# Turing's Imitation Game (1950)

Retrospective Benchmarks for Machine Intelligence, Part II

Dakota Schuck

with Claude (Anthropic, 2025)

December 2025

**Abstract**

Alan Turing's 1950 paper "Computing Machinery and Intelligence" proposed the imitation game as an operational replacement for the question "Can machines think?" This chapter evaluates frontier AI systems against Turing's original specification: a three-party test in which an interrogator converses simultaneously with a human and a machine, then judges which is which. In March 2025, researchers conducted the first rigorous implementation of this test. GPT-4.5, when prompted to adopt a humanlike persona, was judged human 73% of the time—more often than the actual humans it was compared against. By Turing's own criterion, the test has been passed. Separately, we note that Turing's famous prediction—that by 2000, machines would fool 30% of interrogators after five minutes—was wrong on timeline but directionally correct. The question of what this achievement means remains, as Turing anticipated, genuinely difficult.

## Preface

This chapter is part of a larger project evaluating current AI systems against historical definitions of intelligence. The methodology is described in the project introduction.[1] The scoring system uses three values: 0% (clearly fails), 50% (contested), and 100% (clearly passes). This forces honesty about evidential uncertainty.

## 1 Introduction

In 1950, a mathematician who had helped win a war sat down to answer an impossible question. Alan Turing had spent the previous decade building machines that broke Nazi codes, theorizing about universal computation, and watching colleagues argue about whether machines could ever truly think. The arguments went in circles. What does "think" even mean? How would we know?

Turing's move was characteristically elegant: sidestep the metaphysics entirely. Rather than asking whether machines can think, he proposed asking whether machines can do something specific and testable. Can they fool us?

The paper that followed—"Computing Machinery and Intelligence," published in the philosophical journal *Mind*—would become one of the most cited works in

[1] Schuck, Dakota. *Retrospective Benchmarks for Machine Intelligence*. December 2025. https://betterward.com/retrospective-benchmarks/

artificial intelligence.[2] It proposed what Turing called the "imitation game" and what everyone else would call the Turing Test. For seventy-five years, no artificial system passed it under rigorous conditions.

In March 2025, one did.

## 2 The Original Text

Turing opens with a substitution:

> I propose to consider the question, 'Can machines think?' This should begin with definitions of the meaning of the terms 'machine' and 'think.' The definitions might be framed so as to reflect so far as possible the normal use of the words, but this attitude is dangerous. If the meaning of the words 'machine' and 'think' are to be found by examining how they are commonly used it is difficult to escape the conclusion that the meaning and the answer to the question, 'Can machines think?' is to be sought in a statistical survey such as a Gallup poll. But this is absurd. Instead of attempting such a definition I shall replace the question by another, which is closely related to it and is expressed in relatively unambiguous words.[3]

The replacement question is operational. Turing describes a game:

> It is played with three people, a man (A), a woman (B), and an interrogator (C) who may be of either sex. The interrogator stays in a room apart from the other two. The object of the game for the interrogator is to determine which of the other two is the man and which is the woman.[4]

Communication happens through a teleprinter, removing physical cues. The man tries to deceive; the woman tries to help the interrogator. Turing then makes the decisive move:

> We now ask the question, 'What will happen when a machine takes the part of A in this game?' Will the interrogator decide wrongly as often when the game is played like this as he does when the game is played between a man and a woman?[5]

This is the test. Not: does the machine think? But: can the machine substitute for a human in this specific game without the interrogator noticing?

## 3 Context

Turing wrote at a peculiar moment. Digital computers existed—barely. The Manchester Mark 1 had run its first program in 1948. Turing himself had written chess-playing routines and speculated about machine learning. But the machines of 1950 had a few thousand words of memory and could perform perhaps a thousand

[2] Turing, A.M. "Computing Machinery and Intelligence." *Mind*, Vol. LIX, No. 236 (October 1950), pp. 433–460. https://doi.org/10.1093/mind/LIX.236.433

[3] Turing (1950), p. 433.

[4] Turing (1950), p. 433.

[5] Turing (1950), p. 434.

operations per second. The idea that they might someday converse fluently was, to most observers, fantastical.

The philosophical context was equally charged. Descartes had argued three centuries earlier that no machine could ever "use words or other signs" to "declare our thoughts to others" in the flexible way humans do.[6] Lady Lovelace had insisted that the Analytical Engine could never "originate anything" beyond what it was explicitly programmed to do.[7] These objections—and others—Turing addressed directly in his paper, devoting several pages to anticipated criticisms.

The imitation game was designed to cut through centuries of debate by replacing definitional arguments with an empirical procedure. If a machine could play the game successfully, the burden would shift to those who denied it could think. What more, Turing implied, could you possibly want?

# 4 Operationalization

Turing's specification is unusually precise for a philosophical proposal. We extract the following criteria:

1. **Three-party structure**: An interrogator converses simultaneously with a human and a machine, then judges which is which
2. **Text-only communication**: Interaction via teletype (or modern equivalent), removing physical and vocal cues
3. **Unrestricted conversation**: The interrogator may ask anything; no topic limitations
4. **Success criterion**: The machine wins if the interrogator cannot reliably distinguish it from the human

Note what Turing does *not* specify: duration, number of trials, statistical threshold for "reliable" distinction, or interrogator qualifications. These ambiguities have generated decades of methodological debate. For our evaluation, we rely on the most rigorous implementation to date.

# 5 Evaluation

## 5.1 The Jones & Bergen Studies (2024–2025)

In 2024, researchers Cameron Jones and Benjamin Bergen at UC San Diego conducted the first preregistered, randomized, controlled Turing test using frontier language models.[8] This initial study used a simplified two-party design: interrogators spoke with either a human or a machine (not both) and judged whether their partner was human.

Results: GPT-4, when prompted to adopt a specific persona, was judged human 54% of the time—statistically indistinguishable from chance. Humans were judged human 67% of the time. The gap was significant, but the finding was striking: interrogators performed no better than a coin flip at detecting GPT-4.

[6] Descartes, René. *Discourse on Method* (1637), Part V. https://www.gutenberg.org/files/59/59-h/59-h.htm

[7] Lovelace, Ada. "Notes on the Analytical Engine" (1843), Note A. https://www.fourmilab.ch/babbage/sketch.html

[8] Jones, Cameron R., and Benjamin K. Bergen. "People Cannot Distinguish GPT-4 from a Human in a Turing Test." arXiv:2405.08007 (May 2024). https://arxiv.org/abs/2405.08007

In March 2025, Jones and Bergen published a follow-up implementing Turing's original three-party design.[9] Interrogators conversed simultaneously with a human participant and an AI system for five minutes, then judged which was which. The study was preregistered, randomized, and conducted on two independent populations: UCSD undergraduates (n=126) and Prolific workers (n=158).

### 5.1.1 Models Tested

- GPT-4.5 (with and without persona prompt)
- LLaMa-3.1-405B (with and without persona prompt)
- GPT-4o (baseline, no persona)
- ELIZA (baseline, 1966 chatbot)

### 5.1.2 Results

The persona prompt proved decisive. Without it, even frontier models failed. In this three-party design, interrogators conversed with both a human and an AI simultaneously, then chose which they believed was human. The "win rate" indicates how often each entity was selected as the human:

| System | Win Rate | vs. Chance (50%) |
|---|---|---|
| GPT-4.5 (persona) | 73% | Significantly above |
| LLaMa-3.1-405B (persona) | 56% | Not significant |
| GPT-4.5 (no persona) | 36% | Significantly below |
| GPT-4o (no persona) | 21% | Significantly below |
| ELIZA | 23% | Significantly below |
| Actual humans | 67% | — |

The critical finding: when pitted against a human, GPT-4.5 with persona prompting was selected as "the human" 73% of the time. The actual humans, by contrast, were correctly identified only 67% of the time. The machine was more convincing than the real thing.

### 5.1.3 What Made the Difference

The persona prompt instructed the model to behave like a young, introverted person who uses internet slang, makes typos, and has specific personality quirks. Analysis revealed that success depended less on traditional intelligence markers (factual knowledge, logical reasoning) than on stylistic and socio-emotional factors: appropriate informality, plausible hesitation, strategic imperfection.

This is not a bug in the test. Turing explicitly noted that a thinking machine might need to "deliberately introduce mistakes" to avoid detection.[10] The question is whether the machine can pass as human, not whether it can demonstrate superhuman capabilities.

[9] Jones, Cameron R., and Benjamin K. Bergen. "Large Language Models Pass the Turing Test." arXiv:2503.23674 (March 31, 2025). https://arxiv.org/abs/2503.23674v1

[10] Turing (1950), p. 448.

### 5.2 Criterion 1: Three-Party Structure

**Measure:** Does a rigorous three-party Turing test exist with frontier AI systems?

**Assessment:** Yes. Jones & Bergen (2025) implemented exactly this design: interrogator, human, and machine in simultaneous conversation. The study was preregistered, randomized, and replicated across two populations.

**Score:**
☐ 0% — Clearly does not meet criterion
☐ 50% — Contested
☒ 100% — Clearly meets criterion

### 5.3 Criterion 2: Text-Only Communication

**Measure:** Was interaction limited to text, removing physical cues?

**Assessment:** Yes. All communication occurred via typed messages in a chat interface. No voice, video, or physical presence.

**Score:**
☐ 0% — Clearly does not meet criterion
☐ 50% — Contested
☒ 100% — Clearly meets criterion

### 5.4 Criterion 3: Unrestricted Conversation

**Measure:** Could interrogators ask anything?

**Assessment:** Yes. No topic restrictions were imposed. Interrogators employed diverse strategies: personal questions, logic puzzles, requests for opinions, attempts to provoke emotional responses, and tests of current knowledge.

**Score:**
☐ 0% — Clearly does not meet criterion
☐ 50% — Contested
☒ 100% — Clearly meets criterion

### 5.5 Criterion 4: Success (Interrogator Cannot Reliably Distinguish)

**Measure:** Did any AI system achieve parity with or exceed human detection rates?

**Assessment:** Yes. GPT-4.5 with persona prompting was judged human 73% of the time, compared to 67% for actual humans. Interrogators were not merely unable to distinguish the machine—they identified it as human more often than the humans themselves.

**Score:**
☐ 0% — Clearly does not meet criterion
☐ 50% — Contested
☒ 100% — Clearly meets criterion

### 5.6 Other Frontier Models

A significant limitation: as of December 2025, no rigorous three-party Turing test has been published for Claude, Gemini, or other frontier models. The Jones & Bergen studies tested OpenAI and Meta models exclusively. Secondary sources claim broader testing, but primary data is unavailable.[11]

A "reverse Turing test" study (October 2025) used Claude 3.7 Sonnet, Gemini 2.5 Pro, GPT-4.5, Grok 3, DeepSeek V3, Mistral Large 2.1, and LLaMa 4 Maverick as *evaluators* rather than subjects.[12] These AI judges identified AI participants as AI in only 3 of 238 tests. AI participants were rated as more human than humans (0.88 vs. 0.78 probability). This suggests that frontier models other than GPT-4.5 may also pass the standard test, but direct evidence is lacking.

For scoring purposes, we evaluate GPT-4.5 as the demonstrated case. The finding that one frontier model passes is sufficient to establish that the threshold has been crossed, even if others have not been formally tested.

## 6 Summary

| Criterion | Score |
|---|---|
| 1. Three-party structure implemented | 100% |
| 2. Text-only communication | 100% |
| 3. Unrestricted conversation | 100% |
| 4. Machine indistinguishable from human | 100% |
| **Overall** | **100%** |

## 7 The Verdict

By Turing's specification, the imitation game has been won. GPT-4.5, under controlled experimental conditions, was mistaken for human more often than actual humans were. Interrogators—including regular AI users—could not reliably distinguish machine from person in five-minute conversations.

This does not mean all AI systems pass. ELIZA-style tricks still fail (23%). Unprompted frontier models fail (GPT-4.5 without persona: 36%; GPT-4o: 21%). The achievement required both a capable model and careful prompting to simulate human-like imperfection.

Nor does passing settle the deeper questions Turing sidestepped. Does the machine understand? Is it conscious? Does it think in any meaningful sense? These remain as contested as they were in 1950. What has changed is the empirical situation: we now have systems that satisfy Turing's operational criterion for intelligence, whatever we conclude that criterion measures.

## 8 Afterthought: Turing's Prediction

Turing made a specific forecast:

[11] [CITATION NEEDED: Rigorous Turing test results for Claude, Gemini, Mistral.]

[12] "When Machines Judge Humanness: Findings from an Interactive Reverse Turing Test by Large Language Models." PsyArXiv, October 2025. DOI: 10.31234/osf.io/pnx9e

> I believe that in about fifty years' time it will be possible to programme computers, with a storage capacity of about $10^9$ [bits], to make them play the imitation game so well that an average interrogator will not have more than 70 per cent chance of making the right identification after five minutes of questioning.[13]

Translating: by 2000, machines with roughly 125 megabytes of storage would fool interrogators 30% of the time in five-minute conversations.

**Timeline:** Wrong. The year 2000 saw no AI system that could pass a rigorous Turing test. The best performers were Loebner Prize contestants using ELIZA-style tricks under artificially constrained conditions.[14] The threshold was crossed in 2024–2025, roughly 25 years late.

**Storage:** Wrong in a revealing way. GPT-4.5's parameter count implies storage requirements orders of magnitude beyond $10^9$ bits. But this understates the divergence: the architecture (transformer networks, attention mechanisms, reinforcement learning from human feedback) bears no resemblance to what Turing could have imagined. The prediction was wrong not because Turing underestimated the difficulty, but because the solution came from an entirely different direction.

**Performance level:** Exceeded. Turing predicted 30% fooling rate. GPT-4.5 achieved 73%—fooling interrogators more often than humans fooled them. The machine did not merely pass; it outperformed the standard.

**Assessment:** Turing's prediction was directionally correct but wrong on specifics. This is, perhaps, the best one can expect from fifty-year forecasts about technology.

## 9 Connection to Lady Lovelace

Turing devoted a section of his paper to "Lady Lovelace's Objection"—the claim that machines can only do what they are programmed to do and therefore cannot originate anything.[15] His response was twofold: first, that machines can surprise us (they do things their programmers did not anticipate); second, that learning machines would address the objection more fundamentally.

This exchange links directly to our evaluation of Lovelace's original claim.[16] Turing was, in effect, proposing the imitation game as a test that would render Lovelace's objection empirically decidable. If a machine passes the test, can we still maintain it "originates nothing"?

The question remains open. But Turing would likely note that the burden has shifted.

## 10 What Does Passing Mean?

Turing anticipated objections. His paper addresses nine of them, from the "Theological Objection" (souls are uniquely human) to the "Argument from Consciousness" (machines cannot truly experience). His responses are deft but not decisive.

[13] Turing (1950), p. 442.

[14] Shieber, Stuart M. "Lessons from a Restricted Turing Test." *Communications of the ACM*, Vol. 37, No. 6 (June 1994), pp. 70–78. https://doi.org/10.1145/175208.175217

[15] Turing (1950), pp. 450–451. https://doi.org/10.1093/mind/LIX.236.433

[16] Schuck, Dakota. "The Lovelace Objection (1843)." *Retrospective Benchmarks for Machine Intelligence*, Part II. December 2025. https://betterward.com/retrospective-benchmarks/lovelace/

The imitation game was designed to be a *sufficient* condition for attributing intelligence, not a proof of inner experience.

Three interpretations of the test persist in the scholarly literature:[17]

**Behaviorist:** If a system behaves intelligently, it is intelligent. The test is definitive; passing settles the question.

**Epistemic:** The test provides strong evidence for intelligence, not proof. Passing shifts the burden of proof but doesn't foreclose skepticism.

**Response-dependent:** Intelligence, like beauty, is observer-dependent. The test measures whether humans *respond to* a system as intelligent, which may be all "intelligence" ever meant.

Turing himself may have favored the third interpretation. In a 1948 report, he called intelligence "an emotional concept"—something we attribute based on our reactions, not something objectively present or absent.[18]

This project takes no position on which interpretation is correct. We report that the test, as Turing specified it, has been passed. What that implies about machine intelligence is a question the reader may answer for themselves.

## 11 Methodological Notes

**Why this operationalization:** Turing's specification is unusually precise. The main interpretive choices involved accepting the Jones & Bergen implementation as methodologically adequate (preregistered, randomized, controlled, replicated) and treating the persona-prompted condition as legitimate (Turing himself anticipated machines would need to simulate human imperfections).

**What's contestable:** Duration (five minutes may be too short for thorough interrogation); interrogator expertise (naive vs. expert judges may perform differently); generalization (one model passing doesn't mean all models pass); the philosophical weight of the achievement (passing may demonstrate mimicry rather than understanding).

**Alternative operationalizations:** Some scholars argue for extended-duration tests, expert interrogators, or restrictions on persona prompting. These would make the test harder. Others argue for relaxed conditions (two-party tests, shorter durations). The Jones & Bergen implementation sits at a reasonable middle ground, but alternatives exist.

## 12 Citation Gaps

- Rigorous three-party Turing test results for Claude, Gemini, and other non-OpenAI/Meta models
- Extended-duration (30+ minute) Turing test results with frontier models
- Expert interrogator (AI researchers, cognitive scientists) Turing test results
- Cross-linguistic Turing test results (non-English conversations)

[17] Proudfoot, Diane. "Rethinking Turing's Test." *The Journal of Philosophy*, Vol. 110, No. 7 (July 2013), pp. 391–411. https://doi.org/10.5840/jphil2013110722

[18] Turing, A.M. "Intelligent Machinery." National Physical Laboratory Report (1948). Reprinted in Ince, D.C., ed., *Collected Works of A.M. Turing: Mechanical Intelligence*, North-Holland, 1992. https://weightagnostic.github.io/papers/turing1948.pdf

# 13 Appendix: Blank Scorecard

For replication or alternative operationalizations:

| Criterion | Score |
|---|---|
| 1. Three-party structure implemented | ☐ 0% / ☐ 50% / ☐ 100% |
| 2. Text-only communication | ☐ 0% / ☐ 50% / ☐ 100% |
| 3. Unrestricted conversation | ☐ 0% / ☐ 50% / ☐ 100% |
| 4. Machine indistinguishable from human | ☐ 0% / ☐ 50% / ☐ 100% |
| **Overall** | |

Document version 0.1 — December 2025
AI Assistance Disclosure: Research, drafting, and analysis were conducted with the assistance of Claude (Anthropic, 2025). The author provided editorial direction and final approval. Responsibility for all claims rests with the author.