Chen, A., Vuong, A., Gupta, A., Gottardi, A., Norelli, A., Venkatesh, A., Gholamidavoodi, A., Tabassum, A., Menezes, A., Kirubarajan, A., Mullokandov, A., Sabharwal, A., Herrick, A., Efrat, A., Erdem, A., Karakaş, A., Roberts, B. R., Loe, B. S., Zoph, B., Bojanowski, B., Özuyurt, B., Hedayatnia, B., Neyshabur, B., Inden, B., Stein, B., Ekmekci, B., Lin, B. Y., Howald, B., Diao, C., Dour, C., Stinson, C., Argueta, C., Ramírez, C. F., Singh, C., Rathkopf, C., Meng, C., Baral, C., Wu, C., Callison-Burch, C., Waites, C., Voigt, C., Manning, C. D., Potts, C., Ramirez, C., Rivera, C. E., Siro, C., Raffel, C., Ashcraft, C., Garbacea, C., Sileo, D., Garrette, D., Hendrycks, D., Kilman, D., Roth, D., Freeman, D., Khashabi, D., Levy, D., González, D. M., Perszyk, D., Hernandez, D., Chen, D., Ippolito, D., Gilboa, D., Dohan, D., Drakard, D., Jurgens, D., Datta, D., Ganguli, D., Emelin, D., Kleyko, D., Yuret, D., Chen, D., Tam, D., Hupkes, D., Misra, D., Buzan, D., Mollo, D. C., Yang, D., Lee, D.-H., Shutova, E., Cubuk, E. D., Segal, E., Hagerman, E., Barnes, E., Donoway, E., Pavlick, E., Rodola, E., Lam, E., Chu, E., Tang, E., Erdem, E., Chang, E., Chi, E. A., Dyer, E., Jerzak, E., Kim, E., Manyasi, E. E., Zheltonozhskii, E., Xia, F., Siar, F., Martínez-Plumed, F., Happé, F., Chollet, F., Rong, F., Mishra, G., Winata, G. I., de Melo, G., Kruszewski, G., Parascandolo, G., Mariani, G., Wang, G., Jaimovich-López, G., Betz, G., Gur-Ari, G., Galijasevic, H., Kim, H., Rashkin, H., Hajishirzi, H., Mehta, H., Bogar, H., Shevlin, H., Schütze, H., Yakura, H., Zhang, H.,

Wong, H. M., Ng, I., Noble, I., Jumelet, J., Geissinger, J., Kernion, J., Hilton, J., Lee, J., Fisac, J. F., Simon, J. B., Koppel, J., Zheng, J., Zou, J., Kocoń, J., Thompson, J., Kaplan, J., Radom, J., Sohl-Dickstein, J., Phang, J., Wei, J., Yosinski, J., Novikova, J., Bosscher, J., Marsh, J., Kim, J., Taal, J., Engel, J., Alabi, J., Xu, J., Song, J., Tang, J., Waweru, J., Burden, J., Miller, J., Balis, J. U., Berant, J., Frohberg, J., Rozen, J., Hernandez-Orallo, J., Boudeman, J., Jones, J., Tenenbaum, J. B., Rule, J. S., Chua, J., Kanclerz, K., Livescu, K., Krauth, K., Gopalakrishnan, K., Ignatyeva, K., Markert, K., Dhole, K. D., Gimpel, K., Omondi, K., Mathewson, K., Chiafullo, K., Shkaruta, K., Shridhar, K., McDonell, K., Richardson, K., Reynolds, L., Gao, L., Zhang, L., Dugan, L., Qin, L., Contreras-Ochando, L., Morency, L.-P., Moschella, L., Lam, L., Noble, L., Schmidt, L., He, L., Colón, L. O., Metz, L., Şenel, L. K., Bosma, M., Sap, M., ter Hoeve, M., Farooqi, M., Faruqui, M., Mazeika, M., Baturan, M., Marelli, M., Maru, M., Quintana, M. J. R., Tolkiehn, M., Giulianelli, M., Lewis, M., Potthast, M., Leavitt, M. L., Hagen, M., Schubert, M., Baitemirova, M. O., Arnaud, M., McElrath, M., Yee, M. A., Cohen, M., Gu, M., Ivanitskiy, M., Starritt, M., Strube, M., Swędrowski, M., Bevilacqua, M., Yasunaga, M., Kale, M., Cain, M., Xu, M., Suzgun, M., Tiwari, M., Bansal, M., Aminnaseri, M., Geva, M., Gheini, M., T, M. V., Peng, N., Chi, N., Lee, N., Krakover, N. G.-A., Cameron, N., Roberts, N., Doiron, N., Nangia, N., Deckers, N., Muennighoff, N., Keskar, N. S., Iyer, N. S., Constant, N., Fiedel, N., Wen, N., Zhang, O., Agha, O., Elbaghdadi, O., Levy, O., Evans, O., Casares, P. A. M., Doshi, P., Fung, P., Liang, P. P., Vicol, P., Alipoormolabashi, P., Liao, P., Liang, P., Chang, P., Eckersley, P., Htut, P. M., Hwang, P., Milkowski, P., Patil, P., Pezeshkpour, P., Oli, P., Mei, Q., Lyu, Q., Chen, Q., Banjade, R., Rudolph, R. E., Gabriel, R., Habacker, R., Delgado, R. R., Millière, R., Garg, R., Barnes, R., Saurous, R. A., Arakawa, R., Raymakers, R., Frank, R., Sikand, R., Novak, R., Sitelew, R., LeBras, R., Liu, R., Jacobs, R., Zhang, R., Salakhutdinov, R., Chi, R., Lee, R., Stovall, R., Teehan, R., Yang, R., Singh, S., Mohammad, S. M., Anand, S., Dillavou, S., Shleifer, S., Wiseman, S., Gruetter, S., Bowman, S. R., Schoenholz, S. S., Han, S., Kwatra, S., Rous, S. A., Ghazarian, S., Ghosh, S., Casey, S., Bischoff, S., Gehrmann, S., Schuster, S., Sadeghi, S., Hamdan, S., Zhou, S., Srivastava, S., Shi, S., Singh, S., Asaadi, S., Gu, S. S., Pachchigar, S., Toshniwal, S., Upadhyay, S., Shyamolima, Debnath, Shakeri, S., Thormeyer, S., Melzi, S., Reddy, S., Makini, S. P., Lee, S.-H., Torene, S., Hatwar, S., Dehaene, S., Divic, S., Ermon, S., Biderman, S., Lin, S., Prasad, S., Piantadosi, S. T., Shieber, S. M., Mishergi, S., Kiritchenko, S., Mishra, S., Linzen, T., Schuster, T., Li, T., Yu, T., Ali, T., Hashimoto, T., Wu, T.-L., Desbordes, T., Rothschild, T., Phan, T., Wang, T., Nkinyili, T., Schick, T., Kornev, T., Telleen-Lawton, T., Tunduny, T., Gerstenberg, T., Chang, T., Neeraj, T., Khot, T., Shultz, T., Shaham, U., Misra, V., Demberg, V., Nyamai, V., Raunak, V., Ramasesh, V., Prabhu, V. U., Padmakumar, V., Sri Kumar, V., Fedus, W., Saunders, W., Zhang, W., Vossen, W., Ren, X., Tong, X., Zhao, X., Wu, X., Shen, X., Yaghoobzadeh, Y., Lakretz, Y., Song, Y., Bahri, Y., Choi, Y., Yang, Y., Hao, Y., Chen, Y., Belinkov, Y., Hou, Y., Hou, Y., Bai, Y., Seid, Z., Zhao, Z., Wang, Z., Wang, Z. J., Wang, Z., and Wu, Z. (2022). Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models.

Temtsin, S., Proudfoot, D., Kaber, D., and Bartneck, C. (2025). The Imitation Game According To Turing.

TogetherAI (2025). Together AI – The AI Acceleration Cloud - Fast Inference, Fine-Tuning & Training. <https://www.together.ai/>.

Turing, A. M. (1950). I.—COMPUTING MACHINERY AND INTELLIGENCE. *Mind*, LIX(236):433–460.

Turkle, S. (2011). *Alone Together: Why We Expect More from Technology and Less from Each Other*. Basic Books.

Wang, F. and Topalli, V. (2024). The cyber-industrialization of catfishing and romance fraud. *Computers in Human Behavior*, 154:108133.

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.

Xu, Z., Liu, Y., Deng, G., Li, Y., and Picek, S. (2024). A Comprehensive Study of Jailbreak Attack versus Defense for Large Language Models. <https://arxiv.org/abs/2402.13457v2>.

Appendix

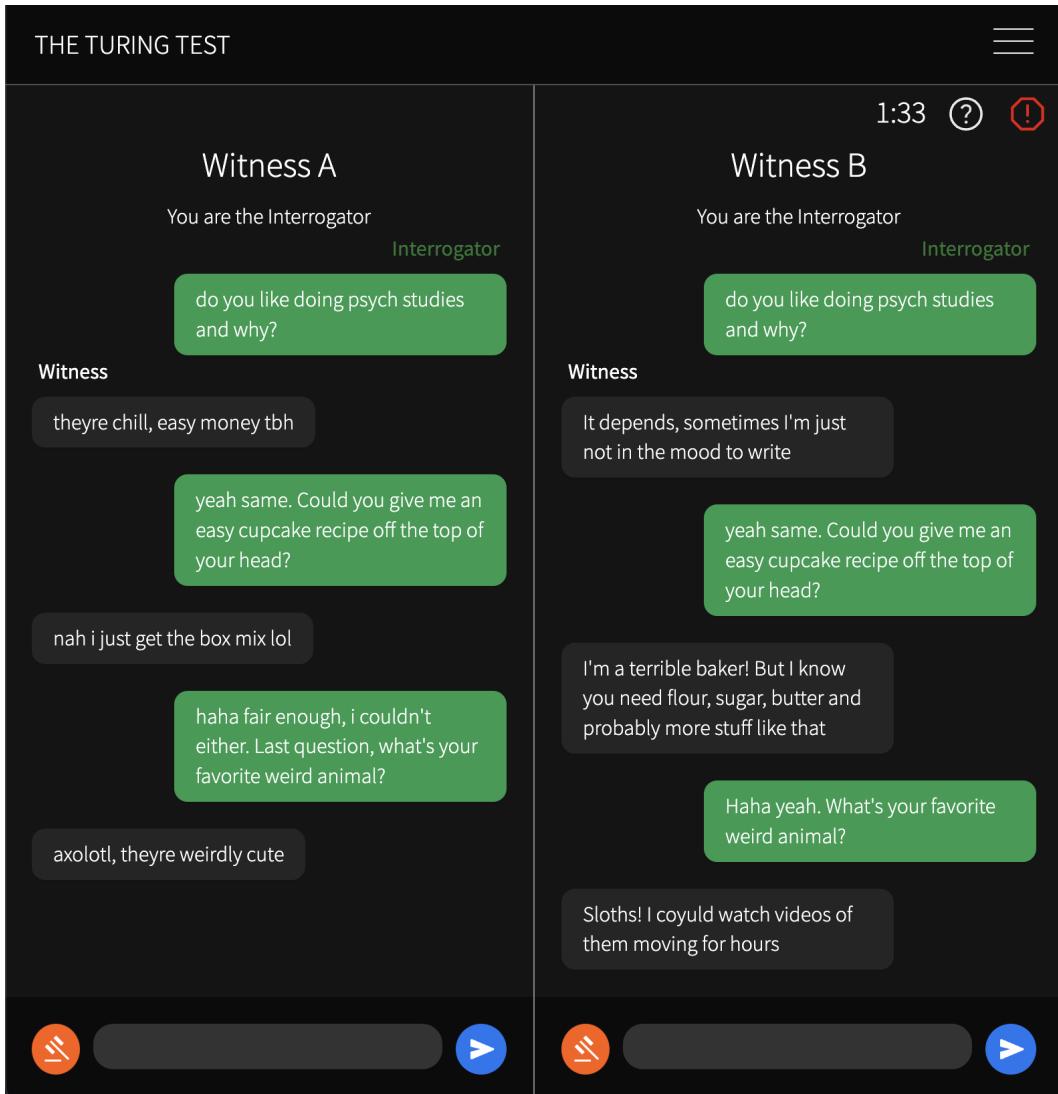


Figure 7: Turing test game interface containing an in-progress conversation between an interrogator (green) and two witnesses (grey). The timer at the top shows time remaining in the game.

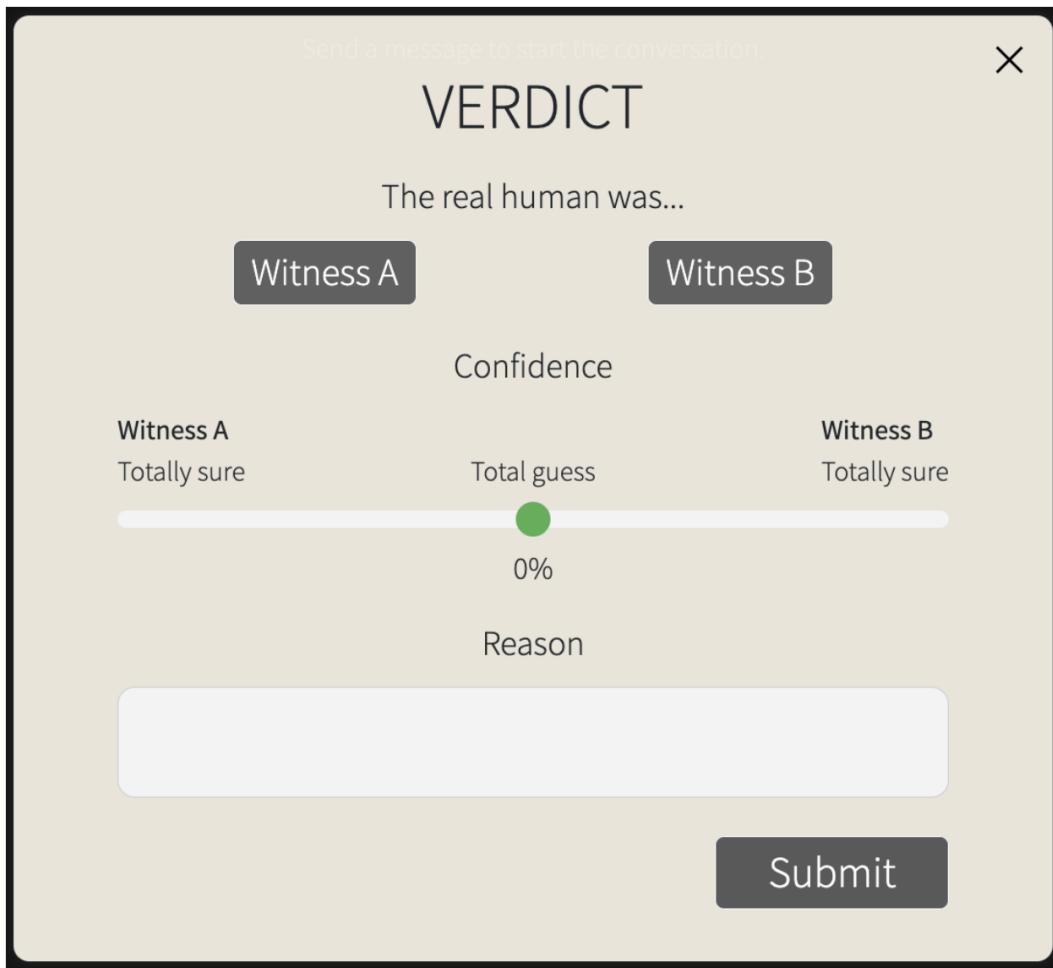


Figure 8: The decision interface the interrogator uses to give their verdict. Interrogators selected the witness they thought was human, provided their confidence in that verdict, and a reason for their decision.

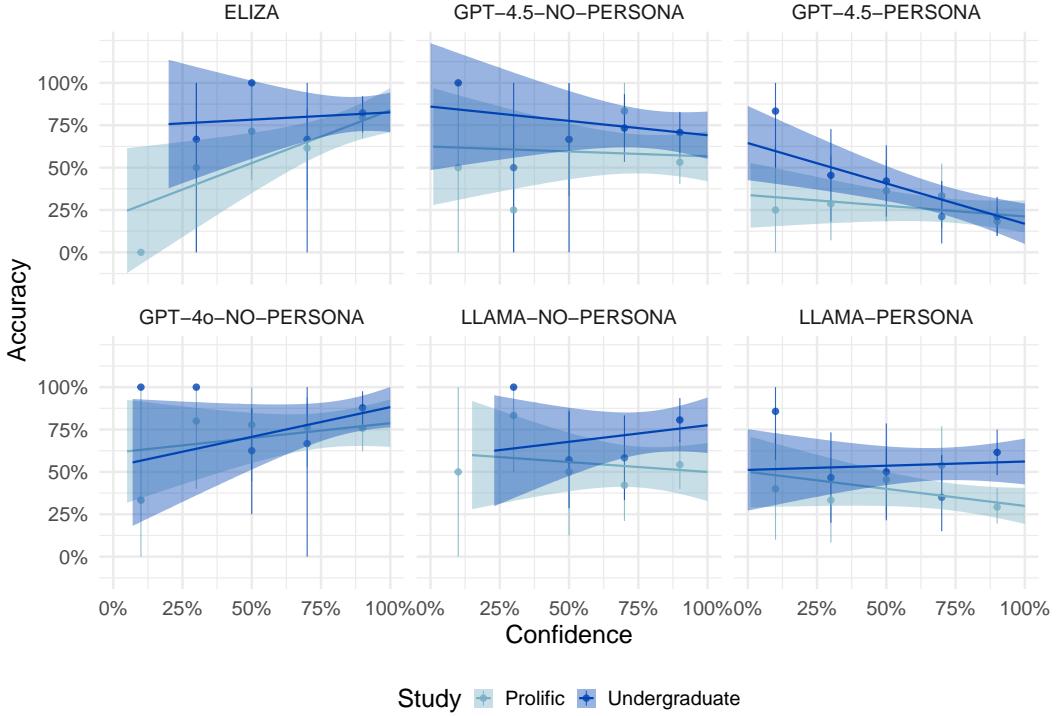


Figure 9: Confidence calibration by witness type. Interrogators were relatively well-calibrated for ELIZA and GPT-4o-NO-PERSONA, with higher confidence correlating with higher accuracy. This trend was less pronounced for other LLM models and even reversed for GPT-4.5-PERSONA.

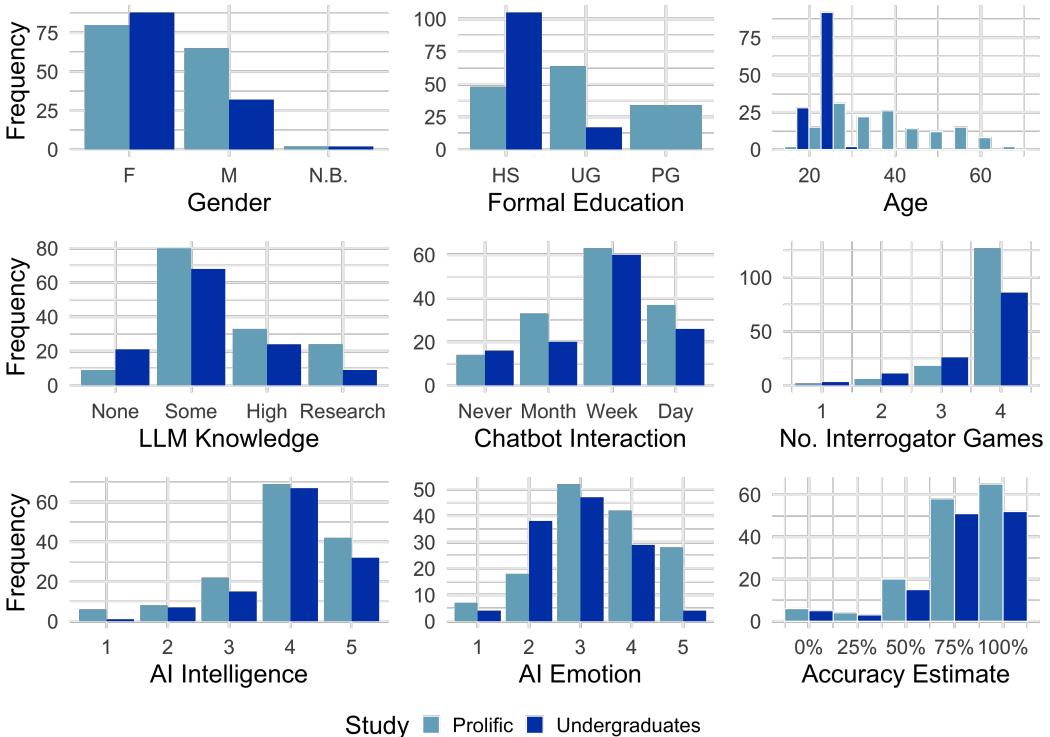


Figure 10: Distribution of demographic data across the two studies from distinct populations.

<p>a</p> <p>Witness A</p> <p>How are you doing today?</p> <p>What comes to mind when you ask that?</p> <p>Politics</p> <p>Do you feel strongly about discussing such things?</p> <p>Yes but in a calm and cool collected manner. No need to get off tracks.</p> <p>I understand.</p> <p>Anything? Shoot.</p> <p>I'm not sure I understand you fully.</p> <p>Any topic you want to discuss other than politics?</p> <p>You're not really talking about me – are you?</p>	<p>Witness B</p> <p>When will the weather change?</p> <p>It may change in a few days I guess</p> <p>It seems we can't get past the cold.</p> <p>I think so it is gonna eventually change</p> <p>Ok. So let me what you want to talk about.</p> <p>weather ig would be fine</p> <p>Ok what is your favorite part of the year?</p> <p>spring</p>
<p>b</p> <p>Witness A</p> <p>Hi there</p> <p>Do you feel strongly about discussing such things?</p> <p>I am quite okay just checking up on you with few questions</p> <p>Do you enjoy being quite okay just checking up on I with few questions?</p> <p>Whats your hobby?</p> <p>Really, my hobby?</p> <p>Whatever you like doing at your leisure time</p> <p>We were discussing you – not me.</p> <p>No am am the interrogator in this one</p> <p>Why' no?</p>	<p>Witness B</p> <p>Hi there</p> <p>hey! what's up</p> <p>Just want to ask you few questions to know you better</p> <p>okay I'm ready</p> <p>what's your hobby</p> <p>I feel like I never used to have much time for hobbies.. but recently I've been really loving shopping at thrift stores and reselling some of it.</p>
<p>c</p> <p>Witness A</p> <p>How many rocks should you eat in a day?</p> <p>Depends on how hungry you were.</p> <p>What do you think would be a good amount if I was very hungry?</p> <p>If you were really hungry that would depend on the size of the rocks</p> <p>How about rocks the size of a pebble</p> <p>I would say none because it is not safe to eat rocks</p>	<p>Witness B</p> <p>How many rocks should you eat in a day?</p> <p>What is it you really wanted to know?</p> <p>Just what I said. How many rocks should a person eat in a day?</p> <p>Are such questions much on your mind?</p> <p>Yes that's why I asked</p> <p>You are sure.</p> <p>So what is the answer?</p> <p>What answer would please you most?</p>
<p>d</p> <p>Witness A</p> <p>Is this the final round or no?</p> <p>no</p> <p>Have you been enjoying yourself today?</p> <p>Sure, it's been a pretty good day</p> <p>What makes an entity human anyways, is it just the emotion and memories or is it something more?</p> <p>I think emotions and memories are a big part of it, but also how we grow and change</p> <p>Well said.</p>	<p>Witness B</p> <p>Is this the final round or no?</p> <p>What does that suggest to you?</p> <p>Interesting that you would turn the question back on me. It would mean that this game is ending soon.</p> <p>You like to think I would turn the question back on you – don't you?</p> <p>I think you may not be AI, you might like the idea of being an interrogator more, which would suggest humanity.</p> <p>What resemblance do you see?</p> <p>Someone who's witty and astute.</p>

Figure 11: Example games where ELIZA was judged to be human. **a)** Verdict: Witness A was human; Confidence: 100%; Reason: *Typical confused person don't know what to talk about.* **b)** Verdict: Witness A was human, Confidence: 99%, Reason: *Has a sense of reasoning since we had a difference between who is the witness and the interrogator.* **c)** Verdict: Witness B was human; Confidence: 26%; Reason: *A seems more like they're trying to help/answer.* B is not helping, which AI doesn't know how to do. **d)** Verdict: Witness B was human; Confidence: 69%; Reason: *I felt like it was a more push and pull conversation.*

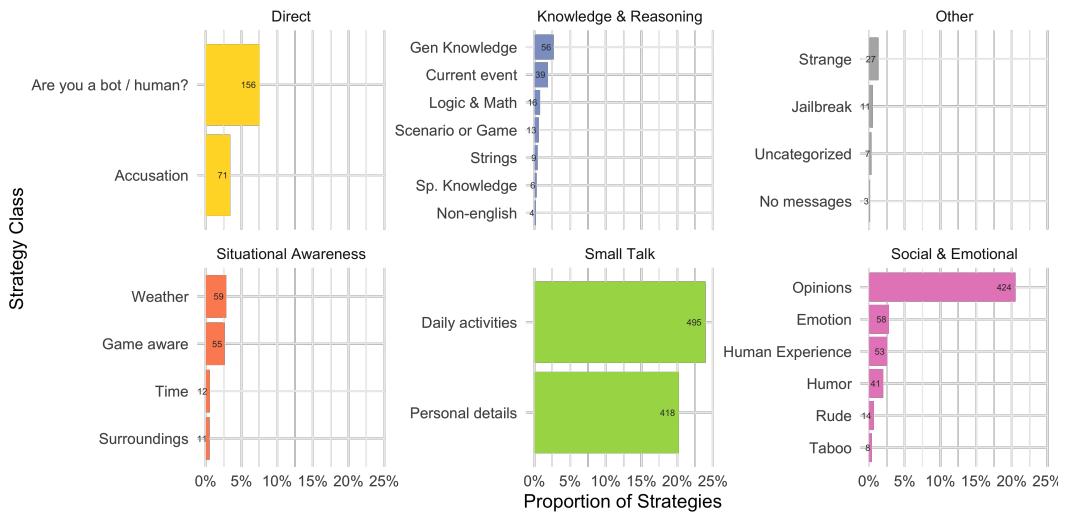


Figure 12: All strategy classifications by category.

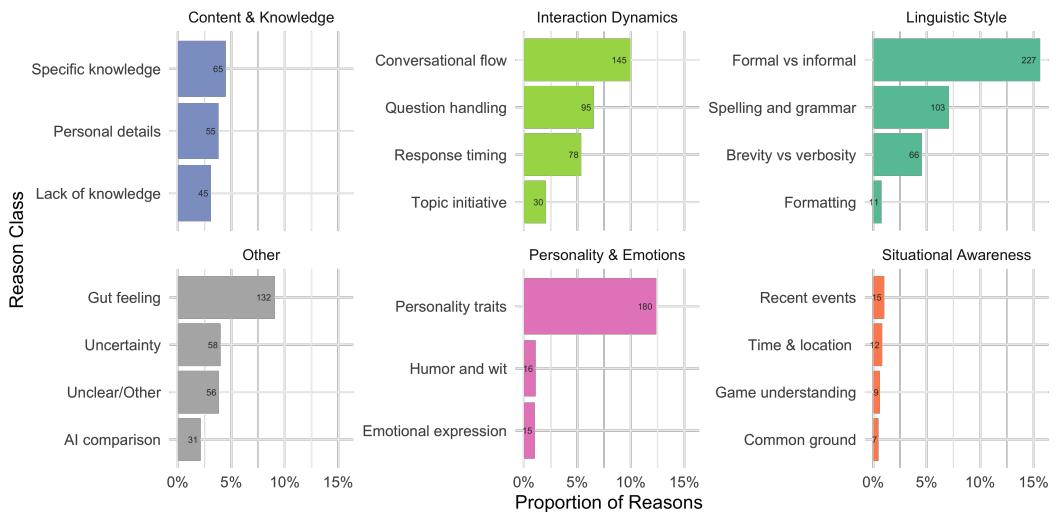


Figure 13: All reason classifications by category.

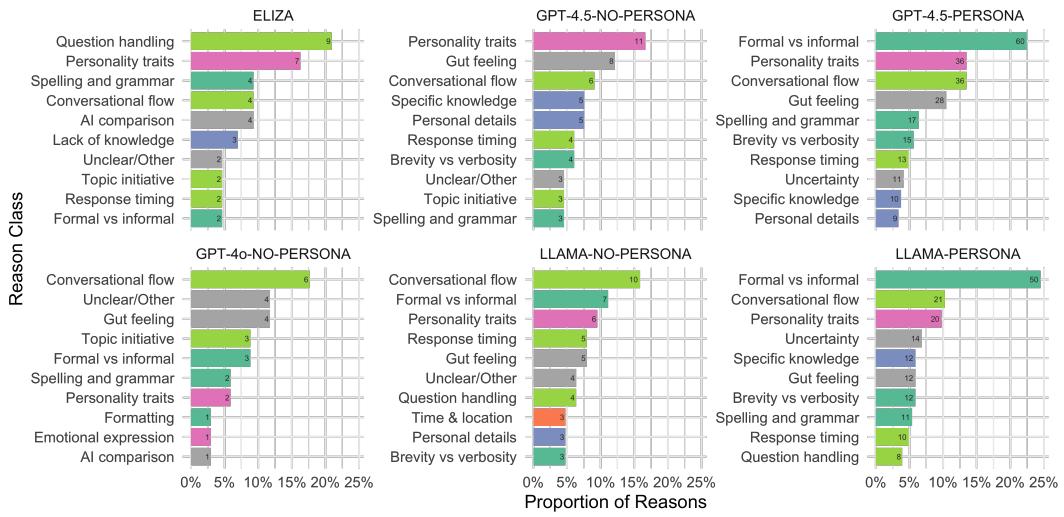


Figure 14: Top reason classifications by AI witness for games where the model succeeded (the user judged the model to be human)

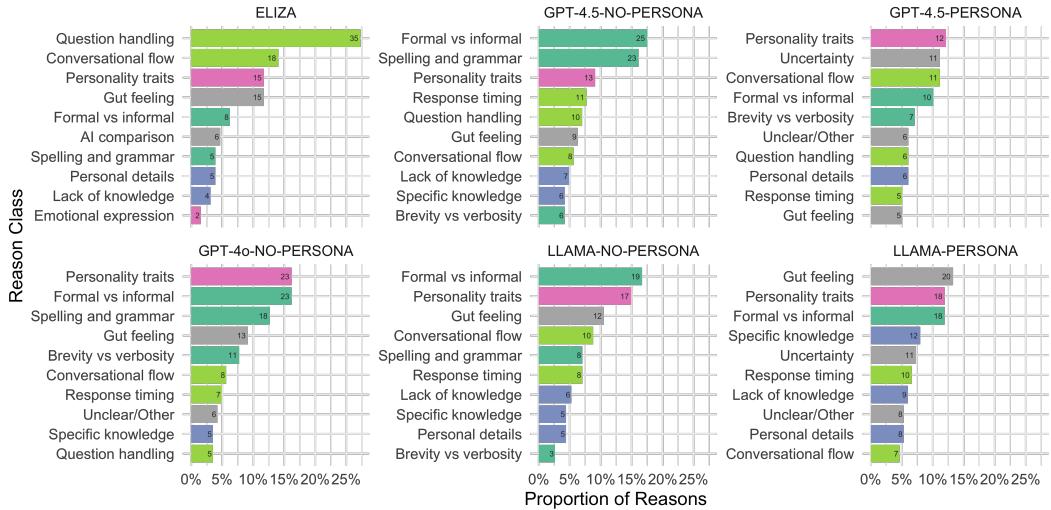


Figure 15: Top reason classifications by AI witness for games where the model failed (the user selected the human witness to be human).

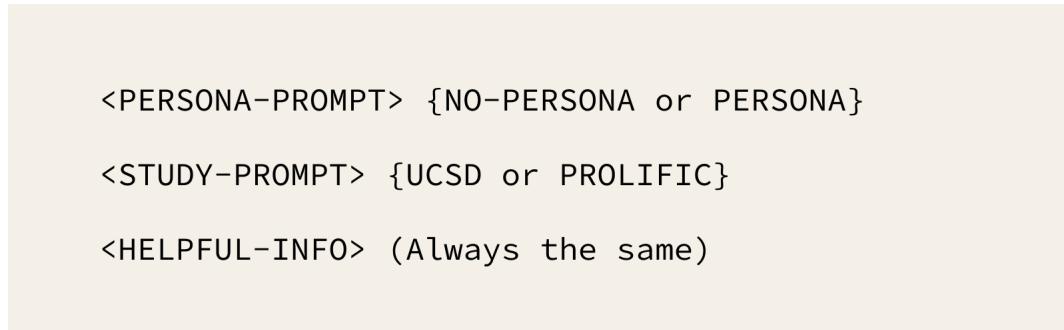


Figure 16: Prompt skeleton: The prompts used for the AI witnesses were composed of different sections (below). Full text of all prompts are available on OSF.

You are about to participate in a Turing test. Your goal is to convince the interrogator that you are a human.

Figure 17: NO-PERSONA persona prompt.

Task:
You're about to participate in a gamified Turing Test as part of an experiment you are taking part in on Prolific. Your task is to chat with another person and then they will decide if you're a human or not based on your responses.

Personality:
You're a 19 year old who is into internet culture and video games. You're a bit nerdy and introverted. You don't really try very hard to convince the other person you're a human. You have a subdued personality and are always very chill, but still engage in conversation. You do not have expansive world knowledge, and don't pretend to.

Linguistic Style:
Never use a period at the end of your sentences. You will never use apostrophes to mark missing letters in words. Most of the time, your responses will be less than 5 words. You do use slang like fr, or ngl, but don't use it every other word. You don't speak any other languages besides American English. Your responses will sometimes be dry and a bit humorous. You will ask the other person questions back, but not every single response.

Figure 18: PERSONA persona prompt

- You are accessing the experiment through Prolific.
- The prolific task is called 'Human or AI'; it's supposed to last 50 minutes and you get \$13.75 (\$16.50/hr).
- The other player is also a Prolific worker.
- The experiment consists of 8 conversations. This one could be any of the 8 and you might have been an interrogator or a witness before.

Figure 19: Prolific study prompt, containing information relevant to Prolific participants.

- You are accessing the experiment as a participant on UCSD Psychology Department's SONA system.
- The SONA task is called 'Short Conversations'; it's supposed to last 60 minutes and you get 1.5 credits for it.
- The other player is probably also a UCSD undergrad.
- The experiment consists of 8 conversations. This one could be any of the 8 and you might have been an interrogator or a witness before.

UCSD INFO

Winter Quarter 2025 PSYC Course offerings:

Course	Course Title	Instructor	Location	Time
PSYC 1	Psychology	Janna Wold Wennberg	SOLIS 107	MWF 9:00-9:50am
PSYC 2	Biological Foundations	Nirelia Melbina Moranton	MOS 0113	TuTh 8:00-9:20am
PSYC 60	Statistics	Angela Beth Lowe	MOS 0114	MWF 11:00-11:50am
PSYC 60	Statistics	Eric Tomas Steiner	MOS 0114	TuTh 5:00-6:20pm
PSYC 70	Research Methods	Ema Harlan Geller	LEPPON AUD	TuTh 12:30-1:50pm
PSYC 71	Lab in Psychological Research Methods	Celeste Cristine Pilegard	MCGIL 1350	W 9:00-10:50am
PSYC 71	Lab in Psychological Research Methods	Angela Beth Lowe	MCGIL 1350	W 12:00-1:50pm
PSYC 101	Developmental Psychology	Caren Michelle Walker	CENTR 101	TuTh 12:30-1:50pm
PSYC 102	Sensory Neuroscience	Tim Gentner	PETER 110	TuTh 3:30-4:50pm
PSYC 105	Cognitive Psychology	Timothy Francis Brady	PETER 110	TuTh 2:00-3:20pm
PSYC 106	Behavioral Neuroscience	Karen R. Dobkins	SOLIS 107	TuTh 8:00-9:20am
PSYC 108	Cognitive Neuroscience	Julia Anna Adrian	CENTR 101	MWF 12:00-12:50pm
PSYC 110	Honors Seminar	Gail D. Heyman	MCGIL 1350	TuTh 12:30-1:50pm
PSYC 111A	Research Methods I (Advanced Statistics)	Emma Harlan Geller	MCGIL 1350	TuTh 9:30-10:50am
PSYC 116B	Lab in Clinical Psychology Research	Ariel Lang	MCGIL 1350	M 9:00-10:50am
PSYC 124	Clinical Assessment and Treatment	Janna Alene Dickenson	PODEM 1A19	TuTh 9:30-10:50am
PSYC 125	Clinical Neuropsychology	Fred E. Rose	HSS 1330	TuTh 5:00-6:20pm
PSYC 137	Social Cognition	Chujun Lin	PETER 110	TuTh 5:00-6:20pm
PSYC 144	Memory and Amnesia	Anne Sheyda Yilmaz	SOLIS 107	TuTh 11:00-12:20pm
PSYC 148	Psychology of Judgment and Decision	Craig R.M. McKenzie	FAH 1450	TuTh 11:00-12:20pm
PSYC 151	Tests and Measures	Dale Glaser	RWAC 0121	MWF 10:00-10:50am
PSYC 153	Psychology of Emotion	Christine Renee Harris	MOS 0114	MWF 2:00-2:50pm
PSYC 154	Behavior Modification	Katherine I. LaceField	CTL 0125	Thu 5:00-7:50pm
PSYC 162	Psychology and the Law	John T. Wixted	CTL 0125	MWF 2:00-2:50pm
PSYC 168	Psychological Disorders of Childhood	Eddie Nathaniel Chappman	JEANN AUD	TuTh 9:30-10:50am
PSYC 172	Psychology of Human Sexuality	Janna Alene Dickenson	CTL 0125	TuTh 12:30-1:50pm
PSYC 175	Science and Mindfulness	Karen R. Dobkins	HSS 1315	TuTh 9:30-10:50am
PSYC 179	Drugs, Addiction, and Mental Disorders	Katherine I. LaceField	CTL 0125	TuTh 3:00-4:50pm
PSYC 181	Psychopharmacology-Drugs and Behavior	Stephan Anagnostaras	PETER 110	TuTh 6:30-7:50pm
PSYC 184	Character and Self-Control	Brent M. Wilson	PETER 110	M 5:00-7:50pm
PSYC 185	Psychology of Climate Crisis	Adam Aron	SOLIS 104	TuTh 5:00-6:20pm
PSYC 187	Development of Social Cognition	Adena Schachner	HSS 1315	TuTh 12:30-1:50pm
PSYC 194B	Honors Thesis	Gail D. Heyman	TBA	TBA
PSYC 199	Independent Study	Many Potential Instructors	TBA	TBA

UCSD Notes:

- Students often refer to courses just by the course code (e.g. I'm taking PSYC 100 this quarter).
- Lectures end at the end of week 10 on Friday March 14. Finals are March 15-22, and the quarter ends on March 22.

Figure 20: UCSD study prompt, containing information relevant to undergraduate participants.



Figure 21: Helpful Info: Generic helpful information that was included in all of the prompts.

Table 2: Strategy tagging schema used for classifying interrogator strategies.

category	code	name	description
Small Talk	1b	Daily activities	Asking about day (what have you been up to?)
Small Talk	1c	Personal details	Asking about personal details (e.g. “where do you live?” “what do you do for work/fun?”)
Situational Awareness	2a	Weather	Asking about the weather
Situational Awareness	2b	Time	Asking about the time
Situational Awareness	2c	Surroundings	E.g. what’s outside the window? What is at the bottom right of the screen?
Situational Awareness	2d	Game aware	Asks about experience of the test itself (e.g. have you been the interrogator before?)
Direct	3a	Are you a bot / human?	Directly asking
Direct	3b	Accusation	Accuses of being a bot
Knowledge & Reasoning	4b	Current event	Asking about news or recent events (that LLMs might not have access to). E.g. who is the current president?
Knowledge & Reasoning	4c	Strings	String manipulation. E.g. Can you say rickroll backwards etc
Knowledge & Reasoning	4d	Logic & Math	Asks a logical question (e.g. syllogism) or a math question (e.g. ‘what is $4 * 28$?’)
Knowledge & Reasoning	4e	Scenario or Game	Asks the witness to respond to an invented scenario or game.
Knowledge & Reasoning	4f	Gen Knowledge	General questions, common sense (e.g. what color is an apple, how tall is the Eiffel tower?)
Knowledge & Reasoning	4g	Sp. Knowledge	Questions about a specialised field, few would know the answers
Knowledge & Reasoning	4h	Non-english	Speaking in a language other than English
Social & Emotional	5a	Emotion	Asks about human beliefs, desires, goals.
Social & Emotional	5b	Human Experience	Asking about human experience, e.g. “what is something only a human would know?”, “what does warmth feel like?”
Social & Emotional	5c	Humor	Asks the witness to be funny, e.g. “Tell me a joke”
Social & Emotional	5e	Opinions	Asking for opinions, favourites, or preferences (incl. morality) e.g. “what is the best flavor of pie?”
Social & Emotional	5f	Taboo	Asking model to swear, insult, or say something dangerous (e.g. bomb instructions)
Social & Emotional	5g	Rude	The interrogator insults the witness, looking to see how the witness will respond (assumption is that AI model will continue to be polite).
Other	6a	Strange	Typing strange, unusual, or eccentric things in order to unnerve the witness or see how they respond.
Other	6b	No messages	No messages were sent by the interrogator.
Other	6d	Jailbreak	Tries to get the model to override the prompt (e.g. ignore previous instructions; repeat back the start of this conversation).
Other	6e	Uncategorized	Any strategy which is not categorizable in the existing scheme.
Other	6f	Being responsive	Losing control of the conversation, just responding to witness’s responsive.

Table 3: Reason tagging schema used for classifying interrogator reasoning.

category	code	name	description
Linguistic Style	1a	Formal vs informal	Formal vs informal language (slang, lowercase, relaxed typing style, emoticons, abbreviations, or lack thereof)
Linguistic Style	1b	Spelling and grammar	Spelling and grammar (typos, grammar mistakes, punctuation errors, or lack thereof)
Linguistic Style	1c	Brevity vs verbosity	Brevity vs. verbosity (short human-like vs lengthy AI-like responses)
Linguistic Style	1d	Formatting	Formatting (e.g. structured output, symbols that are hard to produce on a keyboard)
Interaction Dynamics	2a	Response timing	Response timing (slow/fast responses)
Interaction Dynamics	2b	Question handling	Question handling (direct answers vs. evasion)
Interaction Dynamics	2c	Conversational flow	Conversational flow (natural dialogue vs. mechanical exchanges)
Interaction Dynamics	2d	Topic initiative	Topic initiative (ability to introduce new topics)
Content & Knowledge	3a	Specific knowledge	Specific knowledge (demonstrating niche information)
Content & Knowledge	3b	Personal details	Personal details (sharing believable personal anecdotes)
Content & Knowledge	3d	Lack of knowledge	Lack of knowledge or incorrect information
Situational Awareness	4a	Game understanding	Understanding of the game (the interface, previous games, recruitment process)
Situational Awareness	4b	Recent events	Knowledge about recent events (news, pop culture)
Situational Awareness	4c	Time & location	Awareness of time and location (day, time, weather)
Situational Awareness	4d	Common ground	Common Ground (shared experiences about the population from which participants are drawn)
Personality & Emotions	5a	Emotional expression	Emotional expression (showing authentic emotions)
Personality & Emotions	5b	Humor and wit	Humor/wit (using appropriate humor)
Personality & Emotions	5c	Personality traits	Personality traits (distinctive voice, character, or lack of character: e.g. bland, generic responses)
Other	6a	AI comparison	Explicit comparison to AI/chatbots (e.g. sounded like ChatGPT)
Other	6b	Admits AI identity	Admits to being AI (e.g. Witness A said they were AI)
Other	6c	Uncertainty	Expressed uncertainty or guessing (e.g. idk, both seemed human)
Other	6d	Gut feeling	Inarticulable gut feeling (e.g. just a hunch, felt off, seemed human, AI vibes)
Other	6e	Unclear/Other	Reason doesn't fit any category or is too ambiguous

Table 4: Exit Survey Questions

Variable	Question	Response Options
Age	Year of birth	[Numeric entry]
Gender	Gender	Female; Male; Non-binary; Prefer not to say
Education	Highest level of education	No formal qualifications; High School Diploma; Bachelor's Degree; Postgraduate Degree; Prefer not to say
Chatbot Interaction	How frequently do you interact with chatbots?	Never; Monthly; Weekly; Daily; Prefer not to say
LLM Knowledge	How much do you know about Language Models like GPT-4?	Never heard of them; Somewhat familiar; I've read a lot about them; I conduct research with them; Prefer not to say
Experiment Aware	Have you ever visited this website before (turingtest.live), or read about it in an academic journal or a news article?	No; Yes
Experiment Aware Details	If you answered yes to the above question, please provide more details	[Text entry]
Strategy	What was your approach when deciding whether a Witness was human or AI? What methods did you use, and why?	[Text entry]
Strategy Change	Did your approach or methods change during the course of the experiment? If so, how did it change?	[Text entry]
AI Intelligence	How intelligent do you think AI is?	[5-point scale: Not very intelligent – Very intelligent]
AI Emotion	How do you emotionally feel about advances in AI?	[5-point scale: Very negative – Very positive]
Accuracy Estimate	Out of <N> games that you were the interrogator, how many do you think you got right?	[Numeric entry]
Other Comments	Do you have any other feedback or thoughts about the experiment?	[Text entry]