Exploring Alternative Approaches to Contemporary AI Testing

Tomasz Eisenbardt
University of Warsaw, Poland

teisenbardt@wz.uw.edu.pl

Abstract: While the first artificial intelligence test proposed by A. Turing has a long-standing tradition, it is no longer sufficient to meet today's challenges. There is now a strong need for more multi-contextual testing and comparison of AI solutions. Among other things, the emergence of the new Chinese language model, DeepSeek—designed to compete with Western solutions such as ChatGPT, Copilot, and Gemini—has caused quite a stir. It has not only influenced financial markets but also ignited a broader discussion about the quality of contemporary AI systems. As a result, there is a growing need to systematically test and compare these tools. The aim of this paper is to present an original attempt to identify and systematize contemporary approaches to AI testing. The discussion is framed by a reference to the Turing Test and its relevance in the modern context. The author seeks to identify common features between software testing, human intelligence testing, and AI testing. Subsequently, based on a critical analysis of relevant literature and other available sources, the paper outlines and organizes the types of tests currently in use. The conducted considerations allowed for the identification of three trends in contemporary AI testing: tests imitating or referring to human intelligence testing, tests analogous to approaches used in software engineering, and tests based on parameters.

Keywords: Testing of AI, Turing test, AI, Generative AI, GenAI

1. Introduction – the First Verification of Non-Human Intelligence

When discussing AI testing, it is impossible to overlook the first test in its historical context. In 1950, Alan Turing (Turing, 1950) proposed a test for artificial intelligence that involves determining whether a machine can deceive a human to the extent that the human believes they are interacting with another human.

The test is based on a game that Turing himself called the "*imitation game*." In this game, there are three participants: the interrogator, who is isolated from the other players, a human, and a computer. The interrogator's task is to determine, based on interactions with the other two players, which one is the human and which one is the machine.

For decades, it was believed that the Turing Test was impossible to pass. However, the rapid advancement of artificial intelligence has not only enabled many algorithms to exhibit intelligent traits but also to successfully pass the Turing Test — and even surpass it.

It turns out that AI sometimes surpasses human capabilities. What was once deemed impossible—such as defeating chess grandmasters (Deep Blue in 1997) or winning a Go tournament (Google DeepMind's AlphaGo), a game with even more possible moves than chess—is now history (Eisenbardt, 2023). These achievements are merely milestones in AI development.

Thus, intelligence, once considered an exclusively human trait, is no longer reserved solely for people. The Turing test may prove to be completely insufficient.

In this paper, two research questions will be posed. The first concerns the continued validity of the test. The second question is: should we expect that, in the future, additional AI tests will be needed—designed differently and focusing on aspects other than those considered so far?

2. Literature Background

2.1 Turing Test Deficiencies

Many authors criticize the Turing Test for its limitations and imperfections (Saygin, Cicekli, Akman, 2000; French, 2000), pointing out such flaws as bias related to human assessment, narrow scope of application, failure to take into account the fact that AI is already surprising humans in many areas (Mei, et al., 2024; Cameron, 2024), and also such special and very interesting observations as:

- The Chinese room argument – or "imitation of understanding" (Searle, 1980) – is a philosophical argument that (in simple terms) one can communicate in Chinese using a book without knowing the language at all. This is an apt reference to early AI solutions,
- The ELIZA effect refers to the name of the first chatbot of this name. This is a situation where a chatbot imitates intelligence, but in fact uses certain phrases and a certain conversation technique, based on language engineering, to only pretend to understand the content (Łomża, 2021).

Thus, the flaws or limitations of the Turing test had been demonstrated already many years ago. Nevertheless, it was referenced by a certain competition. The Loebner Prize represents exactly this type of competition. It was initiated in 1990 by Hugh Loebner and consisted of a chatbot annual tournament. It boiled down to the evaluation of conversations with chatbots by judges. The winner was selected whose conversation was considered to be the most similar to a conversation with a human. However, the competition was closed in 2020. Besides, it caused a lot of controversy from the beginning (Artiffical.., 2023; Wakefield, 2019; Floridi, Mariarosaria, Turilli, 2009).

So, the Turing test is not very complicated to perform, but at the same time it has many imperfections. The answer to these imperfections is the Feigenbaum's test (Feigenbaum, 2003) based on the selected field area and Ebert's test (Ostrow, 2011; Lee, 2011), regarding voice simulation. But to conclude this part of the study, let a quotation be used: "*imitating a human is only part of what AI can potentially do*" (Nikolić, 2019).

2.2 Competitors of GenAI

A considerable number of studies have already been dedicated to the history of artificial intelligence (Reynoso, 2024; Eisenbardt, 2023). However, the 2020s undoubtedly mark a period of rapid advancement in Large Language Models (LLMs) and Generative AI. One of the definitions is formulated as follows: „*a class of machine learning technologies that can generate new content—such as text, images, music, or video—by analyzing patterns in existing data*" (Brynjolfsson, Li, Raymond, 2023).

As recently as a few months ago (spring 2025), it was still possible to identify the key players in the LLM market. Above all, and unquestionably, these were ChatGPT[1] (Chat Generative Pre-trained Transformer) - produced by OpenAI. However, alternative solutions were proposed by, for example, Microsoft Corporation: Copilot[2], and Alphabet Inc. (owner of Google): Gemini[3], and also Claude[4] and others. At the beginning of 2025, reports emerged about DeepSeek[5] — a Chinese alternative to chatbots based on advanced language models (DeepSeek w praktyce…, 2025; DeepSeek – ujawnia…, 2025). The unexpectedly strong performance, coupled with significantly lower financial investments allocated for its creation, momentarily disrupted the financial markets.

The increasing competition and abundance of available solutions further emphasize the need for comparative tests to evaluate these solutions relative to one another.

3. Research Methodology

3.1 Research Goal and Questions

The aim of this paper is to present selected methods of testing AI, explore potential development trends in the field of testing artificial intelligence solutions, and carrying out a preliminary systematization of AI tests.

The following research questions were posed:

[Q1] Do contemporary AI tests relate to other tests, particularly to human intelligence testing and software testing?

[Q2] What parameters of artificial intelligence are currently being tested?

To achieve the objective and answer the research questions posed, a critical analysis of the relevant literature and a review of selected tools were conducted. The final outcome, depicted in Figure 1, was attained through the systematic organization of the obtained results and the application of creative thinking techniques.

3.2 Research Procedure

In the first step, a literature review on the Turing Test was conducted, followed by a critical evaluation of the test. This stage is reflected in the sections Introduction and Background. The next step involved identifying areas

[1] OpenAI ChatGPT, https://chat.openai.com

[2] Microsoft Copilot, https://copilot.microsoft.com

[3] Google Gemini, https://gemini.google.com

[4] Anthropic Claude, https://claude.ai

[5] DeepSeek, http://www.deepseek.com

of AI testing and comparing them with testing practices in software engineering (chapter Software Engineering and Testing). The subsequent part of the discussion focused on relating AI tests to human intelligence testing (chapter Human and Non-Human Intelligence Testing), with Howard Gardner's widely known Theory of Multiple Intelligences serving as a conceptual foundation.

Following this, an inquiry into the techniques and methods currently used in AI testing was carried out (chapter Contemporary Artificial Intelligence Tests). The issue of AI hallucinations was also addressed (Hallucinations as a Measure of Quality?), along with the parameters that are currently tested in practice (Parameters-Based Testing). All these steps aimed to answer the two research questions posed and to support the development of a proposed classification of AI tests, presented in the Conclusions chapter of this thesis.

4. Results

4.1 Software Engineering and Testing

Software testing is a part of software engineering. It is one of its pillars, and the importance and significance of testing are difficult to overestimate. There are many concepts and classifications of tests (Śmiałek, Rybiński, 2024; Sommerville, 2020; Sacha, 2010, Pressman, 2004). Among the various types of tests, there are unit tests, integration tests, system tests, and adaptive tests (the last two involving the user). Two opposing methods are also employed:

- "white-box (glass-box) testing", which focuses on the code and structure of the software,
- "black-box testing", where inputs and outputs are examined without delving into the software's internals or having knowledge of what is actually happening inside.

It is assumed that a successful test demonstrates the presence of defects, not their absence. In practice, conducting exhaustive testing to show that the software is completely error-free is, in most cases, impossible. Many literary sources indicate that methods similar to those used in software engineering are also applied to artificial intelligence, for example:

- unit tests and integration tests (Goodfellow, Bengio, Courville, 2016; Géron, 2019),
- accuracy and performance tests (Russell, Norvig, 2021; Murphy, 2012),
- robustness and error tests (Szegedy, et al., 2014; Papernot, et al., 2016),
- ethical and security tests (Binns, 2018; Doshi-Velez, Kim, 2017),
- user tests and A/B testing (Kohavi, Tang, Xu, 2020; Amershi, et al., 2019).

Nonetheless, the discourse continues to focus on the assessment of intelligence. Consequently, it is pertinent to draw parallels with human intelligence.

4.2 Human and Non-Human Intelligence Testing

Howard Gardner is an author of the Theory of Multiple Intelligences (MI). The theory was developed in 1983, and the author described it in his book Frames of Mind: The Theory of Multiple Intelligences. This theory is based on the division of the types of intelligence that characterize humans. The types of intelligence have been categorized into eight types, namely:

- logical-mathematical,
- verbal-linguistic,
- naturalistic,
- interpersonal,
- intrapersonal,
- visual-spatial,
- musical,
- bodily-kinesthetic (Gardner, 2009).

Later, H. Gardner added two more types of intelligence to this list: existential intelligence and moral-ethical intelligence.

Thus, human abilities, stemming from their innate intelligence, can influence future development. With proper self-awareness and the right reinforcement of individual traits through appropriate education, the chances of success in specific professions are increased. Therefore, everyone should make the best use of their own natural predispositions.

This theory remains highly popular to this day, but it has also faced significant criticism. The main critiques have been its lack of empirical evidence and the subjectivity of its assessments (Waterhouse, 2006).

An examination of Gardner's theory through the perspective of AI evaluation is warranted. To what extent can it be considered a foundational inspiration? While critiques regarding its lack of objectivity continue to hold relevance, the typology of intelligences it proposes remains invaluable. Given the proliferation of AI solutions and the extensive array of underlying algorithms, it becomes imperative to categorize their evaluation according to frameworks analogous to those advanced by H. Gardner.

In the subsequent part of the discussion, however, it becomes evident that contemporary methods of testing artificial intelligence frequently draw on human intelligence testing. Nevertheless, they are often even more diversified than what is proposed by the Theory of Multiple Intelligences.

4.3 Contemporary Artificial Intelligence Tests

There are many tests to which artificial intelligence is subjected. Here are some of them:

- AIME (American Invitational Mathematics Examination) This test refers to the prestigious American mathematics competition, which is the second stage of the selection process for the U.S. team for the International Mathematical Olympiad [6],
- MATH-500 is a set of 500 diverse mathematical problems, selected from a larger collection MATH Benchmark[7]. It was in this test that DeepSeek proved to be the best, even though the test was prepared by OpenAI,
- MMLU: is a benchmark created to assess the knowledge acquired during pretraining, by evaluating language models [8],
- SWE-bench: a test, which assesses the language model's ability to perform real-world software engineering tasks [9].

Continuing the discussion on the complexity of human intelligence, it should be noted that artificial intelligence can also be complex and can be classified in many different ways. This is why there is such a wide range of tests currently used to evaluate modern artificial intelligence algorithms.

Do the creators and users of AI recognize this complexity of human intelligence, and does it translate into the evaluation of artificial intelligence? In fact, a team of researchers at Google began implementing an initiative in 2022 called Beyond the Imitation Game (BIG-Bench). BIG-Bench gathered scientists from around the world. A set of approximately 200 tests was created, covering various fields of knowledge (Adamska, 2024).

The BIG-Bench comparative test (BIG-Bench Benchmark) is a very popular set of tests designed to evaluate the capabilities and intelligence of large language models (LLMs). BIG-Bench includes tests aimed at assessing (Srivastava, et al., 2022; Chowdhery, et al., 2022; BIG-Bench GitHub repository):

- language understanding,
- logic and reasoning abilities,
- creativity,
- general knowledge,
- mathematical skills,
- understanding of jokes, metaphors, and riddles,
- analysis of emotions and intentions,
- problem-solving in unusual or novel situations,
- understanding of context and cultural nuances [10].

---

6 AIME Benchmark, https://www.vals.ai/benchmarks/aime-2025-03-13

7 MATH 500 Benchmark, https://www.vals.ai/benchmarks/math500-03-26-2025

8 Multi-task Language Understanding on MMLU, https://paperswithcode.com/sota/multi-task-language-understanding-on-mmlu

9 SWE-bench, https://www.swebench.com/

10 BIG-Bench GitHub repository, https://github.com/google/BIG-benchChowdhery

BIG-Bench Benchmark is a subset of BIG-Bench Hard (BBH). It is consisting of 23 challenging tasks from BIG-Bench, in which previous evaluations of language models did not exceed the performance of an average human evaluator. BBH evaluates models using prompting techniques, both through the method of few-shot, jak i chain-of-thought (CoT) (Suzgun, et al., 2022; Wei, et al.,2022; Kazemi, 2025).

It may thus be concluded that contemporary tests assess artificial intelligence in a comprehensive and multi-dimensional way. While the selected tests partially correspond to the Theory of Multiple Intelligences, they also encompass aspects not originally accounted for by the theory.

4.4 Hallucinations as a Measure of Quality?

Errors made by AI are referred to as hallucinations. In other words: hallucination is a response generated by AI that contains false or misleading information presented as fact (Maynez et al., 2020). The problem of AI hallucinations is a frequently discussed and genuinely fascinating topic (Ji et al., 2023; Hicks, Humphries, Slater, 2024; Ye et al., 2023; Tonmoy et al., 2024).

Is it sufficient to define the quality of machine intelligence by how often the machine makes mistakes? Does this approach make sense when applied to humans? After all, 'to err is human.' Therefore, the fact that we, as humans, make mistakes is completely natural. But should a machine also make errors? If so, why? Is it to prevent us from feeling embarrassed, because suddenly the human will be the one making mistakes, while the machine is always infallible?

4.5 Parameters Based Testing

Hallucinations and their frequency are not the only criteria that may indicate the quality of an AI system. The literature identifies numerous parameters that are relevant to the testing of artificial intelligence. Among them, several are particularly noteworthy (in parentheses, the approximate number of occurrences — in thousands — of the phrase 'AI + [parameter]' found in Google Scholar as of June 8, 2025, is provided):

- accuracy (5.860);
- latency (1.360) / performance (6.070);
- precision (4.400) & recall (2.010);
- hallucinations – mentioned above (146);
- generalization ability (1.540);
- consistency (5.350);
- robustness (3.690) / adversarial resistance (82);
- fairness & bias (406);
- explainability (169);
- resource efficiency (5.410).

It is worth noting that several of these parameters can already be tested using widely available Python libraries, such as scikit-learn. In each case, the number of results retrieved was in the thousands, with the most longstanding and commonly used parameters generating the highest volumes.

5. Conclusions

The Turing Test undoubtedly holds great historical significance. Its proposal marked a pivotal moment—both a milestone and arguably the beginning of artificial intelligence development. However, the discontinuation of the Loebner Prize serves as compelling evidence that deceiving a human into mistakenly believing a machine is not a machine is no longer the primary concern. Intelligent machines have been among us for quite some time, and we are fully aware of their presence. What matters now is their ability to perform—and how effectively they do so.

The tests presented above sometimes refer to the Turing Test, while at other times draw on widely used assessments from the field of software engineering. Artificial intelligence is also frequently evaluated using methods analogous to those employed in the testing of human intelligence.

Human intelligence is not homogeneous, as highlighted in Howard Gardner's Theory of Multiple Intelligences (MI), which distinguishes eight distinct types. Following this line of reasoning, a similar classification framework could be proposed for artificial intelligence. Given the vast diversity of AI applications, modern testing is increasingly losing its universality. The BIG-Bench Benchmark initiative, and particularly the BBH project,

demonstrates that such domain-specific work is already being carried out successfully. Contemporary AI tests are often tailored to specific fields or domains of knowledge.

Joseph Weizenbaum, creator of the ELIZA program developed between 1964 and 1966—the first chatbot in history—was notably skeptical of the growing fascination with machines simulating human conversation (Łomża, 2021). His main concern was that machine responses were not the result of genuine comprehension or learning, but rather a superficial simulation of thought. This concern resonates with the modern issue of AI hallucinations. Perhaps a key indicator of artificial intelligence quality could be its low propensity for hallucination—that is, the highest possible percentage of accurate responses. Still, in the context of testing and evaluating AI systems, hallucination frequency is not the only factor warranting attention.

It is also noteworthy that there is a growing trend toward testing AI systems based on selected performance parameters.

The findings are summarized in Figure 1. Based on the conducted analysis and reasoning, it can be concluded that there are at least three principal paradigms in contemporary AI testing. Artificial intelligence tests may thus be categorized as follows:

- those imitating or referring to human intelligence testing;
- those analogous to approaches used in software engineering;
- those based on specific parameters of AI functionality.

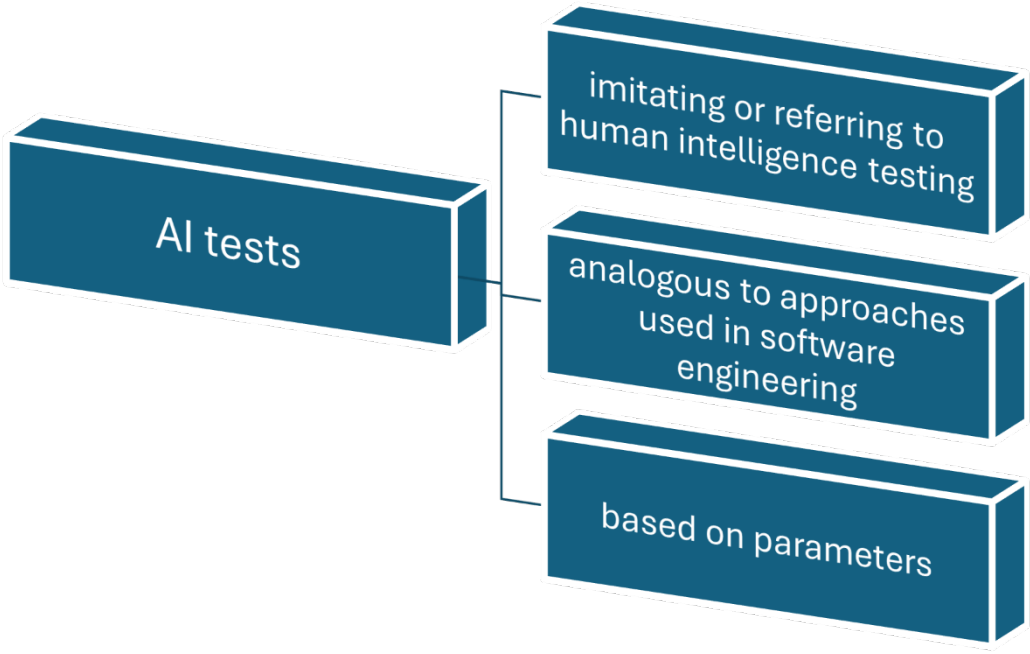

Figure 1. A proposal for the systematization of artificial intelligence testing Source: Author's own work based on the sources listed in the bibliography

In fact, we still view AI through the lens of human intelligence, testing it in ways that mirror human capabilities. However, the range of AI applications is constantly expanding. It was once inconceivable that artificial intelligence could autonomously create artistic content and be utilized in literature, music, painting, and graphic design. This transformation only began with the emergence of large language models (LLMs), and now it is a reality rather than the vision of a futurist. Following this line of reasoning, one can expect that much more will unfold in the realm of AI development and the consequent need for its testing.

6. Limitations and Implications

This study addresses an issue that is undergoing radical technological changes. Some of the presented theses and descriptions of solutions may be temporary and will not be applicable in the case of further, unpredictable developments in artificial intelligence. The presented state of knowledge pertains to the spring of 2025. The author anticipates that the issue of AI testing is currently in its early developmental phase. The described solutions are likely not final.

Ethics declaration: The author declares that there are no conflicts of interest related to the present study. Furthermore, informed consent was obtained from all participants involved in the survey.

AI declaration: A Generative AI tool was employed to assist in the language revision process for this study. The tool was utilized to improve the clarity, coherence, and overall quality of the text, ensuring that the language used throughout the document is free from grammatical errors.

References

Adamska, Z. (2024). Test sztucznej inteligencji: jak sprawdzić, czy maszyna może myśleć? AI Driven, 20 marca, 2024. https://aidriven.pl/ai/biznes/test-sztucznej-inteligencji-jak-sprawdzic-czy-maszyna-moze-myslec/ Access: 2025-04-03.

AIME Benchmark, https://www.vals.ai/benchmarks/aime-2025-03-13, Access: 2025-04-09.

Amershi, S. et al. (2019). Guidelines for Human-AI Interaction. CHI Conference on Human Factors in Computing Systems.

Anthropic Claude, https://claude.ai, Access: 2025-04-09.

Artificial stupidity (2023), Salon.com, 16 February 2003, https://web.archive.org/web/20120720014628/http://www.salon.com/2003/02/26/loebner_part_one. Access: 2025-04-03.

BIG-Bench GitHub repository, https://github.com/google/BIG-benchChowdhery, Access: 2025-04-09.

Binns, R. (2018). Fairness in Machine Learning: Lessons from Political Philosophy. Proceedings of the Conference on Fairness, Accountability, and Transparency.

Brynjolfsson E., Li D., Raymond L. R. (2023). Generative AI at Work. DOI 10.3386/w31161

Cameron, S. (2024). Study finds ChatGPT's latest bot behaves like humans, only better (2024). Stanford University, https://humsci.stanford.edu/feature/study-finds-chatgpts-latest-bot-behaves-humans-only-better. Access: 2025-04-03.

Chowdhery, A., Narang, S., Devlin, J. ., et al. (2022). PaLM: Scaling Language Models with Pathways. arXiv preprint arXiv:2204.02311. https://doi.org/10.48550/arXiv.2204.02311

DeepSeek - ujawnia nowy paradygmat prowadzący do nadludzkiej sztucznej inteligencji. (2025). i-kumam. https://www.youtube.com/watch?v=W0Hs8PGYEsk, Access: 2025-02-03.

DeepSeek w praktyce. Testujemy chińskie AI w porównaniu z ChatGPT i Copilot. (2025). Komputer świat. https://www.youtube.com/watch?v=H7DXhGvVoLY. Access: 2025-04-03.

DeepSeek, http://www.deepseek.com, Access: 2025-04-09.

Doshi-Velez, F., & Kim, B. (2017). Towards a Rigorous Science of Interpretable Machine Learning. arXiv.

Eisenbardt, T. (2023). Intelligent Virtual Assistants and New Perspectives of Knowledge Acquisition. Proceedings of the 24th European Conference on Knowledge Management, ECKM 2023. DOI: https://doi.org/10.34190/eckm.24.1.1421

Feigenbaum, E. A. (2003). Some challenges and grand challenges for computational intelligence. Journal of the ACM. 50 (1): 32–40. doi:10.1145/602382.602400. S2CID 15379263.

Floridi, L., Mariarosaria, T., Turilli, M. (2009). Turing's Imitation Game: Still an Impossible Challenge for All Machines and Some Judges—An Evaluation of the 2008 Loebner Contest. Minds and Machines. 19 (19): 145–150. CiteSeerX10.1.1.672.7417. doi:10.1007/s11023-008-9130-6. S2CID7888490.

French, R. M. (2000). The Turing Test: The First 50 Years. Trends in Cognitive Sciences, 4(3), 115–121.

Gardner, H. (2009). Inteligencje wielorakie. Nowe horyzonty w teorii i praktyce, Laurum, Warszawa.

Géron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow. O'Reilly Media.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.

Google Gemini, https://gemini.google.com, Access: 2025-04-09.

Hicks, M.T., Humphries, J. & Slater, J. (2024). ChatGPT is bullshit. Ethics Inf Technol 26, 38. https://doi.org/10.1007/s10676-024-09775-5

Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y. J.; Madotto, A.; Fung, P. (2023). Survey of Hallucination in Natural Language Generation, https://doi.org/10.48550/arXiv.2202.03629

Kazemi, M., Fatemi, B., Bansal, H., Palowitch, J., Anastasiou, Ch., et al. (2025). BIG-Bench Extra Hard, https://doi.org/10.48550/arXiv.2502.19187.

Kohavi, R., Tang, D., & Xu, Y. (2020). Trustworthy Online Controlled Experiments: A Practical Guide to A/B Testing. Cambridge University Press.

Lee, J. 8. (2011). Roger Ebert Tests His Vocal Cords, and Comedic Delivery. The New York Times. March 7, 2011. Access: 2025-04-03.

Łomża, Ł. (2021). Pogadać z maszyną 2/4 | ELIZA, czyli pierwszy bot i od razu problemy https://www.youtube.com/watch?v=iRs-nC8vEro. Access: 2025-03-29.

MATH 500 Benchmark, https://www.vals.ai/benchmarks/math500-03-26-2025, Access: 2025-04-09.

Maynez, J.; Narayan, S.; Bohnet, B.; McDonald, R. (2020). "On Faithfulness and Factuality in Abstractive Summarization". Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 1906–1919. doi:10.18653/v1/2020.acl-main.173.

Mei, Q. , Xie, Y. , Yuan, W. & Jackson, M.O. (2024). A Turing test of whether AI chatbots are behaviorally similar to humans, Proc. Natl. Acad. Sci. U.S.A. 121 (9) e2313925121, https://doi.org/10.1073/pnas.2313925121

Microsoft Copilot, https://copilot.microsoft.com, Access: 2025-04-09.

Multi-task Language Understanding on MMLU, https://paperswithcode.com/sota/multi-task-language-understanding-on-mmlu, Access: 2025-04-09.
Murphy, K. P. (2012). Machine Learning: A Probabilistic Perspective. MIT Press.
Nikolić, Danko (Apr 9, 2019). "Is the Turing test still relevant? How about Turing time?". Medium. https://medium.com/savedroid/is-the-turing-test-still-relevant-how-about-turing-time-d73d472c18f1. Access: 2025-04-03.
OpenAI ChatGPT, https://chat.openai.com, Access: 2025-04-09.
OpenAI, https://openai.com, Access: 2025-04-09.
Ostrow, A. (2011). Roger Ebert's Inspiring Digital Transformation. Mashable Entertainment. 5 March, 2011. Access: 2025-04-03.
Papernot, N. et al. (2016). The limitations of deep learning in adversarial settings. IEEE.
Pressman R., S. (2004). Praktyczne podejście do inżynierii oprogramowania. Wydawnictwo WNT, Warszawa.
Reynoso R. (2024). Brief History of Artificial Intelligence - From 1900 till Now. https://www.g2.com/articles/history-of-artificial-intelligence. Access: 2025-04-04.
Russell, S., & Norvig, P. (2021). Artificial Intelligence: A Modern Approach. Pearson.
Sacha K. (2010). Inżynieria oprogramowania. Wydawnictwo Naukowe PWN, Warszawa.
Saygin, A. P., Cicekli, I., & 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), 417–457.
Śmiałek M., Rybiński K. (2024). Inżynieria oprogramowania w praktyce. Od wymagań do kodu z językiem UML. Helion, Gliwice.
Sommerville I. (2020). Inżynieria oprogramowania. Wydawnictwo Naukowe PWN, Warszawa.
Srivastava, A., Rastogi, A., Le, Q., 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
Suzgun, M., Scales, N., Kapoor, S., Wei, J., Tay, Y., et al. (2022). Challenging BIG-Bench Tasks and Whether Chain-of-Thought Can Solve Them. arXiv preprint arXiv:2210.09261. https://arxiv.org/abs/2210.09261. Access: 2025-04-09.
SWE-bench, https://www.swebench.com/, Access: 2025-04-09.
Szegedy, C. et al. (2014). Intriguing properties of neural networks. ICLR.
Tonmoy, T.I.; Zaman M.; Jain, V.; Rani A.; Rawte, V.; Chadha, A.; Das, A. (2024). A Comprehensive Survey of Hallucination Mitigation Techniques in Large Language Models, DOI: 10.13140/RG.2.2.11724.39045
Turing A. (1950) Computing Machinery and Intelligence. Mind. vol. LIX. no. 236. October 1950. pp. 433-460
Wakefield, J. (2019). The hobbyists competing to make AI human. BBC 14 September 2019. https://www-bbc-com.translate.goog/news/technology-49578503?_x_tr_sl=en&_x_tr_tl=pl&_x_tr_hl=pl&_x_tr_pto=rq&_x_tr_hist=true. Access: 2025-04-03.
Waterhouse, L. (2006). Multiple Intelligences, the Mozart Effect, and Emotional Intelligence: A Critical Review. Educational Psychologist, 41(4), 207–225. https://doi.org/10.1207/s15326985ep4104_1
Wei, J., Tay, Y., Bommasani, R., Raffel, C., et al. (2022). Emergent Abilities of Large Language Models. Transactions on Machine Learning Research (TMLR). https://arxiv.org/abs/2206.07682. Access: 2025-04-09.
Ye, H.; Liu,T.; Zhang, A.; Hua, W.; Jia, W. (2023). Cognitive Mirage: A Review of Hallucinations in Large Language Models, https://doi.org/10.48550/arXiv.2309.06794