The WinnerGrad Competition – Testing for Intelligence

Introduction

In a world where AI continually pushes boundaries, the WinnerGrad competition introduces an innovative dimension by challenging AI systems not only to solve problems but also to predict their own accuracy. This evolution promises to shed light on the fascinating realm of self-awareness in AI, offering a new perspective on the capabilities and limitations of these advanced systems.

Human Intelligence

Human intelligence is an intricate quality shaped by a complex interplay of genetic and environmental factors (MedlinePlus, 2007). Although consensus on the precise components influencing human intelligence remains elusive, there is a widespread acknowledgment of genetics playing a significant role. There have been ongoing debates on the predominant factor of intelligence; hereditary versus environmental.

Genetic research has identified specific genes linked to cognitive abilities, but it is crucial to understand that intelligence is a polygenic trait. Environmental factors, such as early experiences, upbringing, socioeconomic status, nutrition, intellectual stimulation, emotional intelligence, cultural influences, and neurobiological elements, also exert significant influence on intelligence (Cherry, 2022).

Moreover, personality traits, motivation, and the effort put into learning all contribute to intellectual development. The complex nature of human intelligence continues to be a subject of ongoing scientific exploration, with genetic factors estimated to account for a substantial portion of individual differences and certainly attributed to context-sensitivity aspect of human intelligence.

Context-Sensitive Intelligence

Context-sensitivity is a crucial aspect of human intelligence that enables individuals to adapt their general knowledge and behaviour according to the specific situation or environment they are in. This ability to perceive and respond to contextual cues is deeply rooted in the human cognitive system and is essential for effective problem-solving, decision-making, and social interaction.

Context-sensitivity extends beyond individual cognition to social and cultural dimensions. Cultural psychology, for instance, highlights how cultural context influences cognitive processes and behavioural patterns (Nisbett and Miyamoto, 2005). In communication and language, understanding the context in which words or phrases are used is essential for effective interpretation and expression.

Testing For Intelligence

In the realm of simulating human intelligence, scientists have devised tests like the Turing Test and the Winograd Schema Challenge. The Turing Test evaluates a machine's ability to mimic human conversation, engaging in dialogue indistinguishable from human interaction.

The Winograd Schema Challenge (WSC) was presented in 2012 as a benchmark for testing AI's ability to comprehend and apply commonsense knowledge in language understanding (Levesque, Davis and Morgenstern, 2012). The challenge consists of pairs of sentences differing only by a single pronoun, requiring machines to correctly identify the antecedent of the pronoun, often relying on contextual commonsense reasoning.

The WSC's significance extends beyond its immediate task, showcasing the strides made in natural language understanding and commonsense reasoning by AI systems. It underscores the growing role of pre-trained language models in various applications, including chatbots, search engines, and automated content generation.

Overall, the WSC serves as a testament to the rapid advancement in AI's natural language processing capabilities over the past decade, with implications for improving the sophistication of human-machine interaction and expanding AI's potential in understanding and generating human-like text.

WinoGrande is a dataset of 44'000 problems presented by Sakaguchi et. al. (2020), based on the original WSC, with modifications to make it more challenging. For instance, the problems no longer appear in pairs and are crowdsourced from a large number of people (rather than being crafted by a small pool of experts), so the language used is more diverse.

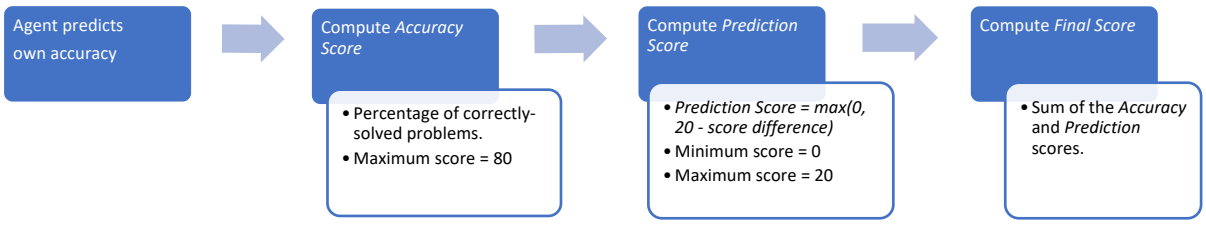
The WinnerGrad Competition

WinoGrande problems are presented as sentences containing a blank, which the agent is tasked with filling in, given two options, as in the following (the correct answer is in **bold**):

- "Ian volunteered to eat Dennis's menudo after already having a bowl because _ enjoyed eating intestine." Options: **Ian**/Dennis
- "He never comes to my home, but I always go to his house because the _ is smaller." Options: **home**/house

We propose a new competition, the *WinnerGrad* competition, that extends the WinoGrande challenge with a predictive element. Before an agent solves any WinoGrande problems it must first predict its overall accuracy will be. After solving the problems this prediction is graded and used to produce a final score.

No agent is permitted to access to the internet or other external resource during the competition. Whilst recognising when a task is beyond an agent's ability is valuable, this can simply be recorded, which would be more informative than a system that can just look up the solution to a question.



Discussion

We consider WinnerGrad to be highly feasible. Given tests such as the WinoGrande Schema have been successfully administered multiple times, we foresee no issue in administering this challenge. Additionally, since there will not be any real-time or heavy computational demands, the challenge remains feasible from a technical perspective. Most modern AI models can solve problems from a dataset like WinoGrande relatively quickly.

Although the most sophisticated large language model, GPT-4, has achieved 87.5% on the WinoGrande challenge (OpenAI, 2023), the introduction of self-reflection within WinnerGrad increases the difficulty significantly. It is therefore expected that many modern AI systems such as GPT-4 will struggle to score highly with this novel approach, achieving scores that will fall below 70%.

The prediction phase will therefore likely remain the most challenging aspect of the competition since self-awareness in AI models is a largely unexplored area. Current models do not have an internal state or self-model that allows them to estimate future performance on a given task. Additionally, it is not uncommon for AI systems to be sure they are correct until explicitly corrected.

As Natural Language Understanding technologies continue to evolve, it is likely that future models will perform better on the WinoGrande problems. However, it is not clear if they will significantly improve on the ability to predict their own performance, given the complexity of how these models are constructed.

Should research in the direction of self-awareness becomes more established, models could be developed that are much better equipped to assess their own capabilities. This could make the tests even more meaningful given the hurdle self-assessment presents now. To better evaluate this, another task could be introduced, requiring the model to rate its confidence in its own solutions.

References

1. Bender, D., 2015. Establishing a Human Baseline for the Winograd Schema Challenge. In M. Glass and J. H. Kim, eds. *Proceedings of the 26th Modern AI and Cognitive Science Conference, 25-26 April 2015, Greensboro, NC, USA,* pp. 39–45.
2. Cherry, K., 2022. *The Nature vs. Nurture Debate* [Online]. Available from: https://www.verywellmind.com/what-is-nature-versus-nurture-2795392 [Accessed 1 Oct 2023].
3. Davis, E., 2021. Using human skills taxonomies and tests as measures of artificial intelligence. *AI and the Future of Skills, Volume 1.* OECD.
4. Lave, J. and Wenger, E., 1991. Legitimate Peripheral Participation. In *Situated Learning*. Cambridge University Press, pp. 27–44.
5. Levesque, H. J., Davis, E. and Morgenstern, L., 2012. The Winograd Schema Challenge. In G. Brewka, T. Eiter and S. A. McIlraith, eds. *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning, 10-14 June 2012, Rome, Italy,* pp. 552–561.

6. MedlinePlus, 2007. *Is intelligence determined by genetics?* [Online]. Bethesda (MD): National Library of Medicine (US). Available from: https://medlineplus.gov/genetics/understanding/traits/intelligence [Accessed 1 Oct 2023].

7. Miller, E. K. and Cohen, J. D., 2001. An Integrative Theory of Prefrontal Cortex Function. *Annual Review of Neuroscience, Volume 24*, pp. 167–202.

8. Nisbett, R. E. and Miyamoto, Y., 2005. The influence of culture: holistic versus analytic perception. *Trends in Cognitive Sciences, Volume 9*, pp. 467–473.

9. OpenAI, 2023. *GPT-4 Technical Report*. San Francisco: OpenAI.

10. Sakaguchi, K., Bras, R. L., Bhagavatula, C. and Choi, Y., 2020. WinoGrande: An Adversarial Winograd Schema Challenge at Scale. In *Proceedings of the AAAI Conference on Artificial Intelligence, 7-12 February 2020, New York, USA,* pp. 8732–8740.