Quantifying Intelligence

Joseph Wargo
Former Student, University of Georgia
josephawargo@gmail.com

Abstract

In this paper, we first explore the ways in which performance is quantified in the realm of Artificial Intelligence. Then, it seeks to find which of these performance-quantifying methods methods, if any, are suitable to judge a model's intelligence, and perhaps establish that model as intelligent.

1 Introduction

Intelligence is the defining characteristic of Artificial Intelligence. It is, well, in the name. Dating back to at least Alan Turing and his eponymous test, and perhaps as far back as the ancient Greeks, (Ovid, 43 B.C.-17 A.D. or 18 A.D.) humans have attempted to define the intelligence of artificial machines. These attempts have resulted in a myriad of definitions of intelligence, which vary greatly across fields of study (Legg and Hutter, 2007). In recent years, as machine learning has become the leading edge of Artificial Intelligence, attempts to quantify model performance benchmarking tests have become essential to the field (Riezler and Hagmann, 2021).

2 Natural Langauge Processing Benchmarks

Large Language Models are the technology de jure of Natural Langauge Processing. To quantify efficacy in certain tasks, authors will test their model's performance on a variety of benchmarks across different fields of NLP. Benchmarks are, essentially, datasets of text on which a Language Model can be tested.

2.1 Grammatical Acceptability

Grammatical Acceptability is, put simply, "detecting whether or not a sentence contains a grammatical error" (Wagner et al., 2013).

One leading benchmark in this field is BLiMP (Warstadt et al., 2020). BLiMP gives a Language Model two sentences that vary only on one grammatical phenomenon. One of the sentences is grammatically acceptable, and the other is not. If the Language Model assigns a higher probability to the grammatically acceptable sentence, then it is judged to have performed the task correctly. The BLiMP dataset contains 67,000 of such pairs.

Another benchmark in this field is CoLA, which works by training a Language Model on acceptability classification (Warstadt et al., 2019). The CoLA dataset is 10,657 sentences.

2.2 Completion

Completion is judged by supplying a Language Model some text with a missing element, and having it try to generate that element. In some benchmarks, there will be a single sentence that is missing a singular word or a punctuation mark. Other times, there will be a set of sentences and the Language Model will be tasked with generating whole sentences to follow.

While Language Models such as BERT (Devlin et al., 2018) performed well on existing Completion benchmarks upon initial release, recent datasets such as HellaSwag, which proved trivial for humans, gave the previously successful Language Models a significant challenge (Zellers et al., 2019).

3 Natural Language Understanding Benchmarks

Natural Language Understanding is a subset of Natural Language Processing which deals

with machine understanding of text. Just like they are on other NLP benchmarks, Language Models are evaluated on NLU benchmarks.

3.1 Winograd

The Winograd Schema Challenge, proposed by Hector J. Levesque, is an edification of the Turing test (Levesque et al., 2012). Levesque took issue with the Turing test's promotion of trickery by the computer, as well as its reliance on free-form conversation, among other things. The challenge consists of a set of problems of textual entailment, which follow the below structure given in the paper:

> **Sentence:** The trophy doesn't fit in the brown suitcase because it's too small.
>
> **Question:** What is too small?
>
> **Answer 0:** the trophy
>
> **Answer 1:** the suitcase

If a Language Model answers the question correctly, it is deemed successful. There are 273 such questions in the original Winograd Schema Challenge.

The WinoGrande Schema Challenge, a 2020 update of the original with a larger amount of more difficult problems, provides a tougher challenge to state-of-the-art Language Models (Sakaguchi et al., 2020).

3.2 Question Answering

Question Answering determines how well an entity can answer a query when given a source of knowledge that contains the answer to said query. For example, an entity is given the Wikipedia page for the United States of America, and is asked:

> How many states are there in the United States of America?

Question Answering is of elevated importance for search engines. Some cutting edge Question Answering benchmarks (Kwiatkowski et al., 2019) use real-life Google searches as questions.

3.3 Higher-Level Question Answering

As Language Models solve Question Answering benchmarks at increasingly higher rates, more complex datasets are made to further challenge inductive capabilities. Benchmarks such as DROP (Dua et al., 2019) aim to judge *reading comprehension* by asking abstract questions instead of more direct queries. SuperGLUE (Wang et al., 2019), an upgraded version of GLUE (Wang et al., 2018) on which Language Models are commonly tested, contains a wide array of Question Answering benchmarks, including different types of *reading comprehension*.

3.4 Chosen Benchmarks

In an effort to make comparisons across models easier, there have been attempts to standardize benchmarks. Three recent Language Models, all of which were created by an industry leader (Google or OpenAI) (Chowdhery et al., 2022), (Du et al., 2021), (Brown et al., 2020), have been tested on the same 29 benchmarks, which span several different NLP and NLU categories.

3.5 Further Benchmarks

Some state of the art Language Models are evaluated on metrics beyond basic NLP and NLU benchmarks that may provide us a deeper understanding of a model's intelligence. Google's Pathways Language Model (PaLM) (Chowdhery et al., 2022), which performed the best of all models on 28 out of the 29 benchmarks referred to as the "most widely evaluated", went further in evaluation by judging the model's ability to solve math word problems, develop code, and translate between languages.

3.6 What they mean

The 29 benchmarks referred to above are the "most widely evaluated English language understanding benchmarks" (Chowdhery et al., 2022). Still, new and effective benchmarks are created all the time (Shaham et al., 2022). Language is a wide category, and there is seemingly always an opportunity to go deeper or

wider in its exploration. This leads back to a question posed in this paper's introduction. Are these benchmarks adequate for evaluating intelligence?

These benchmarks are good at what they do: quantifying Language Model performance. Some benchmarks, such as the Winograd Schema Challenge (Levesque et al., 2012), initially aimed to be a measure of intelligence. However, Language Models have improved and performed increasingly better at the challenge (Kocijan et al., 2022) without becoming definitively intelligent. Thus, Winograd-esque challenges can be viewed similarly to other benchmarks such as Grammatical Acceptability, and not as broader measure of intelligence on the scale of the Turing test. Those who create newer benchmarks are often more conservative when estimating the power of their dataset.

These benchmarks allow for Language Models to be compared to human performance. However, specific benchmarks are narrow, and allow only for human comparison within a particular NLP section. Even a battery of these benchmarks, while effective in judging a Language Model's performance on a wide array of NLP tasks, may fail to capture the breadth and depth of a model's true capabilities. This may very well be by design. Barring a few exceptions, creators of benchmarks do not assert that intelligence is shown through success in their tests.

4 More General Comparisons to Humans

4.1 Turing Test

The most well known example of machine-human comparison is the Turing test (Turing, 1950). The test is Turing's answer to his self-imposed question "Can machines think?". It works as such:

There are three participants in a game, a human (A), a machine (B), and an interrogator (C). Each participant has a separate goal. The human's (A) goal is to convince the interrogator (C) that they are really a human. Similarly, the machine's (B) goal is to convince the interrogator (C) that it is the human (B). The interrogator's (C) goal is to determine which of A and B is a human, and which is a machine. All three entities are in separate rooms, with the interrogator (C) located between the human (A) and the machine (B). Communication is either done via typing, writing, or an intermediary. The interrogator (C) takes turns asking the human (A) and the machine (B) questions, attempting to identify which one is really the human. If the machine (C), over a significant sample of tests, is able to convince a wide array of interrogators that it is indeed the human (B), would that machine (D) be intelligent? In this specific thought experiment, Turing believes so.

There have been several proposed improvements to the Turing Test, such as the Loebner Prize (the). In all these renditions, the core idea remains the same. A machine is intelligent if it is able to convince enough humans that it is a human.

4.2 Chess

A seminal moment in Artificial Intelligence came in 1997, when the IBM chess engine DeepBlue defeated then world number one ranked Gary Kasparov in a match of six games (Campbell et al., 2002). Chess has long been considered a robust expression of human intelligence and intellectual creativity. DeepBlue's besting of Kasparov disrupted this assumption, and foreshadowed the computer dominance of chess to come.

DeepBlue was able to succeed without neural networks or any sort of deep learning. Nowadays, Chess computers that do employ these techniques are leagues ahead of the best human players. In 2017, Google used deep convolutional neural networks to create AlphaZero, which had a then-unprecedented estimated Elo rating of around 3400 (Silver et al., 2017). Stockfish, the current top Chess computer, incorporated many of AlphaZero's techniques into its existing search-based structure. It is now significantly better than AlphaZero, and therefore the best humans, with an estimated Elo rating of 3544.

Nowadays, all kinds of AI compete at chess.

The Language GPT-2 has been tweaked to be trained on a dataset of chess games, and is able to play games in a way that resembles a human (Noever et al., 2020).

4.3 Further Creative Expressions

Every day, AI that produces creative writing (Brown et al., 2020), composes music (Payne, 2019), and creates works of visual art (dal, 2022). As AIs with these capabilities grow in number and competency, their performance against human creation could provide insight into intelligence.

5 Defining Machine Intelligence

In 2007, Shane Legg and Marcus Hutter collected 71 definitions of intelligence (Legg and Hutter, 2007). 18 of the definitions came from collectives like dictionaries or the American Psychological Association. 35 came from psychologists. 18 more came from AI researchers. Using this data, they synthesized their own definition of *universal intelligence*:

> "Intelligence measures an agent's ability to achieve goals in a wide range of environments."

Later that year, they further explored definitions of both *human* and *machine intelligence*, and produced a formal definition of *machine intelligence*, which is a mathematical representation of their *universal intelligence*. I will take inspiration from the similarities between these two definitions for *machine* and *human intelligence* in proposing my own ideas.

We can judge how well an AI does conversing with humans, a la Turing test. We can quantify their performance on a battery of benchmarking tests. Similarly, we can judge how well a human does conversing with humans. We can quantify a human's performance on a battery of benchmarking tests.

But *human intelligence* is not properly defined by these metrics. It is understood via an amalgamation of interpersonal interaction, pattern recognition, academic performance, etc. This is holistic and inherently subjective. It is *human intelligence* that allows humans to complete benchmarking tests. It is *human intelligence* that allows a human to converse with another. But *human intelligence* is not solely, or, I would argue, significantly defined by just those attributes.

So why are we treating machines different than humans? I propose that, in seeking to quantify the intelligence of AI, we treat that AI as we treat a human. What does this entail?

One tricky part of this is that every human has a presumption of intelligence. It is not necessary for you or I or anyone to go through a series of tests to prove whether we are intelligent. While we undergo testing in school, strangers we meet are unaware of our results on those assessments, just as we are unaware of their results, or if they even went to school at all. AI, with good reason, is not given that presumption.

Inherent in the assumption that a human is intelligent is another assumption that if one were to perform a battery of tests on a human to determine if they are intelligent, they would pass. I believe that this presumption of intelligence reveals to us a two pronged system for identifying if an AI is intelligent. An AI must:

> 1. Pass the same battery of tests that would be provided to a human in the event that they needed to prove their intelligence.

and

> 2. Be able to interact with both humans and other machines in a way that does not cause others to question their intelligence.

5.1 Which Tests?

The battery of tests necessary to confirm intelligence, whether that be of a human or a machine, must be both varied and comprehensive. Here is a preliminary list of what I envision:

> 1. Benchmarks currently used to evaluate Language Models
>
> 2. Quantitative reasoning tests (e.g. IQ, SAT)

3. Tests on the ability to learn "creative" skills (e.g. Chess, creative writing, music

5.2 Blending In

The second part of my proposed intelligence test is similar to the Turing test, although it requires a machine to blend in, rather than trick an interrogator. Obviously, robotics are far from the point where an artificial body would, given adequate observation, be confused with a human body. Blending in on the internet would, in my opinion, suffice for this test.

5.3 Conclusion

I believe that *machine intelligence* should be considered fundamentally the same as *human intelligence*. By taking a two-pronged approach with a battery of tests and blending in, we can accurately judge whether an entity, human or machine, is truly intelligent.

Acknowledgements

Once again, I would like to thank Dr. Hale for giving me the idea for this project and leading me down this rabbit hole. It has been a difficult yet rewarding journey, and I have learned so much. Also, once again, I would like to thank Mitchell Ostrow, now an incoming PhD Student at MIT Brain and Cognitive Sciences!!!, for his assistance.

References

The loebner prize.

2022. Dall-e 2.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Murray Campbell, A.Joseph Hoane, and Feng hsiung Hsu. 2002. Deep blue. *Artificial Intelligence*, 134(1-2):57–83.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

Nan Du, Yanping Huang, Andrew M. Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, Yanqi Zhou, Adams Wei Yu, Orhan Firat, Barret Zoph, Liam Fedus, Maarten Bosma, Zongwei Zhou, Tao Wang, Yu Emma Wang, Kellie Webster, Marie Pellat, Kevin Robinson, Kathy Meier-Hellstern, Toju Duke, Lucas Dixon, Kun Zhang, Quoc V Le, Yonghui Wu, Zhifeng Chen, and Claire Cui. 2021. Glam: Efficient scaling of language models with mixture-of-experts.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. pages 2368–2378.

Vid Kocijan, Ernest Davis, Thomas Lukasiewicz, Gary Marcus, and Leora Morgenstern. 2022. The defeat of the winograd schema challenge.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Shane Legg and Marcus Hutter. 2007. A collection of definitions of intelligence. *Proceedings of the 2007 conference on Advances in Artificial General Intelligence: Concepts, Architectures and Algorithms*.

Hector J. Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. pages 552–561. 13th International Conference on the Principles of Knowledge Representation and Reasoning, KR 2012 ; Conference date: 10-06-2012 Through 14-06-2012.

David Noever, Matthew Ciolino, and Josh Kalin. 2020. The chess transformer: Mastering play using generative language models. *CoRR*, abs/2008.04057.

Ovid. 43 B.C.-17 A.D. or 18 A.D. *Metamorpheses*.

Christine Payne. 2019. Musenet.

Stefan Riezler and Michael Hagmann. 2021. Validity, reliability, and significance: Empirical methods for NLP and data science. *Synthesis Lectures on Human Language Technologies*, 14(6):1–165.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. WinoGrande: An adversarial winograd schema challenge at scale. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8732–8740.

Uri Shaham, Elad Segal, Maor Ivgi, Avia Efrat, Ori Yoran, Adi Haviv, Ankit Gupta, Wenhan Xiong, Mor Geva, Jonathan Berant, and Omer Levy. 2022. SCROLLS: standardized comparison over long language sequences. *CoRR*, abs/2201.03533.

David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, Timothy P. Lillicrap, Karen Simonyan, and Demis Hassabis. 2017. Mastering chess and shogi by self-play with a general reinforcement learning algorithm. *CoRR*, abs/1712.01815.

A. M. Turing. 1950. I.—COMPUTING MACHINERY AND INTELLIGENCE. *Mind*, LIX(236):433–460.

Joachim Wagner, Jennifer Foster, and Josef van Genabith. 2013. Judging grammaticality: Experiments in sentence classification. *CALICO Journal*, 26(3):474–490.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. Blimp: The benchmark of linguistic minimal pairs for english.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. Neural network acceptability judgments.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.