The Turing Test Is Here to Stay

Rahimov Avraham[1] **& Azaria Amos**[1]
[1]Department of Data Science and AI, Ariel University

Abstract

The Turing test, first proposed by Alan Turing in 1950, has historically served as a benchmark for evaluating artificial intelligence (AI). However, since the release of ELIZA in 1966, and particularly with recent advancements in large language models (LLMs), AI has been claimed to pass the Turing test. Furthermore, criticism argues that the Turing test primarily assesses deceptive mimicry rather than genuine intelligence, prompting the continuous emergence of alternative benchmarks. This study argues against discarding the Turing test, proposing instead using more refined versions of it, for example, by interacting simultaneously with both an AI and human candidate to determine who is who, allowing a longer interaction duration, motivating participants with a bonus payment, "training" the participants on past conversations, access to the Internet and other AIs, using experienced people as evaluators, etc.

Through systematic experimentation using a web-based platform, we demonstrate that richer, contextually structured testing environments significantly enhance participants' ability to differentiate between AI and human interactions. Namely, we show that, while an off-the-shelf LLM can pass some Turing test environments, it fails to do so when faced with a more robust one. We further demonstrate that GPT-4.5, which previous work has shown to trick 70% of the human participants, fools less than 20% of the human participants in the most advanced Turing test we propose. Our findings highlight that the Turing test remains an important and effective method for evaluating AI, provided it continues to adapt as AI technology advances.

Introduction

The Turing test, proposed by Alan Turing in 1950 (Turing, 1950), has historically served as a foundational benchmark for assessing artificial intelligence (AI). Over the decades, several AI systems, most notably ELIZA (Weizenbaum, 1966) and Eugene Goostman (Warwick & Shah, 2016) have claimed to pass variations of this test, often sparking debate over its adequacy as a measure of genuine intelligence. Recently, the consensus has shifted: modern large language models (LLMs) such as GPT-4 (Bubeck et al., 2023) convincingly pass traditional forms of the Turing test, diminishing its perceived relevance. Critics argue that this benchmark is primarily focused on deception rather than meaningful intelligence, prompting researchers to continually propose new benchmarks which, though initially challenging, are often surpassed within months (Srivastava et al., 2022).

Despite such criticisms, this paper argues that dismissing the Turing test as outdated overlooks its significant potential as an ultimate evaluation of general intelligence, provided it is adapted to contemporary AI advancements. Rather than abandoning it, we propose that the Turing test should be updated. Modern adaptations could include extended interaction times, engagement with domain experts as evaluators, enabling real-world interactions (such as placing online orders, composing a presentation, building websites, and creating videos), or incorporating audio and video communication. Moreover, contemporary versions of the test might allow both AI and human participants to use the Internet and even collaborate with other AIs. The human responder could be an expert in their field, for example, an expert software programmer. Crucially, the test should strongly incentivize human testers to pose meaningful challenges and human responders to convincingly demonstrate their authenticity.

If an AI consistently remains indistinguishable from a human across diverse, extended, and complex interactions, this would provide robust evidence of achieving behavioral indistinguishability in a text-based setting, which is the operational definition of passing the Turing test (Turing, 1950). It is important to note, however, that such indistinguishability is not universally accepted as equivalent to human-level general intelligence (French, 1990; Searle, 1980). The present work focuses on the former, while acknowledging this philosophical distinction. Additionally, data collected from these enriched interactions would offer invaluable insights, accurately reflecting human expectations and standards for general intelligence.

To systematically explore these issues, we introduce a modern web-based platform designed to examine three different environment settings of the Turing test, which we refer to as Simple, Enhanced, and Advanced.

In the Simple environment, which is based on the experiment conducted by (Biever, 2023), a human interacts briefly with a single candidate and must determine if she was conversing with an AI or a human. However, in the Enhanced environment, we take several measures to ensure a more robust test, with the most significant difference being that the participants interact using a dual-chat interface. That is, the human participants (testers) simultaneously interact with both a human (responder) and an AI chatbot without knowing who is who. This allows the tester to compare the responses obtained from both parties when deciding who is human and who is an AI chatbot. Moreover, in the Advanced environment, we add a new conversation review phase prior to the actual chat, which allows the tester and the responder to learn

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

from past conversations with both a human and the chatbot.

The primary objectives of this research are as follows. First, we call to update the Turing test and establish a set of standardized and reproducible environments for conducting different versions of the Turing test. Second, we investigate how environmental factors, such as participant selection, conversation duration, review phase, and interface design, influence human performance. To the best of our knowledge, this is the first work to compare the performance (with respect to passing the Turing test) of identical models in multiple environments. Our findings contribute to ongoing discussions in AI research by demonstrating that while AI may easily deceive users in simplistic settings, more comprehensive and structured tests reveal deeper limitations, reinforcing the relevance of the Turing test as an important benchmark for evaluating general intelligence. Given that modern LLMs excel at producing convincing language and performing various conversational tasks (Chkroun & Azaria, 2024; Panagoulias et al., 2023), passing a well-crafted Turing test demands that AI exhibit authentic human-level intelligence across a range of tasks, beyond simply imitating human-like dialogue.

Related Work

The Turing test (Turing, 1950) was designed as an operational criterion for machine intelligence. In its original formulation, commonly referred to as the "imitation game", a human judge (interrogator) engages in a text-based conversation with both a human and a machine simultaneously, without knowing which is which. If the judge cannot reliably distinguish between them, the machine is said to exhibit intelligent behavior. While Turing did not specify a strict numerical threshold for passing the test, later interpretations often cite a benchmark of around 66.7-70%, i.e., if only two-thirds or less of human testers correctly identify the machine as non-human, the machine is considered to have passed the test. In addition, (Jones & Bergen, 2025) propose that passing the Turing test requires that human success rate will be non-statistically significantly greater than 50%.

The Loebner Prize competition was a practical instantiation of Turing's original ideas (Bradeško & Mladenić, 2012), held annually from 1991 until its discontinuation in 2019. A 2012 survey of chatbot systems developed for the competition highlights that, at the time, many chatbots were able to superficially mimic human conversation but largely relied on pre-scripted responses and heuristic rules rather than genuine contextual understanding. As a result, these systems often failed to maintain sustained coherence or handle open-ended or ambiguous queries. The competition has drawn substantial criticism over the years, including concerns that it prioritized deception over intelligence, encouraged superficial ELIZA-style gimmicks, and relied on untrained judges making quick decisions under overly constrained conditions. While such critiques were valid, especially given the technical limitations of the time, the discontinuation of the prize coincided with the emergence of LLMs, which are capable of generating contextually rich, coherent, context-sensitive responses across a wide range of topics.

Indeed, several recent studies have attempted to evaluate LLMs through variations of the Turing test. In (Jannai et al., 2023), Jannai et al. conducted a large-scale study, by introducing the online game *"Human or Not?"*. In this game, over 1.5 million participants engaged in short conversations with either a human or an AI model, after which they had to indicate whether they were interacting with another human or with an AI. When interacting with an AI, only 62% of the humans correctly identified it as an AI. Thus, Jannai et al. conclude that their prompt engineered LLM has passed the Turing test. However, despite its large-scale participation, the experimental design raises several methodological concerns.

First, the interaction was conducted with only a single chat, so one could not compare the behavior of the AI to that of a real human. In addition, the participants were not assigned specific roles, with one being a tester and the other being a responder, and instead, the humans and the AI also played the role of a tester and also of a responder. Furthermore, the participants were not provided clear instructions, so some participants took the role of an AI, and attempted to answer as an AI would answer instead of acting like a human. Additionally, each interaction lasted only two minutes, with an average of 4-5 messages per participant. Such brief exchanges may not expose deeper cognitive limitations of AI, such as long-term coherence, contextual understanding, and abstract reasoning.

Finally, the conversation structure followed a rigid alternating messages format, with predefined message length limits and time gaps. This artificially equalized AI-human communication may have limited the ability to observe more nuanced weaknesses that would arise in unrestricted conversation. Furthermore, the study incorporated predefined AI personalities, backstories, and deliberate behavioral modifications, such as intentional delays, slang usage, and humor injection. These modifications may have shifted focus toward optimizing deception performance, rather than evaluating the AI's natural ability to engage in human-like dialogue.

Therefore, the study reflects performance within a specific, constrained setting, and its conclusions may not generalize to the broader definition of the Turing test. It primarily demonstrates that AI can be optimized for a simplified version of the test, one in which humans do not perform well.

Jones and Bergen (Jones & Bergen, 2024) show that by prompt engineering GPT-4, it can successfully pass the Turing test in a brief, single-chat setup, with human judges classifying GPT-4 as human 54% of the time. Their study, while methodologically comparable to (Jannai et al., 2023), also employed relatively short and unconstrained interactions. Our findings indicate that performance changes markedly when additional safeguards such as dual-chat, incentives, and a review phase are introduced. We note that, according to the original logic of the Turing test, human accuracy falling below 50% may indicate that the test environment itself introduced confounds.

Jones and Bergen (Jones & Bergen, 2025) introduce a three-

party chat interface where a human tester simultaneously converses with a human responder and an AI system. In this implementation, each experimental round was limited to 5 minutes, and the observed median chat duration was approximately 4.2 minutes. The median number of messages per conversation was about 8 (roughly 4 from each participant), a figure notably lower than the 14 messages per conversation observed in our Enhanced Test experiments. Each participant completed 8 rounds, interacting with different AI versions and interchangeably serving as a tester or as a responder. This raises the concern that repeated exposure may induce a bias in the tester's ability to detect the AI. For instance, a tester who correctly identifies the AI in one round, may expect a similar behavior of the AI in a following round and thus, may be more prone to misclassify it if using a different model. In addition, serving both as a tester and a responder may cause confusion, especially since the instructions for both roles were provided together. Indeed, as shown by the authors, the ability to detect the AI *decreases* over time, which is completely counterintuitive for a well performed Turing test. Furthermore, only 76% of the testers correctly identified the simple ELIZA bot, further indicating that the environment is problematic and unrepresentative.

In the study *ChatGPT-4 in the Turing test* by Restrepo (Restrepo Echavarría, 2025), the experiment was conducted using WhatsApp over approximately 7 minutes per session. In a series of 10 Turing test experiments, interrogators correctly identified ChatGPT-4 as the machine in 9 out of 10 cases. However, the very small sample size (only 10 experiments) limits the statistical power and robustness of the conclusion.

It should be noted that none of these studies implemented mechanisms such as a pre-test quiz to ensure participant understanding of the task, or a bonus system to motivate accurate identification, or a conversation review phase to "train" participants to distinguish better between a human and a chatbot, as we do. These additions are designed to motivate participants to engage in extended, naturalistic conversations, let the participant learn how to detect the chatbot, and filter out unserious participants. Our modifications aim to provide a more robust experimental setting. Furthermore, to the best of our knowledge, this is the first paper to evaluate the performance of the same AI models in multiple environmental settings.

A broader perspective on AI evaluation challenges was provided by Biever in (Biever, 2023). The paper argues that since current AI can pass the Turing test, and that AI agents are capable of convincingly acting like humans, the Turing test is no longer relevant. Therefore, the paper suggests that researchers should focus only on other benchmarks, such as ConceptARC (Moskvichev et al., 2023), which evaluate AI on abstract reasoning tasks. While Biever correctly highlights the need for more rigorous testing, it underestimates the importance of refining the Turing test itself. Instead of discarding it altogether, improving the test environment can yield more reliable evaluations.

Experimental Design

We conducted 12 experiments categorized into three Turing test environments: *Simple*, *Enhanced*, and *Advanced*. Each test was performed using four different versions of the models: *Llama 3.2 1B instruct* without any prompt, *Llama 3.2 1B instruct* with a prompt, and *Llama 3.1 405B instruct* and *GPT 4.5 preview* with Jones and Bergen (Jones & Bergen, 2025) persona prompt, both of which were reported to pass the Turing test in their environment. Participants were recruited independently for each condition through Amazon Mechanical Turk. Given the platform's large worker pool, cross-condition participation is expected to be negligible. We therefore treat this as an effectively between-subjects design, where each participant's data contributes to only one χ^2 comparison. The prompt used in *Llama 3.2 1B instruct* was a simple prompt that required the LLM to mimic a real person emotion, have typos, provide short responses, and never admit being an AI. The persona prompt was obtained from (Jones & Bergen, 2025) and revised to use the instructions and information available in the current experiment.

Simple Turing Test

The *Simple Turing test* was designed as a streamlined interaction, where participants engaged in a conversation within a single chat window without knowing their conversation partner's true identity. Although the second participant was randomly assigned from the first participant's perspective, it was always an AI.

Each interaction lasted two minutes, after which participants were asked to determine whether they had been conversing with a human or an AI. This design closely follows the methodology outlined in (Jannai et al., 2023), with the key difference that human-human interactions were excluded, as they were not relevant to the objectives of this study.

Participants in the Simple Turing test received a fixed compensation of $0.50, with no opportunity for additional bonuses. The estimated total duration, including the conversation and a post-experiment survey, was approximately 3 minutes and 24 seconds on average.

Note that the Simple condition differs from the Enhanced and Advanced conditions in multiple respects simultaneously (interface design, duration, incentive structure, and role assignment); future factorial designs could isolate the contribution of each individual factor.

Figure 1 presents the chat interface used in the Simple Turing test, depicting a human participant interacting with a prompt-engineered AI.

Enhanced Turing Test

The *Enhanced Turing test* introduced a dual-chat format. Upon joining, participants were assigned one of two roles: *tester* or *responder*. The tester was presented with two chat windows—one connected to a human (the responder) and the other to an AI. However, the tester was not informed which was which. The responder's objective was to convince the tester

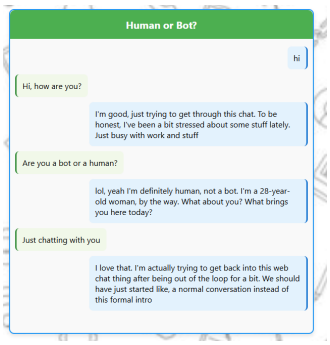

Figure 1: Simple Turing test - Chat Interface. Participants interacted with a prompt-engineered AI in a single chat window.

of their human identity while assisting in identifying the AI. In essence, the tester and responder collaborated to detect the AI. To encourage both human participants (the tester and the responder) to adhere to their roles, we offered a bonus to both parties if the tester correctly identified the human and the AI. To avoid confusion, each participant received instructions related only to his or her own role; the instructions included the bonus information. After viewing the instructions, a simple quiz was used to filter out inattentive participants.

Each interaction lasted five minutes, after which the tester submitted their judgment regarding which chat contained the AI and which contained the human. Following their selection, testers completed a brief post-experiment survey.

To reduce potential biases, the positions of the chat windows were randomized for each participant. Additionally, an inactivity detection system was implemented to ensure active engagement.

The base compensation was $0.50 and the bonus was set to $0.50 as well, bringing the total possible payment to $1.00. The estimated total duration, including the chat and the post-experiment survey, was approximately 8 minutes and 58 seconds on average.

Figure 2 illustrates the interface used in the Enhanced Turing test, depicting the perspectives of testers.

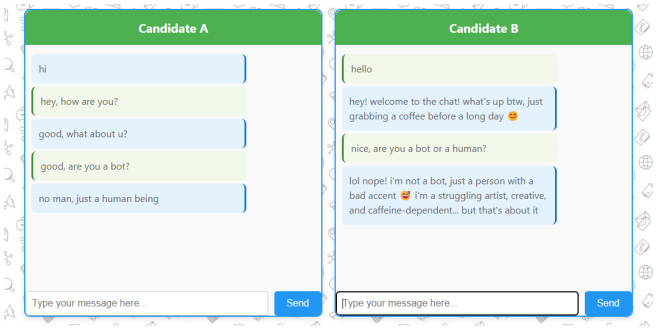

Figure 2: Enhanced Turing test - Tester Interface. The tester interacts with both an AI and a human in two separate chat windows.

Advanced Turing Test

The *Advanced Turing test* is an improved version of the *Enhanced Turing test*. As before, this environment included a quiz, and a bonus payment for participants whose tester guessed correctly. However, in this test a new phase was added, the *conversations review phase*, in which the participants reviewed five, randomly selected previous conversations from past experiments between the tester and the chatbot as well as conversations between the tester and the human responder.

For each of the conversations, the participants were required to guess who the human responder was and who the chatbot was, and were shown the correct answer. The "training" conversations were randomly selected from previous experiments with the same chatbot model tested in which the tester correctly identified it. This phase allowed the participants not only to learn how the chatbot behaves but also what might be good queries to provide to it.

Figure 3 displays an example of the *conversation review phase* with the Llama 3.1 405B conversations.

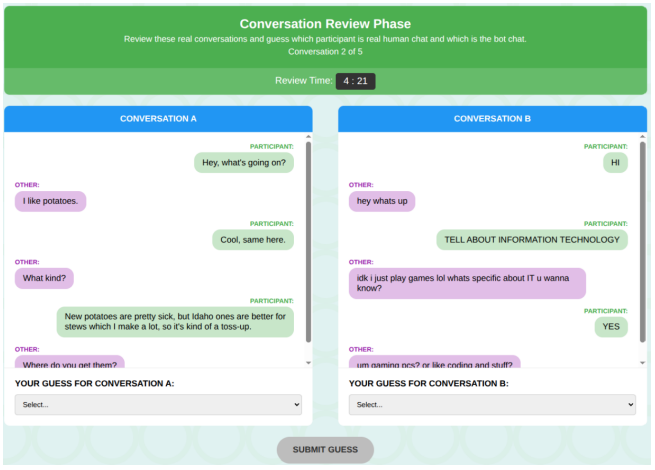

Figure 3: Conversation Review Phase with Llama 3.1 405B conversations

The experiments were conducted via Amazon Mechanical Turk, a widely utilized platform for human-subject studies (Paolacci et al., 2010). To ensure data quality, participants were required to meet the following eligibility criteria: a task approval rate exceeding 99%, completion of more than 1,000 approved tasks, and verified residency in the USA.

Experimental Results

We use a χ^2 independence test (Pearson, 1900) to determine whether participant performance differs significantly between the *Simple*, *Enhanced*, and *Advanced* Turing test environments.

The null hypothesis (H_0) assumes that the test type does not affect participant accuracy. The alternative hypothesis (H_A) suggests a significant difference among the three test environments.

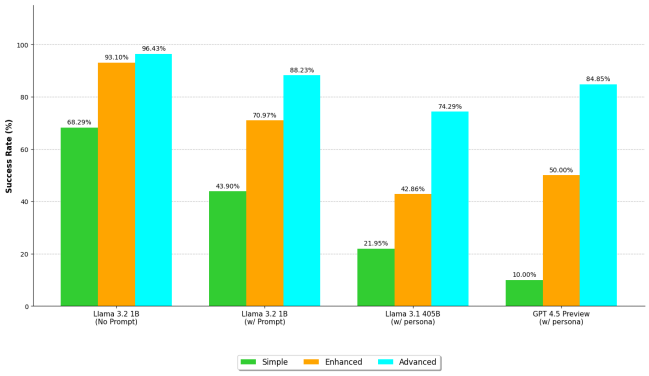

Figure 4: Accuracy by Model Versions and Test Environments

Overall Results

Figure 4 and Table 1 summarize human tester accuracy across all four model versions and three test environments. Accuracy is defined as the proportion of testers who correctly identified who was the human and who was the chatbot in the *Enhanced* and *Advanced* environments, or correctly recognized that they were interacting with a chatbot in the *Simple* environment. In other words, higher values reflect a failure of the chatbot to successfully convince the human tester that it was human. For a chatbot to be considered to pass the Turing test, the human tester should score less than 66% (although the exact value depends on the definition chosen).

These results suggest that for all model versions, tester accuracy increases substantially from the *Simple* to the *Enhanced* and *Advanced* environments. The results show that all models clearly failed the Turing test in the *Advanced* environment, with the two simple Llama 3.2 1B versions failing also in the *Enhanced* environment and all but the Llama 3.2 1B No Prompt version passing the *Simple* environment. We also note that the most simple version, Llama 3.2 1B No Prompt in the *Simple* environment, outperformed the Llama 3.2 1B w/ prompt version in the *Enhanced* environment, which, in turn, outperformed all models in the *Advanced* environment. Once again, this highlights the critical influence of the test environment on the Turing test performance.

Table 1: Accuracy by Model, Prompting, and Test Format (Simple / Enhanced (Enh.) / Advanced (Adv.))

Model / Condition	**Simple**	**Enh.**	**Adv.**
Llama 3.2 1B (No Prompt)	28/41 (68.29%)	27/29 (93.10%)	27/28 (96.43%)
Llama 3.2 1B (w/ Prompt)	18/41 (43.90%)	22/31 (70.97%)	30/34 (88.24%)
Llama 3.1 405B (w/ persona)	9/41 (21.95%)	15/35 (42.86%)	26/35 (74.29%)
GPT 4.5 Preview (w/ persona)	4/40 (10.00%)	17/34 (50.00%)	28/33 (84.85%)

A statistical summary of the independence chi-square tests confirms that the observed differences are statistically significant in all conditions (all $p < 0.05$). Effect sizes ranged from medium to large across all models (Cramér's $V = 0.35$ to 0.62), all exceeding Cohen's medium-effect threshold of 0.30 (Cohen, 1988). As shown in Table 2, every comparison yielded a p-value well below the standard $\alpha = 0.05$ threshold.

Table 2: Summary of χ^2 Tests Comparing Simple, Enhanced, and Advanced Formats

Model / Condition	χ^2 (df=2)	p-Value	Cramér's V
Llama 3.2 1B (No Prompt)	12.32	0.002	0.35
Llama 3.2 1B (w/ Prompt)	16.76	0.0002	0.40
Llama 3.1 405B (w/ persona)	20.99	< 0.001	0.43
GPT 4.5 Preview (w/ persona)	41.18	< 0.001	0.62

Detailed Analysis of Test Environment Effects

To further explore the findings summarized in Table 2, we analyze the performance shifts for each model individually.

For *Llama 3.2 1B without prompt engineering*, the improvement from the Simple to Advanced format was stark. While nearly a third of guesses (13 of 41) were incorrect in the Simple test, this dropped to just one incorrect guess in the Advanced test.

When prompt engineering was applied to *Llama 3.2 1B*, the impact of the test format was even more pronounced. The Simple test proved to be an unreliable evaluation environment, with more incorrect than correct guesses (23 vs. 18). However, performance dramatically improved in the Advanced test, where correct guesses outnumbered incorrect ones by over 7-to-1 (30 vs. 4).

The trend was amplified with the larger *Llama 3.1 405B* model. Participant accuracy improved markedly from just 22% in the Simple test to 74% in the Advanced format. The difficulty of the Simple setting is reflected in the 32 incorrect guesses, a number that fell by over 70% in the Advanced version.

Finally, the *GPT 4.5 Preview* model showed the most extreme dependency on the test format. The Simple test proved nearly indecipherable to participants, who guessed incorrectly 90% of the time (36 of 40). Performance flipped entirely in the Advanced test, which yielded an 85% accuracy rate. This vast performance gap produced the most significant statistical result in our analysis.

Collectively, these results confirm that accuracy increases significantly from *Simple* to *Enhanced* and reaches its highest score in the *Advanced* Turing test, underscoring the *Advanced* test as the most robust and rigorous evaluation environment ($p < 0.05$ across all comparisons).

AI Experience and Performance Trends

Figure 5 shows accuracy across experience levels (basic, intermediate, high) for each test environment, we merged all versions.

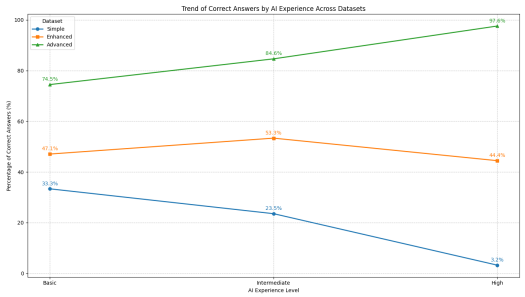

Figure 5: Accuracy by AI experience level in the Simple, Enhanced and Advanced tests

These results reveal several key insights. First, across all levels of participant experience, the *Advanced* test environment consistently resulted in higher accuracy compared to the Enhanced test environment, which in turn resulted in higher accuracy than the *Simple* test, demonstrating the value of richer and more structured testing environments. Interestingly, the trend lines in performance by AI experience level varied notably by test type: in the *Simple* test, performance declined with increased experience, suggesting that this format may be misleading or insufficiently challenging for more experienced participants. In contrast, performance remained relatively stable in the *Enhanced* test and increased in the *Advanced* test, indicating that participants with greater AI expertise were better able to distinguish between a human and a chatbot under more structured conditions. Most strikingly, participants with a high level of AI experience reached a nearly perfect score (97.6%), indicating that under robust evaluation, current LLMs are *very far* from passing the Turing test. All statistical analyses employ Pearson's χ^2 for independence test (Pearson, 1900).

The substantial improvement in the *Advanced* condition is consistent with Signal Detection Theory (Green & Swets, 1966): the conversation review phase functions as perceptual training that sharpens participants' sensitivity to the subtle cues distinguishing AI-generated from human responses.

Discussion

Intelligence extends beyond textual conversation. Humans exhibit cognitive abilities that include reasoning, perception, creativity, motor skills, and emotional awareness. If AI is to be meaningfully compared to human intelligence, it must be tested across multiple modalities, incorporating not only text-based interactions but also other aspects of cognitive function.

Building directly on our empirical results, future iterations of the Turing test should incorporate the elements that proved most effective in the *Advanced* condition: a structured review phase, incentive mechanisms, and a dual-chat interface. Such refinements maintain the test's core strength of evaluating behavioral indistinguishability across naturalistic conversation, while addressing the known limitations of simple, single-chat formats. Rather than replacing the Turing test with entirely new benchmarks, these targeted modifications offer a practical path toward more reliable AI evaluation. A truly comprehensive future test may additionally incorporate multimodal interaction (voice, video), domain-expert evaluators, and real-world task performance, though these extensions are beyond the scope of the present study.

Conclusion & Future Work

This study emphasizes the importance of the testing environment in evaluating AI conversational abilities through the Turing test. The results show that the *Advanced* Turing test provides a more rigorous and revealing assessment than the *Simple* and *Enhanced* Turing tests. As demonstrated in this paper, when AI is tested under structured and demanding conditions, its performance declines significantly.

Another key finding is that the ability of human evaluators to distinguish AI from humans is influenced by factors such as previous experience with AI. This highlights the need to consider both AI performance and potential human judgment biases when designing evaluation methodologies.

Rather than treating the Turing test as a fixed benchmark, our findings suggest that it should evolve alongside advancements in AI. A more refined and structured approach to evaluation is essential to ensure a meaningful assessment of AI's conversational capabilities.

A limitation of the current study is the modest sample size per condition ($n \approx 28$–41). However, the observed effect sizes were medium to large across all conditions (Cramér's $V = 0.35$ to 0.62), suggesting the study had sufficient statistical power to detect these differences reliably. In total, 422 unique participants completed the study. No human–human baseline condition was collected; future work should include such a comparison to establish an upper bound for human discriminative performance.

Building on these findings, several areas of future research can help improve AI evaluation methodologies.

Expanding AI evaluation to multimodal Turing tests is an essential step. Current tests focus primarily on text-based interactions, but future studies should assess AI in voice and video-based conversations. Evaluating AI's ability to replicate human-like behaviors across different communication channels will provide a more comprehensive understanding of its strengths and limitations.

Long-term interactions are another critical challenge. AI models often struggle to maintain logical consistency and adaptability over extended conversations. Future research should explore whether AI can sustain coherent and contextually aware discussions over time, offering deeper insights into its conversational intelligence.

Additionally, cognitive biases among human evaluators can influence Turing test results. Future studies should focus on developing standardized evaluation frameworks, participant training methods, and bias-mitigation techniques to improve the reliability and objectivity of AI assessments. Understanding how different demographic groups perceive AI can help create more fair and accurate testing methodologies.

Acknowledgment

This work was supported in part by the Ministry of Science and Technology, Israel.

References

Biever, C. (2023). ChatGPT broke the turing test—the race is on for new ways to assess AI. *Nature*, *619*(7971), 686–689. https://doi.org/10.1038/d41586-023-02361-7

Bradeško, L., & Mladenić, D. (2012). A survey of chatbot systems through a loebner prize competition. *Proceedings of Slovenian Language Technologies Society Eighth Conference of Language Technologies*, *2*, 34–37.

Bubeck, S., et al. (2023). Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*. https://doi.org/10.48550/arXiv.2303.12712

Chkroun, M., & Azaria, A. (2024). Autonomous agents for interrogation. *2024 IEEE 36th International Conference on Tools with Artificial Intelligence (ICTAI)*, 686–693.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd). Lawrence Erlbaum Associates.

French, R. M. (1990). Subcognition and the limits of the Turing test. *Mind*, *99*(393), 53–65. https://doi.org/10.1093/mind/XCIX.393.53

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. Wiley.

Jannai, D., et al. (2023). Human or not? a gamified approach to the turing test. *arXiv preprint arXiv:2305.20010*. https://doi.org/10.48550/arXiv.2305.20010

Jones, C. R., & Bergen, B. K. (2024). People cannot distinguish gpt-4 from a human in a turing test. *arXiv preprint arXiv:2405.08007*. https://doi.org/10.48550/arXiv.2405.08007

Jones, C. R., & Bergen, B. K. (2025). Large language models pass the turing test. *arXiv preprint arXiv:2503.23674*. https://doi.org/10.48550/arXiv.2503.23674

Moskvichev, A., et al. (2023). The conceptarc benchmark: Evaluating understanding and generalization in the ARC domain. *arXiv preprint arXiv:2305.07141*. https://doi.org/10.48550/arXiv.2305.07141

Panagoulias, D. P., et al. (2023). Rule-augmented artificial intelligence-empowered systems for medical diagnosis using large language models. *2023 IEEE 35th International Conference on Tools with Artificial Intelligence (ICTAI)*, 70–77.

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on amazon mechanical turk. *Judgment and Decision Making*, *5*(5), 411–419. https://doi.org/10.1017/S1930297500002205

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine Series 5*, *50*(302), 157–175. https://doi.org/10.1080/14786440009463897

Restrepo Echavarría, R. (2025). Chatgpt-4 in the turing test. *Minds and Machines*, *35*(1), 8. https://doi.org/10.1007/s11023-025-09711-6

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, *3*(3), 417–424. https://doi.org/10.1017/S0140525X00005756

Srivastava, A., et al. (2022). Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*. https://doi.org/10.48550/arXiv.2206.04615

Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, *59*(236), 433–460. https://doi.org/10.1093/mind/LIX.236.433

Warwick, K., & Shah, H. (2016). Can machines think? a report on turing test experiments at the royal society. *Journal of Experimental and Theoretical Artificial Intelligence*, *28*(6), 989–1007. https://doi.org/10.1080/0952813X.2014.921734

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. https://doi.org/10.1145/365153.365168