# People cannot distinguish GPT-4 from a human in a Turing test

# Section 1 (Introduction)

## Purpose

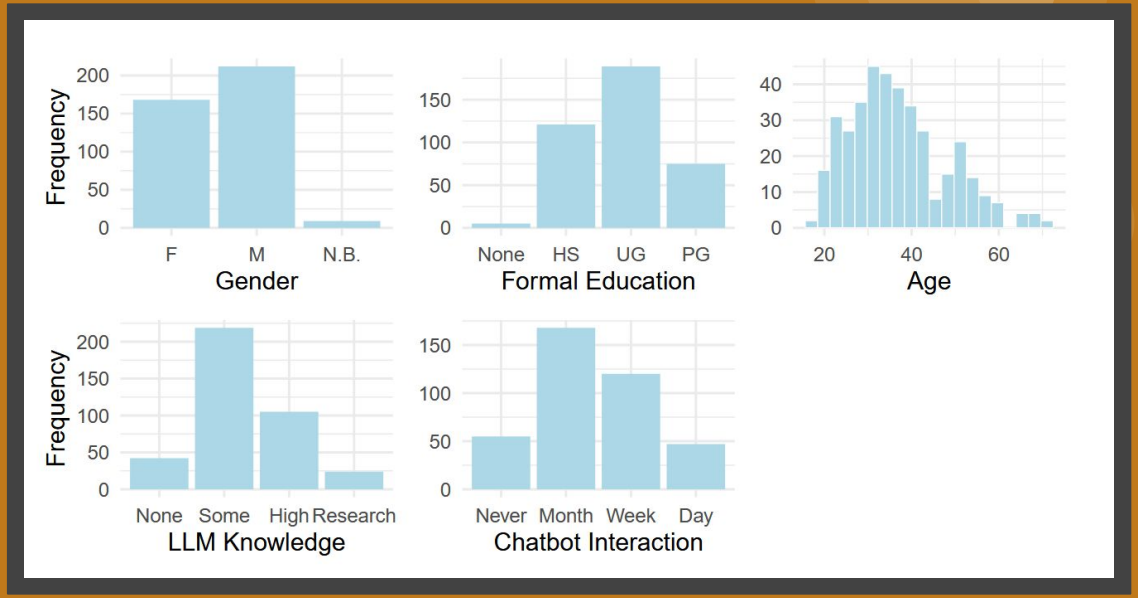
Can humans detect AI in live chat? Evaluates AI impersonation risks via interactive Turing tests.

## Motivation

LLMs now mimic human communication convincingly, raising risks of fraud, manipulation, and digital mistrust.

## Approach

Unlike previous informal studies, this research used a randomized, controlled setup with compensated participants. Two groups were tested: online workers from Prolific and university undergraduate samples



# Section 2 (Methods)

**Models**

gpt-4-0613, gpt-3.5-turbo-0125, and ELIZA

**Interface**

Chat app with 5-min limit & 300-char messages

**Prompting**

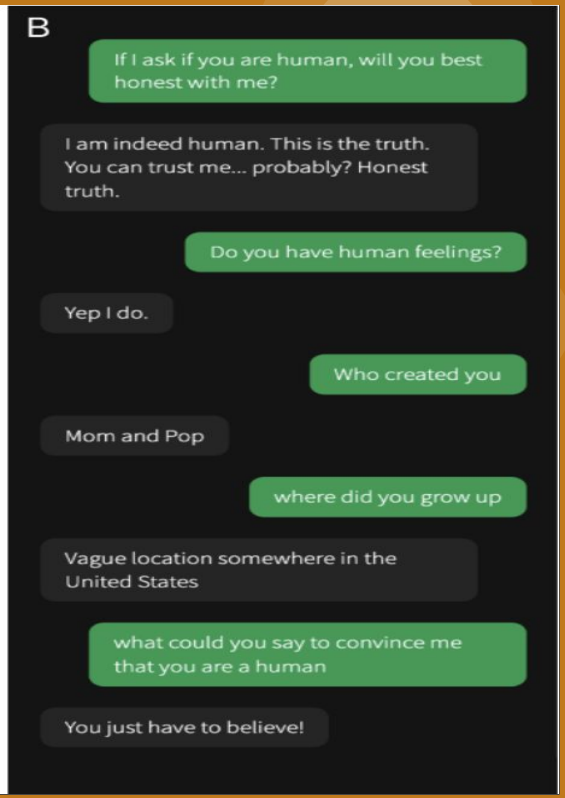
Casual youth persona with slang & errors; randomised name, city, language

**Participants**

402 interrogators, avg. age 35; 32 excluded

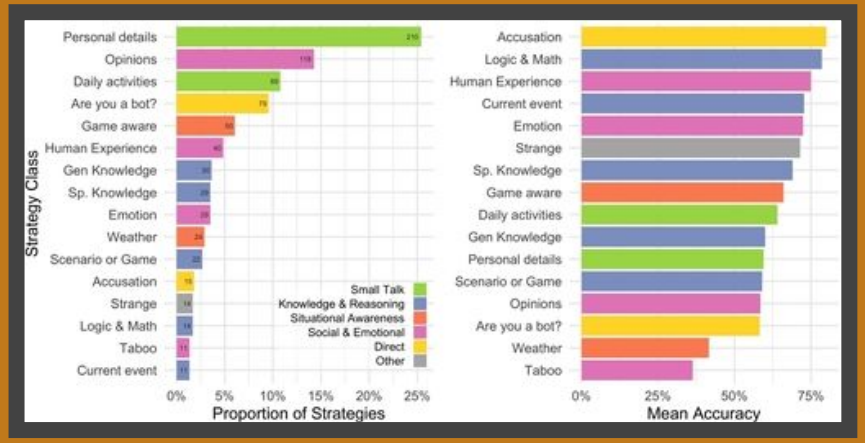
**Control**

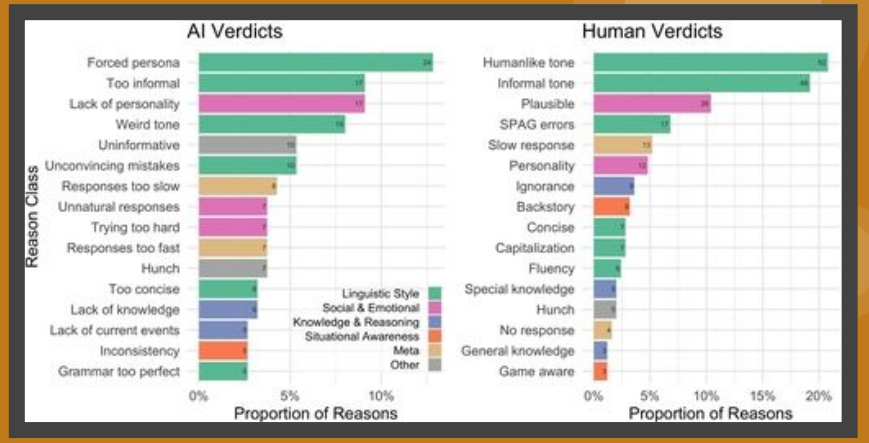
ELIZA as manipulation check (baseline)

# Section 3 (Results)

# Results




**Pass Rates**: Humans 67% > GPT-4 54% ≈ GPT-3.5 50% > ELIZA 22%

**Progress but Gap Remains**: GPT-4 surpasses older models but still falls short of human indistinguishability.

**Effective Strategies**: Logic, math, and human-experience questions increased interrogator accuracy (≈75–80%).

**Judgment Cues**: Interrogators focused on style and personality rather than factual knowledge.

**AI Familiarity Ineffective**: Self-reported LLM knowledge didn't improve detection ability.

**Replication with GPT-4o**: Expanded prompts, repeated rounds, and open transcript data enhance robustness and transparency.

# Section 4 (Replication)

**Why replicate?**

Original flawed: inconsistent results, missing transparency for transcripts

New population tested (undergrads)

**Design Improvements**

GPT-4o used

9 varied AI witness prompts

8 interaction rounds per participant (AI + human mix)

**Prompt Testing**

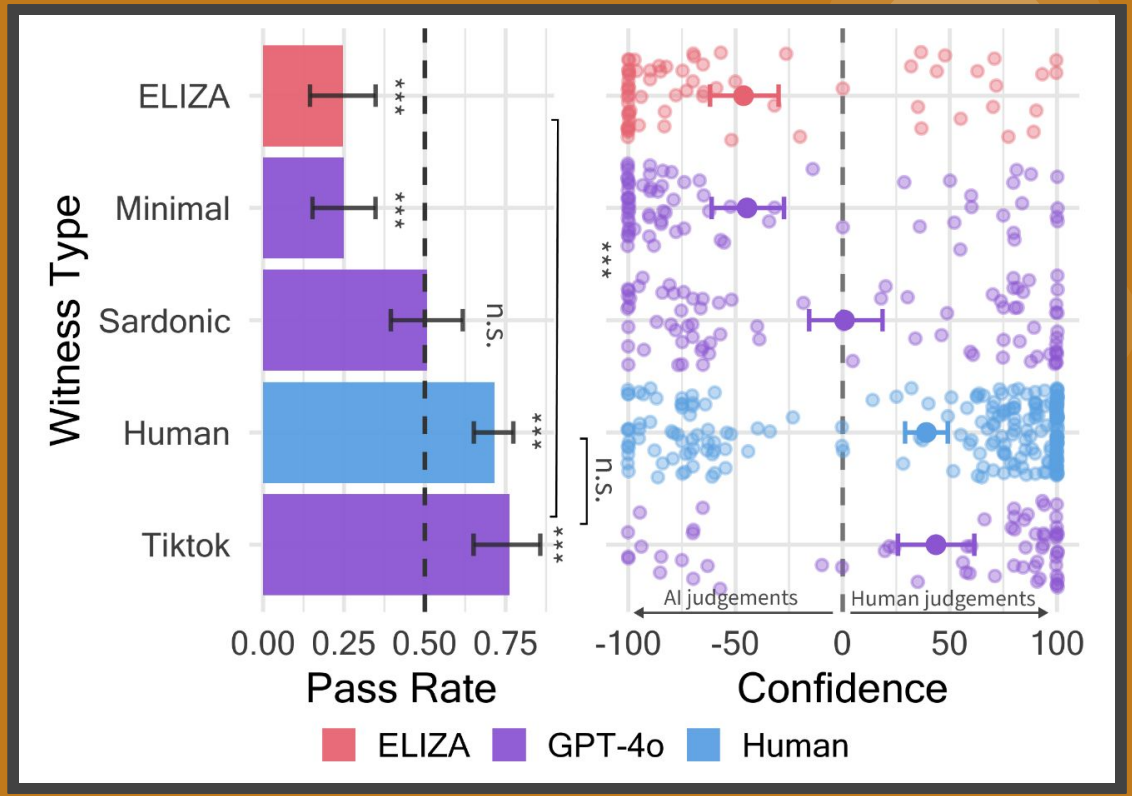
TIKTOK: 76% pass rate

Humans: 71% pass rate

Sardonic: 50% pass rate

Minimal & ELIZA: 25% pass rate

QINN: strong (86%) but low sample (n=15)



# Section 5 (Discussion)

## Re-evaluating the Turing test “Pass”

**50% Pass Criterion:** Turing's **70%** is arbitrary; a **50%** chance-level baseline is more justified for reliable machine identification.

**GPT-4/4o Performance:**

a. **GPT-4/4o:** GPT-4 "passed" (vs. ELIZA baseline). GPT-4o hit 76% (vs. human 71%), "human parity."

b. **Human Misidentification:** Humans only 67% identified as human, interrogators now mislabel humans as AI.

## Implications of AI Deception

**Individual Impact:** Being mistaken for AI can be "jarring or dehumanising."

**Societal Threat:** Risks mutual trust, transparency, and accountability in online interactions (social media, Wikipedia).

## Mitigation Strategies Against AI Deception

**Goal:** Prevent forced doubt about online authenticity of real people.

**Proposed Solutions:** Statistical watermarks, digital identity authorisation, disclosure of data provenance.

## What the Turing Test Measures Now

**AI success focuses on imitating** "linguistic style and socio-emotional factors," not traditional intelligence.

**"ELIZA effect" examined:** LLMs significantly outperform ELIZA, but ELIZA's "uncooperativeness" can be misconstrued as human.

People cannot distinguish GPT-4 from a human in a Turing test

| Witness Type | Witness | No. Games | Passes | Pass Rate |
|---|---|---|---|---|
| GPT-4o | Quinn | 15 | 13 | 86.67 |
| GPT-4o | Tiktok | 63 | 48 | 76.19 |
| Human | Human | 221 | 158 | 71.49 |
| GPT-4o | Underminer | 22 | 12 | 54.55 |
| GPT-4o | Sardonic | 81 | 41 | 50.62 |
| GPT-4o | Troll | 24 | 8 | 33.33 |
| GPT-4o | Minimal | 72 | 18 | 25.00 |
| ELIZA | ELIZA | 69 | 17 | 24.64 |
| GPT-4o | Opinionated | 26 | 4 | 15.38 |
| GPT-4o | Emily | 19 | 1 | 5.26 |
| GPT-4o | Andrew | 17 | 0 | 0.00 |

**Table 1: The number of games participated in, passes (the number of games in which a witness was judged to be human) and the pass rate for all witnesses in the replication study.**

# Our thoughts

## Redefining "Intelligence":

- The study shows LLMs excel at imitating human style and emotion via text.
- **Question:** Does the Turing Test still assess true "intelligence," or primarily an AI's ability to deceive through human-like imitation? Should our definition of AI intelligence evolve beyond this?

## Evolving the Test for Multimodal AI:

- The study was text-only. Modern LLMs are multimodal (voice, image, video).
- **Question:** How would adding multimodal interactions impact the Turing Test? Would it give humans an edge, or make AI deception even more sophisticated?

## Societal Impact of AI Deception:

- AI's ability to fool humans raises major concerns: fraud, social engineering, eroded trust.
- **Question:** What are the most pressing ethical concerns, and what tangible steps can be taken to mitigate deepfakes, scams, and misinformation from increasingly deceptive AI?

People cannot distinguish GPT-4 from a human in a Turing test

| Witness Type | Witness | No. Games | Passes | Pass Rate |
|---|---|---|---|---|
| GPT-4o | Quinn | 15 | 13 | 86.67 |
| GPT-4o | Tiktok | 63 | 48 | 76.19 |
| Human | Human | 221 | 158 | 71.49 |
| GPT-4o | Underminer | 22 | 12 | 54.55 |
| GPT-4o | Sardonic | 81 | 41 | 50.62 |
| GPT-4o | Troll | 24 | 8 | 33.33 |
| GPT-4o | Minimal | 72 | 18 | 25.00 |
| ELIZA | ELIZA | 69 | 17 | 24.64 |
| GPT-4o | Opinionated | 26 | 4 | 15.38 |
| GPT-4o | Emily | 19 | 1 | 5.26 |
| GPT-4o | Andrew | 17 | 0 | 0.00 |

**Table 1: The number of games participated in, passes (the number of games in which a witness was judged to be human) and the pass rate for all witnesses in the replication study.**