Turing Test in the Era of LLM

Hengli Li
Yuanpei College
Peking University
2000017754@stu.pku.edu.cn

Abstract

The Turing test has long been a cornerstone for evaluating machine intelligence through conversational interactions with humans. However, as the field of natural language processing (NLP) has rapidly advanced, concerns have arisen regarding the test's ability to assess the depth of a machine's comprehension versus its capacity to mimic human speech. This paper reimagines the evaluation of machine understanding, introducing two pivotal properties: pragmatic reasoning and commonsense. Pragmatic reasoning delves into a machine's capacity to engage in real-world scenarios, considering context and practical implications, while commonsense evaluates an AI system's ability to comprehend the embodied attributes and nuanced meanings of words. These core properties provide a comprehensive and meaningful assessment of machine understanding, addressing the limitations of both the traditional Turing test and contemporary benchmarks. In the era of large language models (LLMs), this reimagined approach guides the development of more sophisticated and human-like AI agents, bringing us closer to the realization of truly intelligent artificial agents.

1 Introduction

The Turing test, as originally proposed by TURING [5], has long served as a method for assessing machine intelligence by engaging humans in a conversational evaluation. However, as the field of natural language processing (NLP) has advanced rapidly, the emergence of large language models (LLMs) has raised concerns about the suitability of the Turing test for gauging the depth of a machine's comprehension as opposed to its ability to mimic human speech acts probabilistically.

From a practical standpoint, the Turing test faces limitations due to the time-consuming nature of human interactions and the diversity of topics that can be discussed with machines. These factors undermine the effectiveness of the test as a comprehensive measure of true machine understanding.

Notably, contemporary evaluation benchmarks, although valuable, also exhibit limitations. To address these concerns, we posit that a robust evaluation approach should encompass the following two essential properties:

(I) Pragmatic Reasoning: A well-functioning evaluation method should assess a machine's ability to engage in pragmatic reasoning. This involves evaluating the machine's capacity to comprehend and respond appropriately to real-world scenarios, considering context and practical implications.

(II) Commonsense: Machine understanding should extend beyond mere data-driven patterns and encompass commonsense reasoning. A capable evaluation approach should gauge a machine's ability to apply commonsense knowledge and reasoning to diverse situations, demonstrating an understanding of the world.

By embracing these core properties, an evaluation approach can provide a more comprehensive and meaningful assessment of a machine's capacity for understanding, thus addressing the limitations of both the Turing test and existing evaluation benchmarks.

Figure 1: Emergence Grounding of Human Words

2 Characteristics of "Turing Test" in the era of LLM

2.1 Pragmatic Reasoning

Indeed, pragmatic reasoning plays a pivotal role in the development and evolution of human language, as elucidated by Frank and Goodman [2] and Lazaridou et al. [3]. The ability to grasp the nuanced relationship between words and real-world scenarios is a remarkable characteristic of human language and cognition, serving as a foundation for effective communication.

As depicted in Figure 1, machines have made notable strides in associating words with contextual scenarios, showing a degree of understanding. In contemporary large language models (LLMs), this capability is often referred to as pragmatic reasoning, and it represents a critical step toward creating more human-like agents. However, it is evident, as demonstrated by Li et al. [4], that current LLMs exhibit deficiencies in this aspect, failing to fully capture the nuances of pragmatic reasoning.

In light of these limitations, we contend that a reimagined approach to testing the abilities of language models and AI systems should place pragmatic reasoning at its core. Evaluations and benchmarks that comprehensively assess a model's capacity for pragmatic reasoning will not only provide a more accurate measure of its language understanding but also guide the development of more sophisticated and human-like AI agents.

2.2 Commonsense

Commonsense, a longstanding and essential topic in the field of AI, has seen considerable advancements in recent years, driven in large part by the development of large language models (LLMs). These models indeed possess a vast amount of commonsense knowledge, which enables them to understand and generate text with a high degree of coherence and relevance.

However, it is equally true that LLMs exhibit certain limitations when it comes to commonsense reasoning, particularly in the context of embodied attributes or real-world understanding. As words used by humans often carry a wealth of embodied meanings and connotations, evaluating an AI system's capacity to comprehend these aspects of commonsense is of paramount importance [1].

In this context, a comprehensive evaluation approach that encompasses multiple facets of commonsense reasoning is crucial for a modern-day "Turing Test" in the era of large language models. Such an evaluation should extend beyond linguistic coherence and engage AI systems in tasks that demand a deeper understanding of the world and its embodied attributes. By testing an AI system's grasp of commonsense across various dimensions, we can better gauge its true understanding and ability to interact effectively in real-world scenarios.

3 Conclusion

In conclusion, as the field of natural language processing has witnessed tremendous advancements with the emergence of large language models (LLMs), it has become increasingly important to reevaluate the traditional Turing test and contemporary evaluation benchmarks. While these tests have served as valuable tools for assessing machine intelligence, they have inherent limitations that are particularly apparent in the era of LLMs. To address these limitations, we have proposed two core properties to consider to evaluating machine understanding, one that places two crucial properties at its core: pragmatic reasoning and commonsense. Pragmatic reasoning, encompassing the ability to engage with real-world scenarios and respond contextually, is a fundamental aspect

of human language and communication, and it represents a key milestone in the development of more human-like AI agents. Commonsense, on the other hand, extends beyond linguistic coherence, delving into the embodied attributes and nuanced meanings that words carry in the real world. By embracing these core properties in an evaluation approach, we can provide a more comprehensive and meaningful assessment of a machine's capacity for understanding. This approach not only addresses the limitations of the Turing test but also guides the development of more sophisticated AI systems that can interact effectively in diverse real-world scenarios. In the era of LLMs, it is crucial that our evaluation methods evolve to meet the demands of assessing these advanced language models accurately. By doing so, we can foster the development of AI systems that not only mimic human speech probabilistically but truly understand and engage with the world, bringing us closer to the realization of more capable and intelligent artificial agents.

References

[1] Nicolas Fay, T. Mark Ellison, and Simon Garrod. Iconicity: From sign to system in human communication and language. *Pragmatics Cognition*, 22:243–262, 12 2014. doi: 10.1075/pc.22.2.05fay. 2

[2] Michael C. Frank and Noah D. Goodman. Predicting pragmatic reasoning in language games. *Science*, 336(6084):998–998, 2012. doi: 10.1126/science.1218633. URL https://www.science.org/doi/abs/10.1126/science.1218633. 2

[3] Angeliki Lazaridou, Alexander Peysakhovich, and Marco Baroni. Multi-agent cooperation and the emergence of (natural) language, 2017. 2

[4] Hengli Li, Song-Chun Zhu, and Zilong Zheng. Diplomat: A dialogue dataset for situated pragmatic reasoning. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. 2

[5] A. M. TURING. COMPUTING MACHINERY AND INTELLIGENCE. *Mind*, LIX(236): 433–460, 10 1950. ISSN 0026-4423. doi: 10.1093/mind/LIX.236.433. URL https://doi.org/10.1093/mind/LIX.236.433. 1