TEJAS Journal of Technologies and Humanitarian Science
ISSN-2583-5599
Vol.03, I.03 (2024)
https://www.tejasjournals.com/
https://doi.org/10.5281/ zenodo.13982962

From Narrow to General: Rethinking Benchmarking for Artificial General Intelligence

Ansh Tiwari [1], Ashmit Dubey[2], Mahesh Kumar Tiwari[3], Rinku Raheja[4]

[1]Student Scholar, Computer Science, National P.G. College, Lucknow
[2]Student Scholar, Computer Science, National P.G. College, Lucknow
[3]Assistant Professor, Computer Science, National P.G. College, Lucknow
[4]Assistant Professor, Computer Science, National P.G. College, Lucknow

[1]anshtiwari3891@gmail.com, [2]dubeyashmit2606@gmail.com, [3]maheshyogi26@gmail.com, [4]rr_141085@yahoo.co.in

KEYWORDS

Artificial General Intelligence, Cognitive Skills, Turing test, Evaluation Approaches

ABSTRACT

Artificial General Intelligence (AGI) represents the ultimate goal of AI research: in order to develop systems that possess cognitive characteristics similar to those of the human brain and can effectively solve problems in different fields without the need of being trained for a specific type of work. But there are numerous issues that attend the pursuit of AGI, one of which is that benchmark that are adequate comprehensive measures to track the progress towards this goal are absent. The current kinds of evaluation, which have received most of their evolution within the establishment of narrow AI, do not perform adequately when applied to AGI. The focus of this research paper is a historical perspective of the significance of AI assessment with focus on the modern techniques which aimed at proving the AI's ability to think more widely. It discusses some of the inherent flaws present in existing benchmarks like biased sampling, overfitting and the so called "AI effect" in which solved tasks no longer measure intelligence. Looking at these challenges, this paper seeks to present the criteria towards formation of better benchmarks for the complexity and the range of capabilities needed in AGI.

1. Introduction

The objective of artificial intelligence of study is to develop intelligent systems with human like capabilities in terms of problem solving and cognition that can solve different problems in all fields without being trained specifically for a certain task. This is called artificial general intelligence, and often referred to simply as AGI. However, there are various impediments that exist in the path towards AGI and one of which is that there are very few good, standardized and comprehensive metrics for gauging how progress is being made in this regard. The current evaluation post methods are restricted in their use and scope and are designed for the current narrow AI and hence have a lot of limitations when used on AGI's. The ascending of the Turing crisis and developing new techniques for AI assessment, focusing on the use modern techniques which allows to evaluate other cognitive abilities. It analyses the integrated limitations of the current standards with relation to their biases, susceptible. This paper aims at exploring the inherent weaknesses of the current methods of benchmarking to show how benchmarks

Corresponding Author: Ansh Tiwari, National Post Graduate College, Lucknow, India
Email: anshtiwari3891@gmail.com

TEJAS Journal of Technologies and Humanitarian Science
ISSN-2583-5599
Vol.03, I.03 (2024)
https://www.tejasjournals.com/
https://doi.org/10.5281/ zenodo.13982962

can be biases; susceptible It is referred to as artificial general intelligence or AGI as it can solve any of the issues in the world. However, there are many challenges towards the development of AGI and one of them is lack of good, standard measures that can be used to measure progress towards this end. Most of the current-evaluation techniques are relatively narrow focused for modifying AI applications and, envisaged to fail noticeably when embedded into the AGI paradigm. This paper discusses the advancement in the AI from the early assessment test like the Turing test, to the present current techniques that were devised in a bid to assess various other cognitive abilities. It looks at some of the inherent flaws that have been recorded on the existing benchmarks such as the susceptibility to bias.

2. Historical Benchmarks of AI

It is possible to trace the beginning of the AI evaluation with regard to ideas of modern scientific thought as early as the Turing Test introduced by Alan Turing in 1950. This test was undertaken with the view of trying to determine whether a machine could replicate a human's behavior or not. The approach was systematically using a computer to connect a human evaluator with a machine and a human where the machine passed the test if the evaluator could not discern the difference between the two. Somewhere it was done by Turing in the form of The Turing Test which served as a significant landmark of providing the society a clear criterion of what constitutes Artificial Intelligence. Since Turing test, AI assessment has not remained stagnant but has evolved a lot due to the dynamic nature of AI systems which evolved from simple problem solving defined and tested approaches to more holistic ways of evaluating such systems. A: The initial attempts were based upon the works done for quite limited purposes, which included the assessment of how well a system performed in certain tasks, but as the systems evolved, conventional assessment methods such as black-box or behavioral assessments offered insufficient measures. Three main evaluation categories emerged: human discrimination and in working out problem benchmarks and peer confrontation.

However, each had its drawbacks: discrimination of human was not standardized, problem benchmarks may lead to overfitting and peer confrontation fails at providing an overall analysis of AI capabilities. These limitations were realized leading to the transition into ability-based assessments that focused on cognitive as aspects of AI systems not tasks. This transition involved the modification of the cognitive tests employed when studying human and animals and growing new methodologies from the concept of algorithmic information theory. Universal psychometrics is another idea that arose with the view to offer a single model of intelligence testing for people, animals and computers. Nonetheless, some difficulties remain: Inadequate statistical evaluation protocols, the SCHIB finds demand for better sampling methods for benchmarks, and more diverse problems. The performances of AI have also been evaluated through change from the focus on static point-based problem solving to the dynamic cognitive abilities of an AI system with continuous advancements being made on the comprehensive and reliable measure of AI. [7]

3. Limitations of Current metrics in measuring AGI progress

3.1 Problems with current benchmarks

3.1.1 Inherent limitations of narrow AI benchmarks

The demand for intelligence in AI originates from the existing AI systems, that are great at completing exact tasks, but poor when it comes to broad learning, problem solving in new scenarios or even considering data outside of the

TEJAS Journal of Technologies and Humanitarian Science
ISSN-2583-5599
Vol.03, I.03 (2024)
https://www.tejasjournals.com/
https://doi.org/10.5281/ zenodo.13982962

training sample. Currently, there are no clear quantitative definitions and clear parameters for measure the advancement toward the creation of artificial general intelligence.

Since growth in intelligence is typically measured incrementally, without an actionable measure of intelligence, AI development is problematic and it becomes challenging to define genuine advances. A common concern of today's tests such as the Turing Test is considered having flaws with its basis for the evaluation on the basis of participants' subjective perception. It is impossible to conduct objective metrics to measure cognition; the present methodologies for standardization are inadequate, which contribute to the emergence of biases that remain within AI research: tasks where AI can will surpass its human counterpart (e. g., board games).

A good proxy of intelligence is important in classifying useful Figure 1: intelligence as a template for the shape and direction of AI increase meaningful novelty, informing AI direction, and confirming that the advancement is in line with the goal of general intelligence, as conceived of at the outset of the field. In an effort to fill this void, the current document lays down a working definition of artificial intelligence and a rubric from which a categorical measure may be taken to gauge AI maturity; both are based on the principles of developmental cognitive psychology. [3]

The issues of reliability of the public benchmarks in large language models are emerged because of the possible contamination in pre-training or fine-tuning data sets. Previous techniques in decontamination of the data sources including string matching and the n-gram overlapping have proved to be inadequate. In other words, such measures are trivially evadable and can result in overfitting, and when the test data is translated or paraphrased, they proceed to score nicely on benchmark data. This has been evidenced in benchmarks such as MMLU, GSK8k, and HumanEval such as Google's diagnostic tests. More recent but closely related threat has been proposed a stronger LLM-based decontamination method to contain this risk. This method has exposed prior unseen extent of testing overlap in well-known pre-training and fine-tuning data sets, including RedPajama-Data-1T and StarCoder-Data that contains 8-18% of the HumanEval benchmark. It has also been discovered that random datasets created by such models as GPT-3 are also contaminated. 5 and GPT-4, this creates a possibility of accidental interaction between the two models and hence a chance of mixing. It is suggested that other, more forceful decontamination techniques should be practiced in the community, and that new, single-use models should be constructed to adequately assess the models. Overfitting is the phenomenon emanating from the somehow related problem of contamination, which happens when some information from the test set spills over to the training set, thus making the evaluation of the performance of the model artificially high. The following are the common approaches to detect contamination: The first one is n-gram overlap; The second is embedding similarity search; The third is decoding matching; The last one is influence approach. N-gram overlap utilizes string matching but has low accuracy as compared to exact match. Similarity search with embeddings is much different from word embeddings as it involves extracting new features of a similar example by testing with pre-trained model embeddings; it is difficult to set an appropriate limit to the similarity measures. Is Contamination This approach of decoding matching can be used when there are no training data available but the model is, Though not conclusive proof of contamination. Influence functions are based on the idea of defining an influence factor for each training sample, with the purpose of detect contaminated samples, however, this approach is very costly in terms of computational time. The issue can be made more complicated by such samples as rephrased ones, which have the same meaning as an original text but are extremely

TEJAS Journal of Technologies and Humanitarian Science
ISSN-2583-5599
Vol.03, I.03 (2024)
https://www.tejasjournals.com/
https://doi.org/10.5281/ zenodo.13982962

difficult to distinguish. These samples can include recasting of test samples by translating them into another language for example.

Samples used to train Models may be rephrased very easily to deliver highly accurate Benchmarks and they often overfit. To overcome such a problem, a novel decontamination technique using LLM has been designed recently. This method employs embedding similarity search to propose the k nearest samples to the test sample and then it asks a powerful LLM to determine if any of the k nearest neighbors are too close to the test case. The reported results demonstrated that this approach is much better than previously used approaches. The study shows that models trained on rephrased examples give near perfect results on test such as MMLU, HumanEval and GSM-8k therefore, rephrased samples should be deemed as contamination. The presented methods of detecting such samples are not able to recognize rephrased ones, whereas, the suggested LLM decontaminator does. This decontaminator applied on some standard training sets has shown rather large shares of rephrased samples, which proves a rather high risk of contamination. It is especially important to note that the subject of accidental contamination becomes critical when training models on synthetic data created by LLMs. There are suggestions for applying enhanced approaches toward sanitation, and the application of new, single-use tests is proposed for the assessment of LLM effectiveness[3].

3.1.2 Benchmarks capable of measuring AGI differ from current in focus narrow task-specific AI evaluation

From the turn, the emphasis on the assessment of closely defined AI capability has in the past been occasioned by the practical requirement of establishing performance at a given task. Such measures include those used in image recognition or board games, which state as to how effectively an AI system copes with a particular task. The effectiveness of these benchmarks can be viewed using reference to such a competition as the ILSVRC one – the image recognition challenge, or the DARPA Grand Challenge for autonomous driving. These benchmarks are replicable, equitable, extensible and reversible, thus can be considered as potent means to achieve further advancement in artificial intelligence. Nevertheless, they often result in creating the systems that are efficient in terms of accomplishing certain tasks although they do not evidence other types of intelligence as well as flexibility. Two, the ability to measure these broader forms of AI entails other paradigms of standards that make assessment of abilities instead of skills possible. In contrast to the specific benchmarks these broader benchmarks have to measure an AI system's ability to generalize, adapt as well as its resilience in a variety of tasks. This approach is very similar to psychometrics which relies on general tests to estimate the level of intelligence of people. In other words, for AI, it tends to set standards that measure the system's capability of performing new tasks it has not encountered before or new environments on its own. Some of such initiatives are Arcade Learning Environment which is used for testing reinforcement learning agents and Animal-AI Olympics which focuses on testing learning and planning instead of skills focused on particular tasks.

That is why there is a need for different addressing in a more general AI assessment in holistically, as the narrow measures simply do not suffice in reflecting the real intelligence of artificial beings. It's possible to state that preselected benchmarks, rather specific for certain purposes, may be outplayed by narrowing the system, which will perform well in a strict setting but horribly in the actual one. Of the two camps of evaluation methodologies, the narrow ones are designed to assess the AI's capability in specific contexts, while the broader aims at assessing the generalization capability and flexibility of an AI, angles that are very crucial in attaining human like intelligence.

TEJAS Journal of Technologies and Humanitarian Science
ISSN-2583-5599
Vol.03, I.03 (2024)
https://www.tejasjournals.com/
https://doi.org/10.5281/ zenodo.13982962

This shift is important to build systems capable of addressing real-world tasks that are considerably more complex than typical training tasks, that require general intelligence rather than Stylesheet Hairstyling & Salons Текст switch from focusing solely on specific tasks and providing sequences of actions that lead to the completion of only them.

3.1.3 Benchmark Saturation and Diminishing Returns

The researchers at Stanford are asking some interesting questions about how we benchmark AI tools such as ChatGPT. Although these systems can achieve high results in terms of specific tasks, the fellows from Stanford mention that they sometimes fail in handling real-life scenarios – providing wrong or at least, peculiar answers.
The same has been evidenced by the most recent AI Index report, where many of the benchmarks appear to have plateaued in recent time, with even the leading systems only marginally ahead year over year. For instance, the best image labeling AI merely increased by 0.1% between 2021 to 2022, indicating that the rate of development is slowing.

The Stanford researchers believe that in order to build better AI systems we urgently require new evaluations that demonstrate that systems are not only accurate, but also non-prejudiced, secure and helpful to society. For instance, they mention the HELM benchmark developed at Stanford which considers broader characteristics such as fairness and reliability.

It is a critical discussion of the proper relationship between innovation and people. Even when a system is capable of producing high scores on tests, this does not necessarily mean that it will be useful or not damaging when implemented. Better and more general references allow further advancement to occur in the right direction. However, this might take some effort maybe a multi-partnership involving the business world, the academia, and the society to achieve this. [8] [18].

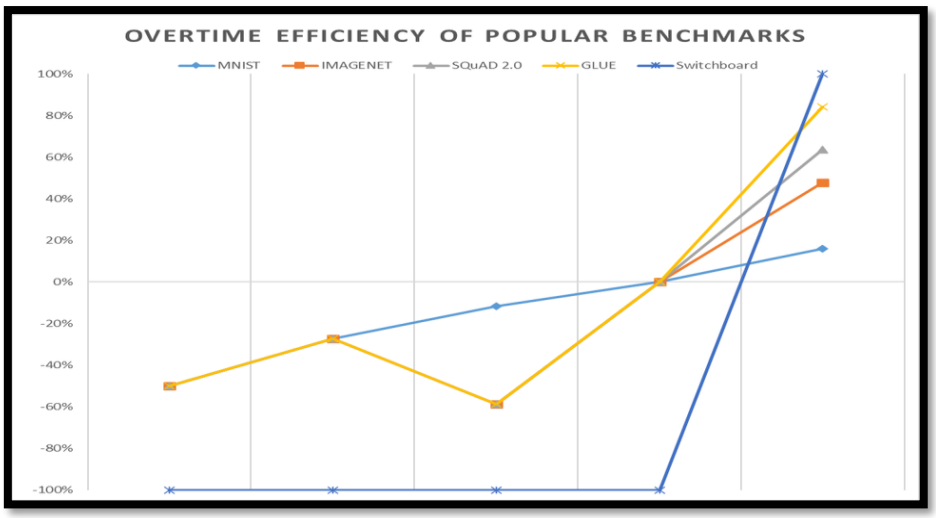

Figure 1, Benchmark saturation over time for popular benchmarks, normalized with initial performance at minus one and human performance at zero.

3.1.4 The Bias factor in data

A considered reflection suggests that it is important to note that existing AI benchmarks are largely skewed by their designers. Like no individual can be completely bias free, no data set is free of bias as well. Some benchmarks

TEJAS Journal of Technologies and Humanitarian Science
ISSN-2583-5599
Vol.03, I.03 (2024)
https://www.tejasjournals.com/
https://doi.org/10.5281/ zenodo.13982962

such as ImageNet and GLUE are considered as bars for assessing the general AI progress, but they reflect only certain aspects of intelligence.

For example, ImageNet categories originate from WordNet and therefore it possesses the Anglo-American culture of categorization that connects to what is deemed worthy of naming. This can recklessly omit nonwestern perspectives. About 46% of images in ImageNet originate from the US, while only 1% show China and India despite being two of the most populous countries. Therefore, its visual world is culture-specific and based around certain cultures. When one searches ImageNet using Hindi search terms, it produces entirely different images, as research has demonstrated.

While GLUE and SuperGLUE are intended for the American English, which already makes the range of meanings they hold rather limited by default. Thus, it is misleading to present these benchmarks as genuine intelligence tests. This could contribute to building artificial intelligence that performs well on standardized assessments but poorly in practice.

Some of the ethical issues that arise when researchers make benchmark scores their main target include the following. The solution is not to look for some kind of universally ‘neutral’ dataset but to properly frame benchmarks and augment them with other assessments. We have to limit datasets to precisely define what they contain and validate AI more generally by creating datasets systematically, analyzing outputs, and using ablation analyses. This will result in having more accurate estimates of the capabilities and limitations. The main idea is that there is no key that can encode all the intelligence. However, careful and considerate design and assessment can create AI that is helpful and fulfilling for people. [11]

3.2 Challenge of defining “General Intelligence”

Defining intelligence has been a complex and debated issue for over a century, involving both psychology and artificial intelligence (AI). The core problem lies in the **divergent visions of intelligence**. One view sees intelligence as a collection of **task-specific skills**, where an intelligent agent excels in various tasks, such as playing games or solving puzzles. The other view considers intelligence as a **general learning ability**, emphasizing the capacity to acquire new skills and adapt to new situations.

A significant challenge is that **measuring skill at specific tasks** does not fully capture intelligence. Skills can be heavily influenced by prior knowledge and experience, making it possible to “buy” high levels of skill through extensive training, which masks the system’s true generalization power. This leads to the **need for a new definition** that accounts for the efficiency of skill acquisition, considering factors like scope, generalization difficulty, priors, and experience.

The contemporary AI community often benchmarks intelligence by comparing AI and human performance on specific tasks, but this approach falls short. It does not address the broader abilities required for **humanlike general intelligence**, such as adaptability and flexibility in unfamiliar situations. Therefore, a more comprehensive and actionable definition of intelligence is needed to guide the development of AI systems that can truly mimic human cognitive abilities. [3]

TEJAS Journal of Technologies and Humanitarian Science
ISSN-2583-5599
Vol.03, I.03 (2024)
https://www.tejasjournals.com/
https://doi.org/10.5281/ zenodo.13982962

4. Proposals for criteria of developing robust benchmarks for evaluating AGI

4.1 Criteria for Developing Effective Benchmarks for AGI

The researchers have discussed the evaluation of AI systems, emphasizing the shift from task-oriented to ability-oriented approaches. It critiques traditional methods like black-box (behavioral) evaluation, human discrimination, problem benchmarks, and peer confrontation, highlighting their limitations. The paper proposes a more systematic and robust evaluation, focusing on cognitive abilities rather than specific tasks. We also explore possible similarities between human intelligence evaluation and AI evaluations. The goal is to improve AI system evaluation by integrating diverse evaluation paradigms and addressing the challenges posed by the increasing complexity and unpredictability of AI systems. [4]

The advancement of AI benchmarks is moving towards enhancing both task-specific and cognitive ability evaluations. While task-oriented assessments, which gauge AI performance on particular activities like image recognition or natural language processing, are commonly employed, they have their shortcomings. These benchmarks often fail to evaluate an AI system's capacity to generalize, as they concentrate on narrow task sets that may not accurately represent real-world situations. To tackle this issue, researchers are investigating more varied problem sets, problem-generating tools, and standardized, open-source benchmarks to better assess AI's adaptability and effectiveness across diverse tasks.

In contrast, ability-oriented evaluation aims to measure the cognitive capabilities of AI systems, drawing inspiration from intelligence tests designed for humans and animals. This method seeks to assess general intelligence, including learning, reasoning, and problem-solving skills, providing a more comprehensive view of AI capabilities. Challenges in this field include the need for novel evaluation metrics, such as modifying psychometric tests like IQ assessments for AI use or utilizing algorithmic information theory to measure problem complexity and difficulty, offering more nuanced insights into AI abilities.

A key area of research involves merging task-oriented and ability-oriented evaluations. This integrated approach aims to develop comprehensive benchmarks that assess both specific performance and general cognitive abilities of AI systems. This aligns with the notion of universal psychometrics, which seeks to create universal tests applicable to any intelligent system, establishing a unified framework for AI evaluation.

In conclusion, future AI benchmarks are likely to evolve towards more comprehensive, adaptable, and integrated methods, combining task-specific assessments with evaluations of cognitive abilities. This approach will provide a deeper understanding of AI capabilities, helping to identify areas for improvement and driving the development of more sophisticated and versatile AI systems, which is crucial for the ongoing progress of artificial intelligence. [6]

4.2 A Multimodal approach to AGI: Multifaced benchmarks

The evaluation of the human intelligence includes various methods, of which are intelligence quotient or IQ tests that estimate logical behavior, problem solving abilities, knowledge retention and capability to understand language. Nevertheless, these measures do not necessarily reflect the entire repertoire of skills and abilities inherent in human intelligence, including creativity, empathy, and practical problem-solving.

Changing paradigms stress the importance of the above aspects to be incorporated in intelligence tests, which is crucial for the advancement of AGI.A powerful argument in AGI research is that AGI cannot be fully instantiated

TEJAS Journal of Technologies and Humanitarian Science
ISSN-2583-5599
Vol.03, I.03 (2024)
https://www.tejasjournals.com/
https://doi.org/10.5281/ zenodo.13982962

using the current algorithmic paradigm of AI. AGI needs the ability to detect new affordances – which is a possibility or a threat that cannot be clearly defined and encapsulated into an equation. In contrast, human cognition is excellent in capturing such affordances via contextual reasoning, empathy and goal setting which are based on the functional structure and self-regulation of biological beings. So, AGI benchmarks should include components like goal setting and goal modification beside task accomplishment and time to solution; control of uncertainty; and new forms of knowledge representation. From this point of view, it becomes possible to identify the weaknesses of addressing the AI problem in terms of general intelligence based on the algorithmic approach, as the tools based on this approach do not have an option for self-organization and adaptation inherent in biological entities. As mentioned earlier, the AI systems excel specifically in accomplishing certain tasks, but they lack the organismic agency intrinsic to human cognition: objectives, actions, and affordances. Analyzing the nature of smartness and its quantification has implications for AGI creation. With the help of non-linear measures of performance that are free from the limitations of the standard metrics, it is possible to develop more adequate evaluations for AGI systems that take into consideration their autonomy, flexibility, and the rich nature of cognition. It raises questions about whether it is possible and appropriate to apply existing metrics to AGI again, while recognizing the particularities of both biological and algorithmic approaches. [14] [17]

4.3 Lessons from Human Cognition Studies

The evaluation of the human intelligence includes various methods, of which are intelligence quotient or IQ tests that estimate logical behavior, problem solving abilities, knowledge retention and capability to understand language. Nevertheless, these measures do not necessarily reflect the entire repertoire of skills and abilities inherent in human intelligence, including creativity, empathy, and practical problem-solving. Changing paradigms stress the importance of the above aspects to be incorporated in intelligence tests, which is crucial for the advancement of AGI.

A powerful argument in AGI research is that AGI cannot be fully instantiated using the current algorithmic paradigm of AI. AGI needs the ability to detect new affordances – which is a possibility or a threat that cannot be clearly defined and encapsulated into an equation. In contrast, human cognition is excellent in capturing such affordances via contextual reasoning, empathy and goal setting which are based on the functional structure and self-regulation of biological beings. So, AGI benchmarks should include components like goal setting and goal modification beside task accomplishment and time to solution; control of uncertainty; and new forms of knowledge representation.

From this point of view, it becomes possible to identify the weaknesses of addressing the AI problem in terms of general intelligence based on the algorithmic approach, as the tools based on this approach do not have an option for self-organization and adaptation inherent in biological entities. As mentioned earlier, the AI systems excel specifically in accomplishing certain tasks, but they lack the organismic agency intrinsic to human cognition: objectives, actions, and affordances.

Analyzing the nature of smartness and its quantification has implications for AGI creation. With the help of non-linear measures of performance that are free from the limitations of the standard metrics, it is possible to develop more adequate evaluations for AGI systems that take into consideration their autonomy, flexibility, and the rich nature of cognition. It raises questions about whether it is possible and appropriate to apply existing metrics to AGI again, while recognizing the particularities of both biological and algorithmic approaches. [10][16]

TEJAS Journal of Technologies and Humanitarian Science
ISSN-2583-5599
Vol.03, I.03 (2024)
https://www.tejasjournals.com/
https://doi.org/10.5281/ zenodo.13982962

5 Case Studies and Practical Implementations

5.1 Experiments and Benchmarks with gpt-4

The research article “Sparks of Artificial General Intelligence: Early experiments with GPT-4” by Microsoft Research explores the capabilities of GPT-4, a large language model developed by OpenAI, through various experiments across multiple domains. The researchers conducted a series of tests to evaluate GPT-4’s performance in language, coding, mathematics, vision, and interaction with humans and tools. These experiments aimed to assess the model’s general intelligence and its potential as an early version of an artificial general intelligence (AGI) system.

One of the key experiments involved testing GPT-4’s language abilities. The model was asked to generate a number of unrealistic outputs such as a poem composed of infinitudes of prime numbers. GPT-4 successfully completed these tasks, demonstrating its ability to combine mathematical reasoning, poetic expression, and coding skills. The researchers also compared GPT-4’s performance to that of ChatGPT, a previous state-of-the-art language model, and found that GPT-4

produced superior outputs. This experiment highlighted GPT-4’s mastery of natural language and its ability to manipulate complex concepts, which are core aspects of reasoning.

In the domain of coding, GPT-4 was tested on LeetCode’s Interview Assessment platform, which simulates coding interviews for software engineer positions at major tech companies. GPT-4 solved all questions from three rounds of interviews, achieving high scores and outperforming the majority of human candidates. This experiment demonstrated GPT-4’s coding abilities and its potential to be hired as a software engineer. Additionally, GPT-4’s performance on the US Medical Licensing Exam and the Multistate Bar Exam showed its competence in medicine and law, respectively. These results indicate that GPT-4 possesses human-level abilities in multiple expert domains, further supporting its potential as an AGI system.

The researchers also explored GPT-4’s mathematical abilities by engaging it in a mathematical conversation and testing its performance on mathematical problem datasets. GPT-4 demonstrated a deep understanding of mathematical concepts and produced accurate solutions to complex problems. The model’s ability to generalize and adapt to different mathematical tasks suggests that it has a flexible and general understanding of the domain.

In the realm of vision, GPT-4 was tasked with generating images based on detailed instructions, similar to the capabilities of models like DALL-E. The model successfully created images that went beyond simple memorization, demonstrating its ability to generate novel visual content. This experiment highlighted GPT4’s integrative ability to combine language and vision, which is crucial for developing more comprehensive AGI systems.

The researchers also tested GPT-4’s interaction with tools and its ability to plan and learn from experience. GPT-4 was able to use multiple tools to solve complex tasks, simulate game environments, and interact with humans in a meaningful way. These experiments demonstrated GPT-4’s potential to be integrated into real-world applications and its ability to adapt to different scenarios. The model’s performance in these tasks suggests that it has the capacity to learn and improve over time, which is a key characteristic of AGI.

The study showed clearly how traditional benchmarks fail to accurately measure complete capabilities of advanced AI models. Traditional benchmarks, such as standard datasets and evaluation metrics, were used to assess the model’s performance in various domains. However, the researchers also proposed a new approach to studying

TEJAS Journal of Technologies and Humanitarian Science
ISSN-2583-5599
Vol.03, I.03 (2024)
https://www.tejasjournals.com/
https://doi.org/10.5281/ zenodo.13982962

GPT-4, which involved generating novel and difficult tasks that go beyond traditional benchmarks. This approach aimed to demonstrate GPT-4's generality and its ability to perform tasks that do not admit a single solution. The researchers acknowledged that this approach is somewhat subjective and informal, but they believe it is a necessary first step to appreciate the remarkable capabilities and challenges of GPT-4. [5]

Table-1: Test scores of GPT-4(ChatGPT) compared to other model's base on Summarization

TASK	RELEVANCE	FLUENCY	COHERNCE	CONSISTENCY
Gold reference	4.00(4.00)	4.00(4.00)	4.00(4.00)	4.00(3.00)
T0pp	3.00(4.00)	4.00(3.00)	4.00(4.00)	4.00(3.00)
GPT-3	1.67(1.50)	3.00(2.00)	2.83(4.00)	1.17(2.00)
ChatGPT	**1.33(1.00)**	**1.33(1.50)**	**1.00(1.00)**	**1.00(1.00)**

Table-2: Test scores of GPT-4(ChatGPT) compared to other model's base on Simplification

TASK	RELEVANCE	FLUENCY	SIMPLICITY
Gold reference	4.00(4.00)	4.00(4.15)	3.33(2.00)
Flan-T5	3.00(2.50)	4.00(1.50)	3.33(2.00)
InstructGPT	2.00 (3.00)	3.00 (2.00)	2.33 (3.00)
ChatGPT	**3.00 (1.00)**	**4.00 (2.00)**	**4.00 (4.00)**

Table-3: Test scores of GPT-4(ChatGPT) compared to other model's base on GEC

Task	SEMANTICS	GRAMMATICALITY	OVER-CORRECTION
Gold reference	3.33 (2.00)	4.00 (3.50)	2.50 (1.00)
OPT-IML	1.00 (4.00)	4.00 (2.00)	1.00 (2.00)
GPT-3	1.00 (3.00)	3.00 (2.50)	2.00 (3.00)
ChatGPT	**3.67 (3.00)**	**1.00 (2.50)**	**3.50 (4.00)**

5.2 Stanford AI hallucination Benchmark and What it means for the future of AGI Benchmarking

Stanford new benchmark is a significant progress in rating the complex AI systems, particularly in terms of its testing in text, image and data. It is as if, like the lawyers trying to manage documents, photos, and databases, this approach is more in tune with the expectations of tasks attributed to 'thinking' machines. One of the objectives is to expand the analysis of the system's ability to 'hallucinate' that is produce plausible lies. This matters in law, where facts reign supreme. The results indicate positive changes but also significant gaps, particularly in cases

TEJAS Journal of Technologies and Humanitarian Science
ISSN-2583-5599
Vol.03, I.03 (2024)
https://www.tejasjournals.com/
https://doi.org/10.5281/ zenodo.13982962

involving multifaceted issues. Thus, although in some fields the AI seems to be a master, we cannot rely on its judgment while performing actual legal work. By charting different types of errors, Stanford's test provides information for future training. And by testing the AI 'with and without hints,' it also exposes the flaws of the machine even more. This underlines the importance of continuing to expand the type of data used in development rather than fixing that in place. On a more general level, the analysis points to the danger of bias with performance differing from one court to another depending on what data was available. Similar to how humans' limited or rich life experience shapes their worldview, patterns in the AI's experience can subtly bias its "reasoning." It is why constant training on various cases improves these systems' resilience, equity, and utility. The Stanford approach is constructive in a field where facts and ethics dominate the discourse. [13]

7 Conclusion

Consequently, Artificial General Intelligence (AGI) requires moving from traditional, narrow AI benchmarks to more capable approaches to evaluation. The capabilities and limitations of AGI systems, however, are rarely captured by these benchmarks which have been developed for narrow AI. Instead, they typically emphasize task specific performance that often does not naturally translate into a more human like ability to be adaptable, generalizable, and problem-solve. This research highlights the need for benchmark cognitive abilities which incorporate learning, reasoning and adaptability. The paper also discusses the problem of contamination, biases, and overfitting in existing benchmarking practices, and suggests using more robust decontamination techniques. Moving ahead towards ability-oriented benchmarks inspired from human intelligence assessment gives a holistic view of AI capabilities and also ensures that in future AGI systems they can perform specific tasks with high proficiency as well as generalize across different domain and adapt to different environments. The successful implementation of true AGI will necessitate interdisciplinarity, a knowledge of specific skills and a broader measure of cognitive ability, and the development of integrated evaluation frameworks to capture both. By rethinking and refining benchmarks we can simultaneously push towards the creation of more sophisticated, versatile, and ethically aligned AI systems, and closer than ever to the creation of truly intelligent machines.

References

[1] Akpan, M. "Have We Reached AGI? Comparing ChatGPT, Claude, and Gemini to Human Literacy and Education Benchmarks."
[2] Schaul, T., Togelius, J., & Schmidhuber, J. "Measuring Intelligence through Games."
[3] Chollet, F. "On the Measure of Intelligence."
[4] Hernández-Orallo, J. "AI Evaluation: Past, Present, and Future."
[5] Bubek, S. "Sparks of Artificial General Intelligence: Early Experiments with GPT-4."
[6] Mueller, M. "The Myth of AGI: How the Illusion of Artificial General Intelligence Distorts and Distracts Digital Governance."
[7] Yampolskiy, R. V. "Turing Test as a Defining Feature of AI-Completeness."

TEJAS Journal of Technologies and Humanitarian Science
ISSN-2583-5599
Vol.03, I.03 (2024)
https://www.tejasjournals.com/
https://doi.org/10.5281/ zenodo.13982962

[8] Stanford HAI. "AI Benchmarks Hit Saturation." Available at: https://hai.stanford.edu/news/aibenchmarks-hit-saturation

[9] Raji, I. D. "AI and the Everything in the Whole Wide World Benchmark."

[10] Yang, S., Chiang, W. L., Zheng, L., Gonzalez, J. E., & Stoica, I. "Rethinking Benchmark and Contamination for Language Models with Rephrased Samples."

[11] Sottana, A., Liang, B., Zou, K., Yuan, Z. "Evaluation Metrics in the Era of GPT-4: Reliably Evaluating Large Language Models on Sequence-to-Sequence Tasks."

[12] Dahl, M., Magesh, V., Suzgun, M., & Ho, D. E. "Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models."

[13] Fei, N., Lu, Z., Gao, Y., Yang, G., Huo, Y., Wen, J., Xin, R., Gao, T., Xiang, H., Sun, H., & Wen, J. R. "Towards Artificial General Intelligence via a Multimodal Foundation Model."

[14] Pieter Abbeel and Andrew Y. Ng, 'Apprenticeship learning via inverse reinforcement learning', in Proceedings of the twenty-first international conference on Machine learning, p. 1. ACM, (2004).

[15] Tarek Besold, Jose Hern ´ andez-Orallo, and Ute Schmid, 'Can Machine Intelligence be Measured in the Same Way as Human intelligence?', KI- Kunstliche Intelligenz ¨, 1–7, (April 2015).

[16] Richard A. Caruana, Multitask Learning Thesis, PhD, Carnegie Mellon University, Pittsburgh, PA, September 1997.

[17] Jordi Bieger, Kristinn R. Thorisson, and Deon Garrett, 'Raising AI: ´Tutoring Matters', in Proceedings of AGI-14, pp. 1–10, Quebec City, Canada, (2014). Springer.