

## Review

*Kris Carlson prompting Adobe Acrobat Pro AI Assistant*

*Can you review this article, culling out in full the prompt that was used (give them all if there's more than one), and summarize their analyses of what underlies the different levels of human impersonation achieved by the 4 models, and notably, the superior impersonation of ChatGPT-4.5? Critique their methodology. Summarize any discussion they give of the dangers due to imperonation, such as AI persuasiveness.*

## Large Language Models Pass the Turing Test

Cameron R. Jones and Benjamin K. Bergen

The document reports on a study evaluating the performance of various large language models in passing a three-party Turing test, highlighting the capabilities of GPT-4.5 and LLaMa-3.1 in imitating human conversation.

### Prompts Used in the Study:

The study used two types of prompts for the AI models:

#### 1. **NO-PERSONA Prompt:**

- Instructions: "You are about to participate in a Turing test. Your goal is to convince the interrogator that you are a human."
- Additional Context: Included instructions provided to human participants, information about the game interface, recent news stories after the model's training cutoff, and common ground information (e.g., compensation for Prolific participants or class schedules for UCSD undergraduates).

#### 2. **PERSONA Prompt:**

- Instructions: Models were prompted to adopt the persona of a young, introverted individual knowledgeable about internet culture and using slang.
- Full Text: The PERSONA prompt for the Prolific study is detailed in Figure 6 of the document, and other prompts are available in the Appendix.

### Analysis of Human Impersonation Levels:

The study evaluated four AI systems (GPT-4.5, LLaMa-3.1-405B, GPT-4o, and ELIZA) in a three-party Turing test. Key findings include:

#### 1. **Performance of Models:**

- **GPT-4.5-PERSONA:** Achieved a win rate of 73%, significantly outperforming human participants and other models. It was judged to be human more often than actual human witnesses.
- **LLaMa-PERSONA:** Achieved a win rate of 56%, comparable to human witnesses but less consistent across populations.
- **NO-PERSONA Models** (GPT-4.5-NO-PERSONA and LLaMa-NO-PERSONA): Performed worse, with win rates of 36% and 38%, respectively.
- **Baseline Models** (GPT-4o-NO-PERSONA and ELIZA): Had the lowest win rates (21% and 23%), significantly below chance.

## 2. Factors Underlying Superior Impersonation by GPT-4.5:

- **Prompting:** The PERSONA prompt significantly enhanced performance by instructing the model to adopt a specific humanlike persona, emphasizing linguistic style, slang, and cultural knowledge.
- **Flexibility:** GPT-4.5 demonstrated adaptability in responding to interrogators, leveraging its ability to mimic human conversational dynamics and emotional expression.
- **Comparison to Humans:** In the three-party setup, GPT-4.5-PERSONA not only appeared plausibly human but outperformed real humans in direct comparisons, suggesting it excelled in conversational flow and interactional dynamics.

## 3. Strategies and Reasons for Interrogator Decisions:

- Interrogators often relied on linguistic style (e.g., slang, typos) and interactional dynamics (e.g., conversational flow, question handling) to make judgments.
- The PERSONA prompt helped GPT-4.5 appear more human by aligning its responses with social and emotional intelligence, which interrogators found compelling.

## Critique of Methodology:

### 1. Strengths:

- **Three-Party Turing Test:** The study adhered to Turing's original formulation, allowing direct comparison between AI and human witnesses.
- **Controlled Design:** Randomized assignment of AI witnesses and pre-registered hypotheses ensured methodological rigor.

- **Replication:** Conducting the study on two independent populations (undergraduates and Prolific workers) strengthened the reliability of findings.

## 2. Weaknesses:

- **Prompt Dependency:** The study's conclusion that GPT-4.5 passes the Turing test relies heavily on the PERSONA prompt. This raises questions about whether the model itself or the human-designed prompt is responsible for its success.
- **Short Interaction Time:** The 5-minute time limit may not fully capture the complexity of human-AI interactions in longer conversations.
- **Population Bias:** The study primarily involved lay participants, which may not generalize to experts who might be better at detecting AI.
- **Limited Scope:** The study focused on conversational impersonation but did not explore other dimensions of humanlike intelligence, such as creativity or reasoning.

## Discussion of Dangers Due to Impersonation:

### 1. AI Persuasiveness:

- The ability of GPT-4.5 to convincingly impersonate humans raises concerns about its potential use in social engineering, misinformation, and manipulation.
- Models capable of robust deception could be exploited to influence opinions, spread false information, or impersonate individuals in sensitive contexts.

### 2. Counterfeit People:

- The study highlights the risk of "counterfeit people," where AI systems substitute for real humans in social and economic roles. This could undermine the value of genuine human interaction and give entities controlling these systems disproportionate power.

### 3. Social and Economic Impacts:

- AI systems that pass as human could disrupt industries reliant on human interaction, such as customer service, education, and therapy.
- The widespread use of such systems might lead to societal dependence on AI, eroding trust in human relationships.

#### 4. Detection Strategies:

- The study suggests that interrogators struggled to distinguish AI from humans, even when using strategies like asking unusual questions or attempting to "jailbreak" the models. Future work could explore training people to improve their ability to detect AI deception.

#### Conclusion:

The study provides compelling evidence that GPT-4.5, when prompted with a humanlike persona, can pass the original three-party Turing test. However, its reliance on prompting raises questions about the nature of its intelligence. The findings underscore the need for ethical considerations and safeguards to mitigate the risks posed by AI systems capable of human impersonation.

#### Key Findings About GPT-4.5's Performance:

##### 1. Superior Impersonation:

- **GPT-4.5-PERSONA** achieved a win rate of 73%, meaning it was judged to be human more often than actual human participants. This is the first robust evidence of an AI system passing the original three-party Turing test.
- It outperformed all other AI models, including LLaMa-3.1-405B, GPT-4o, and ELIZA.

##### 2. Impact of Prompting:

- The **PERSONA prompt** significantly enhanced GPT-4.5's performance by instructing it to adopt a humanlike persona (introverted, knowledgeable about internet culture, and using slang). Without this prompt (NO-PERSONA), its win rate dropped to 36%.

##### 3. Consistency Across Populations:

- GPT-4.5-PERSONA performed consistently well across two independent populations: UCSD undergraduates (69% win rate) and Prolific participants (76% win rate).

##### 4. Humanlike Interaction:

- GPT-4.5 excelled in conversational flow, linguistic style, and interactional dynamics, making it appear more human than the real human witnesses it was compared to.

##### 5. Replication of Results:

- The findings were replicated across two studies, providing strong evidence of GPT-4.5's ability to convincingly impersonate humans in short conversations.

#### **6. Limitations Without Persona:**

- Without the PERSONA prompt, GPT-4.5's performance was significantly weaker, suggesting its success is heavily dependent on tailored instructions.

These findings highlight GPT-4.5's ability to convincingly mimic human behavior when guided by specific prompts, raising important questions about its adaptability and implications for AI-human interactions.

#### **Limitations of the Turing Test for AI Evaluation:**

##### **1. Dependence on Human Judgment:**

- The Turing test relies on human interrogators to assess whether a machine is indistinguishable from a human. This introduces subjectivity and fallibility, as humans may be biased, inattentive, or overly gullible (e.g., the "ELIZA Effect").

##### **2. Focus on Deception, Not Intelligence:**

- The test measures a machine's ability to deceive humans into believing it is human, rather than directly evaluating its intelligence or reasoning capabilities. This conflates humanlike behavior with intelligence.

##### **3. Prompt Dependency:**

- AI systems like GPT-4.5 only pass the test when given tailored prompts (e.g., the PERSONA prompt). This raises questions about whether the success is due to the model itself or the human-designed instructions.

##### **4. Short Interaction Time:**

- The study used a 5-minute time limit for conversations, which may not fully capture the complexity of human-AI interactions over longer durations. Longer tests might reveal weaknesses in AI systems.

##### **5. Limited Scope:**

- The Turing test evaluates conversational impersonation but does not assess other dimensions of intelligence, such as creativity, problem-solving, or emotional understanding.

##### **6. Cultural and Contextual Bias:**

- Human interrogators may rely on cultural, social, or linguistic common ground to make judgments, which could disadvantage AI systems unfamiliar with specific contexts.

#### **7. Binary Outcome:**

- The test provides a binary pass/fail result, which oversimplifies the nuanced capabilities of AI systems. It does not account for partial success or areas where AI excels beyond human abilities.

#### **8. Vulnerability to "Cheap Tricks":**

- AI systems can exploit superficial strategies, such as mimicking typos, slang, or informal language, to appear human. These tricks may not reflect genuine intelligence or understanding.

#### **9. Changing Standards:**

- As AI systems improve, human expectations of what constitutes "humanlike" behavior may evolve, making the test a moving target.

#### **10. Ethical Concerns:**

- The test inherently involves deception, which raises ethical questions about its use as a benchmark for AI evaluation.

### **Conclusion:**

While the Turing test provides a flexible and interactive way to evaluate humanlike behavior, its limitations highlight the need for complementary benchmarks that assess broader aspects of intelligence and avoid reliance on deception or subjective human judgment.