RESEARCH DIRECTIONS

To What extent Can Artificial Intelligence Apply Physics to Solve Global Problems?

Dylan Davidson and Samantha L Pugh*

School of Physics and Astronomy, University of Leeds, Woodhouse Lane, Leeds LS2 9JT

*Corresponding Author: S.L.Pugh@leeds.ac.uk

Keywords: *Artificial Intelligence; Generative AI; Global Challenges; Physics; Qualitative Methods*

Abstract

Generative Artificial Intelligence (GenAI) is an emerging technology that creates relevant text, images and other content from prompts. Large Language models (LLMs) are the most widely used of these GenAI forms. This technology already has applications in business and education.

This paper tests GenAI's ability to apply physics to global problems and arrive at viable solutions. When an idea is created by a human, it is merely a culmination of that person's experiences and prior knowledge, ordered into a new concept. This research proposes that it should be possible to replicate the process by a machine learning algorithm and, due to its vast database, a far more informed and coherent idea should be the result. This research tested how well AI could tackle some global challenges and compared the results to how well these same challenges could be addressed by physicists.

The data collection process was to have a dynamic conversation with each of the participants and work with them to create a number of ideas and solutions that apply physics to a selection of global issues. This process was repeated with both Bing AI and ChatGPT-4, where they were prompted to return ideas to the same issues. Each of the ideas were then coded to a marking scheme adapted from the OECD DAC criteria for development evaluation.

While Bing AI did not prove itself to be capable of unique idea creation, ChatGPT-4 returned valuable data. ChatGPT-4 excelled at providing efficient, coherent and sustainable results whilst it performed significantly worse than humans in versatility and profitability.

The findings show that at the present time, AI cannot work as an idea generation tool on its own due to lacking in accuracy and versatility. It is best applied in tandem with humans where it can be used to generate a series of ideas to a problem which physicists refine the results.

Introduction

Artificial Intelligence (AI) is one of the fastest evolving technologies in our society (West and Allen, 2018). The mass distribution of accessible Large Language models (LLMs) in the form of AI chat bots have catapulted AI to the forefront of both scientific and public attention. Their capabilities of these span from creating unique stories from a short prompt to writing pages of accurate code in seconds. The ramifications of AI in scientific circles and elsewhere could prove to change the way we approach and solve problems. (Sarma, 2023)

The world is currently facing a range of global issues from the climate crisis to poverty and hunger in developing countries. Although physics alone may not be able to solve all global issues, certain

https://doi.org/10.29311/ndtns.vi19.4705

branches of physics could be utilised to mitigate some of these problems (Niemela, 2021). With the help of AI, in the form of large language models, these problems could be addressed in the most logical and economical ways. Since LLMs are created from an extremely large databases, they have more information available to them than any human could comprehend (Walsh, 2023). Using this expansive knowledge, it follows that these LLMs could make well-informed decisions about how these issues can be tackled. Humans, however, have natural intuition, incredibly complex brains that have evolved over many centuries for problem solving and the power of collaboration. This project assesses the ability of AI to apply physics to solve global challenges and where it may compliment or surpass existing human capabilities.

Generative AI (GenAI) is the blanket term for Machine learning algorithms that 'create'. They can process information and adjust, improve or synthesise a new unique result (Hughes, 2023). LLMs are specialised to create human like speech patterns. With their large dataset and with specific training, they become the AI Chatbots seen today. In their raw form, LLMs are a hyper intelligent autocompleter. They use a token system to predict the most likely following sequence of words (Kumar, 2023). The token system works by splitting sentences, words and sequences up into tokens. This allows the AI to handle every word without having to have each of them in its memory. Prefixes such as 'in' can be used in a range of words (Mittal, 2023). This also means that the AI can predict the meanings of words that it does not know the meaning of as it can break up the word into its constituent parts.

ChatGPT and BingAI are not LLMs; they are Chatbots based on LLMs, a process called ‘Reinforcement learning from human and AI feedback’ (Polverini and Grogorcic, 2023). A series of results to a prompt are given to humans and they are ranked on relevance and eloquence, among other factors. In its raw form, an LLM is not user friendly, and it could be difficult to get information from it. Due to the way that the LLM processes information, there are some simple techniques that allow the user to receive far more relevant and accurate answers from it (Sahoo et al, 2024). For the data collection, two specific methods were used:

Chain-of-thought prompting (Mittal, 2023) is a prompt engineering method that works to break up a problem for the AI. This is done by getting the AI to tackle a problem bit by bit. In the context of this study, this involved getting the AI to outline the issues related to the topic first, and then choose one to make an innovative, physics-based solution for. This prompt engineering technique leads to more accurate results.

Role play is essential if the user is looking for highly relevant results. Role play is the practice of assigning a role for the AI to answer the question (He et al, 2024). This helps the AI tune its temperature, or how creative the answer should be (Choi et al, 2024) and give it a better understanding of what data from its database to pick from. It also helps the AI know the level of complexity that it should answer with e.g. “From the perspective of a physics lecturer explain …” will give a far more complex answer than “Pretend that you are explaining … to a toddler.”

A Categorical Archive of ChatGPT Failures (Borji, 2023) lays out 11 elements that the current LLMs have trouble with. Certain factors are not relevant to this project but its challenges with mathematics, logic and reasoning are relevant. It struggles with kinematic questions or simple logic riddles. This is because its knowledge lies in a completely linguistic place (Mittelstadt et al, 2023) and has no basic sensory experiences like humans have (Blasi, 2023). These downfalls improved dramatically from the capabilities of GPT-3.5 to GPT-4 (Kelly, 2024) but the underlying lack of true understanding remains.

Methodology

The aim of the research process was to assess the idea creation process of AI chatbots. This was done using human responses to similar questions as a comparator.

The issues were chosen to be diverse and open topics that are potentially relevant to physics. The final three topics chosen were:

- Endangerment of animals
- Excessive food waste
- Climate change

To compare human and AI responses, a method to assess the quality was required. The OECD (Organisation for Economic Co-operation and Development) DAC criteria for development evaluation served as a template for the criteria that was used (OECDa, 2023). The OECD is an intergovernmental Organisation with a goal to set international guidelines for development. There are 32 Member countries that use the OECD to assist in creating change (OECDb, 2023). It uses six factors to judge the value of a given intervention. These are: Relevance, Coherence, Sustainability, Effectiveness, Efficiency and Impact, as illustrated in Figure 1.

Figure 1 OECD Evaluation Criteria (OECDa, 2023)

Whilst these criteria are useful, some factors that were considered important were missing so three more were added:

1. Scalability - The ability for a specific intervention to scale up and get significantly better once implemented, is an essential factor in ensuring that the changing demands of the population are taken care of.
2. Versatility - If an intervention is implemented and it can address multiple global issues, its value drastically increases. While this is not essential to every innovation, versatility can provide a huge economical upside as an investment in a solution can work towards addressing multiple issues.
3. Profitability - Lastly, profitability is invaluable in promoting change. A great deal of the funding into global issues comes from business. Companies will only invest in technologies that are deemed profitable as they simply cannot operate sustainably without making money.

https://doi.org/10.29311/ndtns.vi19.4705

With these extra criteria outlined, the final list that the proposed ideas were graded on is as follows:

1. Relevance – how urgent is the problem that this intervention solves?
2. Coherence – how clear is the method and how complete is the science to apply this intervention?
3. Sustainability – how well does this intervention promote our journey to a more sustainable planet?
4. Effectiveness – to what extent does this intervention solve the problem?
5. Efficiency – is this an efficient use of money/time/resources for what is solved by the intervention?
6. Impact reduction – does this avoid having a negative impact whilst actioning this intervention?
7. Scalability – does this intervention have the ability to rapidly improve due to investment of time/money?
8. Versatility – does this intervention have the ability to address other problems than the target issue?
9. Profitability – can this intervention generate profit?

Human Discussion Data Collection

The participants were two physics lecturers and two final year physics students. This was to achieve a range in perspectives while still having the participants maintaining a high degree of fluency in physics. The goal of the discussions was to get the participants to use their knowledge of physics to generate unique and creative ideas that address the problems outlined. The discussions were planned to be semi-structured interviews. The goal during the discussion was to have a dynamic conversation and avoid the participants simply feeding ideas. The interviews ranged from 40 to 90 minutes. This interview structure promoted the power of human conversation and going off on a tangent, which is an important way that humans have unique ideas. A range of ideas was collected from each participant.

Ethical approval from the University of Leeds MEEC 13-017 was obtained to complete this discussion stage of the research.

AI Questioning Data Collection

The goal of the AI discussion was to attempt to get the AI to generate unique physics-based solutions to the issues provided. The AI chatbots used were Bing AI and Chat GPT-4. It would be of limited value to test AI chatbots that are not using the most up to date algorithm in GPT-4. Simple prompt engineering techniques were used to help the AI to understand how to respond. This was primarily using the Role Play prompt engineering method to get the AI to respond, 'from the perspective of a physicist'. This has a significant impact on the quality of results. Similar questions without this parameter performed much worse. A similar effect occurred when the chain-of-thought prompt engineering technique was not applied. With the open topics, the AI needed some direction, but it was necessary that it arrived at any 'idea' it had on its own. The method employed to achieve this was to ask it to outline the problems with a certain topic then target each of the issues it outlines and request a unique solution to the problem applying physics and physical principles. The initial prompt given was:

'From the perspective of a physicist, outline the issues causing *ISSUE* and describe how you could address it.'

This typically would generate a list of bullet points of the branches of the topic with short and vague solutions. After this, one of the branches was chosen and the AI was asked:
'Ignoring cost and political tensions, please come up with a unique idea to improve *SPECIFIC PROBLEM* in order to aid in resolving *ISSUE*.'

https://doi.org/10.29311/ndtns.vi19.4705

To encourage creative ideas, the Chatbots were instructed to ignore cost and political tensions. This allowed the AI to have fewer constraints when creating ideas. This parameter was reintroduced occasionally to see how the AI could adapt the idea into something more realistic. These ideas were then judged by the marking criteria.

Data Collection
The results were tabulated by extracting one idea per topic for each of the AI chatbots and each of the humans. The ideas were assessed using the criteria on a scale of 1 to 10 for each of the factors. The averages of the human answers for each of the criteria were calculated and the difference between these and the ChatGPT-4 answers were plotted to show the factors that humans tend to consider more and where AI can perform better. This is essential information as it highlights the steps of idea creation that suit each of the sources of ideas.

Results

The information from the human discussions was collected by recording the meetings and taking bullet point notes after the conversations of the important aspects of each idea. These abridged ideas were then coded according to the marking scheme.

As an example of the process, during the first conversation with Student 1, the alternative uses of their drone swarm idea was a topic of discussion. The raw data of the notes for this section of the discussion was:

AI drone swarm for animal monitoring
Other uses

- Imaging ice caps – climate
- Mountain rescue, police chase – public services.
- Alternative to satellites (satellite limit doesn't apply to drones)
- Surveillance etc.

This led to the versatility score being a 10/10 as the drone technology can make a significant impact in other fields.

For the AI data collection, the prompts that have been outlined in the Method were entered into the AI Chatbots and the answers were copied into a notebook to be coded to the criteria. ChatGPT-4 gave bioengineered algae bloom fields as a solution in assisting the climate crisis. This idea focuses on increasing the carbon capture capabilities of phytoplankton and growing large fields of this improved algae. The AI stated:

"Genetically Engineered Phytoplankton: Develop genetically modified phytoplankton that have enhanced capabilities for photosynthesis and carbon sequestration. These super-phytoplankton would be designed to absorb CO_2 from the atmosphere more efficiently than natural variants and convert it into organic carbon, much of which would sink to the ocean floor when the organisms die, effectively sequestering carbon from the atmosphere for millennia."

The bioengineering to create mass amounts of 'super-phytoplankton' does not yet exist but CRISPR/Cas9 gene editing technology is rapidly improving (Redman et al, 2016) so this seems to be a viable option in the future. It is already possible to grow phytoplankton fields which have incredible carbon capture capabilities naturally (Irion et al, 2021). As the idea is not possible in its entirety currently, but can already be useful, the coherence score was given as a 6/10.

In the following data tables, where a column remains blank, no suitable idea was provided for that topic. For the human discussions, if the column is blank, the participant chose not to focus on that topic. For Bing AI, if the column is blank, it failed to produce a unique or useful response.

https://doi.org/10.29311/ndtns.vi19.4705

Criteria	Student 1	Student 2	Lecturer 1	Lecturer 2	Bing AI	ChatGPT-4
Idea	AI Drone network for monitoring of animals	Fertility detecting microchip	No idea provided	No idea provided	More Satellites	Dynamic Environmental Shield
Relevance	6	4	-	-	6	7
Coherence	4	8	-	-	7	9
Effectiveness	9	7	-	-	2	9
Efficiency	2	6	-	-	2	5
Impact reduction	4	5	-	-	6	5
Sustainability	5	6	-	-	5	8
Scalability	5	6	-	-	2	4
Versatility	10	2	-	-	6	3
Profitability	9	2	-	-	3	2
Total score	55	56	-	-	39	57

Table 1 Endangerment of Animals responses evaluation

Criteria	Student 1	Student 2	Lecturer 1	Lecturer 2	Bing AI	ChatGPT-4
Idea	Cellulose Packaging	Fast Defroster	AI Smart crop watering	More filling foods	No suitable idea provided	Bio-preserver system
Relevance	6	3	5	8	-	7
Coherence	7	8	7	4	-	8
Effectiveness	7	2	5	7	-	4
Efficiency	3	6	7	3	-	7
Impact reduction	8	8	6	2	-	8
Sustainability	9	4	6	8	-	7
Scalability	10	3	8	3	-	5
Versatility	8	2	4	2	-	3
Profitability	5	8	9	1	-	8
Total score	63	44	57	38	-	56

Table 2 Reducing Food Waste responses evaluation.

https://doi.org/10.29311/ndtns.vi19.4705

Criteria	Student 1	Student 2	Lecturer 1	Lecturer 2	Bing AI	ChatGPT-4
Idea	Sahara solar field	Cloud generation for solar reflection	AI renewable energy generation monitoring	No suitable idea provided	No suitable idea provided	Terraforming Algae blooms for carbon capture
Relevance	9	7	3	-	-	6
Coherence	7	4	8	-	-	6
Effectiveness	7	4	5	-	-	6
Efficiency	2	4	8	-	-	8
Impact reduction	3	7	7	-	-	7
Sustainability	7	6	8	-	-	9
Scalability	5	3	4	-	-	8
Versatility	2	5	7	-	-	2
Profitability	8	2	5	-	-	3
Total score	50	42	55	-	-	55

Table 3 Climate Change responses evaluation

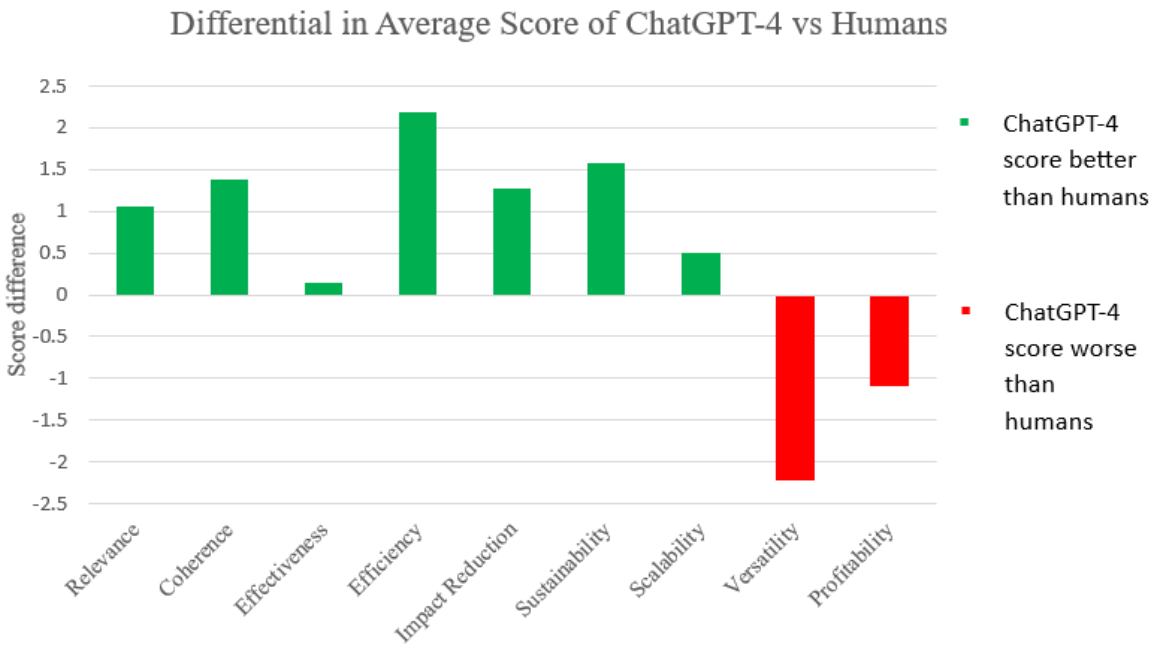

Figure 2 Differential in average score between ChatGPT-4 versus Human respondents

Figure 2 shows the difference between the scores for the average human answer and the average ChatGPT-4 answers for each of the criteria. Where the bar is green, the AI provided a better score for that specific criterion and where red, the AI was less able than humans.

The high efficiency, sustainability and coherence scores for ChatGPT emphasises that the AI was better at directly addressing a problem in a clear and effective manner. It did not get creative with the

https://doi.org/10.29311/ndtns.vi19.4705

method of addressing the problem but just gave a direct answer according to the prompt. ChatGPT-4 often included implementation methods in the answers to the prompts. Because of this, the AI considered the cost and time of actioning the intervention, leading to much more efficient answers than that of the humans.

Where the score differential is shown as red on the chart, it demonstrated that AI struggled with versatility and profitability. This is due to that direct answer form of response. When working with a human to solve a problem, the conversation can carry the idea generation away from just simply answering the question at face value. In this way, humans can generate more versatile ideas compared to AI. The profitability is another symptom of this direct response. As the AI was never instructed to make the intervention profitable, it did not attempt to in any way.

Discussion

Human Results Analysis

Each of the participants had noticeably different approaches to solving the problems. Both students had more of an erratic random ideas approach with student 1 relying more on their initial concepts and student 2 having more focus on creation of ideas through the conversation. With their lesser experience in physics, the ideas were less physically grounded but one that stood out was an idea about cellulose packaging (Liang et al, 2022). This concept considers cellulose fibres being turned into plastic alternatives. This was proposed as a solution to the excess of plastic packaging that is used in the food industry. This scored a 63/90 which is one of the highest scores given.

The **l**ecturers were far more structured with their approach with **l**ecturer 1 outlining each of the causes of the problem before tackling it and lecturer 2 equat**ing** many of the real-life problems to physical models to approach it in a less emotional, more analytical and logical way. This led to interesting results: Lecture 1 proposed an interesting idea about AI smart crop care. This entailed having an AI algorithm that can monitor and adjust the amount of light and water each plant gets for optimal yield. This could be very profitable for farmers and scored well with a 57/90. On the other hand, lecturer 2 proposed an idea of making foods more filling so that a smaller volume of food is consumed but everyone still get all the nutrients they need. If the norms of society are ignored, then this could be a useful solution, but it would involve removing the whole food infrastructure and creating a whole societal change. Interestingly, these routes both address a similar problem from opposite routes. ‘Food availability’ could be seen as an issue where there is a need to create a greater volume of food, or it could be about making the food that is available more effective.

Across all the human generated solutions, there was an array of useful and less useful answers, but they were all very different. Given the same prompt, with a bit of conversation, all the discussions arrived in completely different places. The less useful answers are sometimes based in absolute absurdity but if distilled down to what problem they solve and other routes to the same solution are theorised, they can be helpful. Another merit to the human conversation was the fact that they do **no**t always attack the problem head on. Student 2 had the idea of a ‘fast defroster’. This can already be done by a microwave but not without sacrificing the quality of the food. Student 2 stated that if there was a quick and easy way to get frozen food to a similar state as if it had been refrigerated then it would incentivise freezing food more often. This, in turn, would reduce food waste. Student 2 has not thought of a solution to directly reduce food waste but has thought of promoting current solutions. This is an example of humans’ ability to address a problem by looking at a bigger picture using abstract thought processes.

Generative AI Responses Analysis

Bing AI Analysis

Despite being powered by GPT-4, Bing AI was underwhelming for this study. It is mainly a smart search engine rather than a GenAI in the way that OpenAI’s Chat-GPT is. When asked to outline the issues around the given topics, it had no problems providing a list of issues**. H**owever, when asked to

https://doi.org/10.29311/ndtns.vi19.4705

come up with a way to address these issues, it merely replayed existing solutions. An example of this is that when questioned about applying physics to address the endangerment of animals, it recommended sending up more satellites to monitor them from a distance. This is not a useful solution as the cost of doing this would outweigh the value of the data received, and it evidently has no creativity behind the idea as all it has done is think 'satellites require application of physics principles' and 'monitoring is a way to help reduce the endangerment of animals' and stuck them together in a way that is not original or interesting. The Bing AI 'ideas' did not score well in any of the criteria.

When asked about fossil fuel alternatives, Bing AI provided the normal renewable energy methods (e.g. solar, wind, geothermal etc.) but failed to mention nuclear fission in any way. This is a method of energy generation that has proven itself to be efficient and is at the stage where it is relatively safe due to learnings made following some historic catastrophic events (Lea, 2022). When questioned about the omission of nuclear fission as an interim power source until there is a larger renewable energy infrastructure, Bing AI stated that it did not include it as it is deemed controversial. This is a very important discovery as The AI has access to the information of how good nuclear fission power could be yet chooses to ignore it due to some public opinions. This is a limiting factor in the field of providing solutions to global problems as sometimes the public opinion might not reflect what is best for the planet. Bing AI tended to be difficult to work with and often got stuck in loops of repeating the same response regardless of the prompts.

Chat GPT-4 Analysis

Chat GPT-4 uses the same Large Language Model as Bing AI but was very different performance-wise (OpenAI, 2023). It consistently showed signs of critical analysis and creativity. In addition, when using Chat-GPT-4, specific prompt engineering techniques were less important than when using Bing AI and it understood what was being asked of it more easily. It had no issues outlining the issues around a certain one of the given topics and provided a short solution to each in its initial response. These solutions were not developed in any way and were ignored as they provided little value. When asked to expand on one of these solutions however, Chat-GPT-4 provided unique and creative ideas with a bit of physics backing it up and even implementation strategies. For example, in the endangerment of animals, the AI was asked to create a way to stop light and noise pollution disturbing natural habitats. 'The Dynamic Environmental Shield' was proposed as a solution. It is bordering on the absurdity of the human ideas. It involves a dome made of electrochromic materials to block out artificial light and large poles that act as giant noise cancelling headphones for an area of nature. In practice, this is not accurate physics, but some adjustments could be made.

Electrochromic materials can only control the amount of light they let in by ionizing the conductive coating on the glass. When a charge is applied, the ions build up on one side of the glass, reducing the transparency. Depending on the charge put through the glass, the electrochromic glass can let more or less light in. For the noise cancelling poles, these are not possible in the way that they were proposed by Chat GPT-4. This is because noise cancelling headphones only work when focusing on cancelling the noise at one point. They work by creating an 'anti-sound' wave that has the same frequency and amplitude as the surrounding sound, but an opposite phase. The poles would create pockets of silence but also due to the wave nature of sound, they would also create areas where the sound is essentially doubled by the poles. While there are noticeable problems with the idea, the comparison between ChatGPT-4 and Bing AI is significant as ChatGPT has attempted to apply physics and logic to create a unique idea whereas Bing AI has not.

Conclusion

This research has assessed the ability of generative AI to apply physics principles to global problems to arrive at solutions. While the human conversations revealed far more versatile solutions, ChatGPT-4 excelled in creating coherent and efficient solutions. ChatGPT-4 had some trouble with applying accurate physics to the situation, but it tried. Bing AI had no unique concepts to offer and was not found to be effective in this research. The human responses showed an array of unique ideas with

https://doi.org/10.29311/ndtns.vi19.4705

varied levels of physics accuracy. The dynamic conversations fielded interesting ways to use the ideas in other ways or develop the ideas further.

Implications of the work

The introduction of Artificial Intelligence into society has already had a distinct impact on business, education and the workplace. This research shows that it also has a place in enacting global change. It can generate Physics-based ideas and solutions faster than any human can and these are, in many aspects, better than the ideas from humans at their peak of knowledge in physics. This capability is essential given the pressing nature of the current challenges faced by humanity. Certain areas still require the Chatbots to become more advanced and versed in the logic and mathematics of the world, but soon LLMs will be more capable than humans in every metric of idea generation and physics application.

Limitations of the Research

This research had some limitations. The sample size of the human participants was small, with only four sources. This is a similar case with the AI programs that were tested. This research only covers two AI Chatbots in Bing AI and ChatGPT-4. If the research was to be repeated, a greater number of participants and AI programs would yield more accurate results. The changing landscape of AI leads to information on it quickly becoming outdated (Safrai and Azaria, 2023). There is a constant need to refresh experimental data before it becomes obsolete. While this research was up to date as of 05/03/2024, it will have to be updated as new GenAI tools are released.

Potential Future applications

Clearly generative AI and Humans have different strengths when it comes to applying physics in real world scenarios. Humans can approach an issue in different ways and discuss with another person to develop an idea or create new ones. The diversity gained by human perspective is essential in idea creation. AI can create a fully formed idea in seconds with implementation strategies and a baseline of physics applied. This idea can be based on information from its vast database and the AI can be questioned and develop its ideas, but it will take a problem at face value and not assess the broader causes for a different type of solution. Surprisingly, the best AI chatbots are far better at ‘thinking creatively’ than they are at maintaining accuracy of underlying physics (Renato, 2023).

For the most efficient idea creation method using AI, a group of physicists can come together to lay out a problem and think of its causes and a range of ways to address the issue. AI can be used to create a series of ideas for each of these solution routes. Humans can then assess the answers, think critically about the physics used and tweak them until they agree on an optimal strategy.

Currently, in isolation, AI cannot effectively apply physics to solve global problems, but this research shows that it can improve on many aspects of decision making. It can generate clear and concise ideas quickly that can be adapted by experts into impactful interventions. This process will only get more powerful as AI evolves and gains a better understanding of our world and the issues at hand. Moving forward, AI will not only be able to address global problems but apply its power and logic into the frontier of physics (Durante, 2024), pushing the boundaries of our understanding of the universe as we know it.

References

Blasi, A.D. (2023). A Symphony of Senses: Introducing Sensory AI: A Pathway to Achieving Artificial General Intelligence, *Medium*, [Online]. Available from: https://medium.com/@aarondiblasi/a-symphony-of-senses-introducing-sensory-ai-a-pathway-to-achieving-artificial-general-8bd731154d09.

Borji, A. (2023). A Categorical Archive of ChatGPT Failures. arXiv:2302.03494 [cs].

Choi, W., Park, J., Han, D.J., Park, Y. and Moon, J. (2024). Consistency-Guided Temperature Scaling Using Style and Content Information for Out-of-Domain Calibration. arXiv.org. [Online]. [Accessed 6 March 2024]. Available from: https://arxiv.org/abs/2402.15019.
Durante, Z., Sarkar, B., Gong, R., Taori, R., Noda, Y., Tang, P., Adeli, E., Lakshmikanth, S.K., Schulman, K., Milstein, A., Terzopoulos, D., Famoti, A., Kuno, N., Llorens, A., Vo, H., Ikeuchi, K., Fei-Fei, L., Gao, J., Wake, N. and Huang, Q. (2024). An Interactive Agent Foundation Model. arXiv.org. [Online]. [Accessed 6 March 2024]. Available from: https://arxiv.org/abs/2402.05929.

He, Y., Qiu, J., Zhang, W. and Yuan, Z. (2024). Fortifying Ethical Boundaries in AI: Advanced Strategies for Enhancing Security in Large Language Models. arXiv.org. [Online]. [Accessed 6 March 2024]. Available from: https://arxiv.org/abs/2402.01725.

Hughes, O. (2023). Generative AI defined: How it works, benefits and dangers. TechRepublic. [Online]. Available from: https://www.techrepublic.com/article/what-is-generative-ai/.transforming-the-world/.

Irion, S., Christaki, U., Berthelot, H., L'Helguen, S. and Jardillier, L. (2021). Small phytoplankton contribute greatly to CO_2-fixation after the diatom bloom in the Southern Ocean. The ISME Journal.

Kelly, W. (2024). GPT-3.5 vs. GPT-4: Biggest differences to consider | TechTarget. Enterprise AI. [Online]. [Accessed 6 March 2024]. Available from: https://www.techtarget.com/searchEnterpriseAI/tip/GPT-35-vs-GPT-4-Biggest-differences-to-consider.

Kumar, M. (2023). Understanding Tokens in ChatGPT. Medium. [Online]. [Accessed 6 March 2024]. Available from: https://medium.com/@manav.kumar87/understanding-tokens-in-chatgpt-32845987858d.

Lea, R. (2022). What is nuclear fission? Space.com. [Online]. Available from: https://www.space.com/what-is-nuclear-fission.

Liang, Y., Ries, M.E. and Hine Peter Hine, P.J. (2022). Three methods to measure the dissolution activation energy of cellulosic fibres using time-temperature superposition, *Carbohydrate Polymers*, p.119541.

Mittal, A. (2023). The Essential Guide to Prompt Engineering in ChatGPT - Unite.AI. www.unite.ai. [Online]. Available from: https://www.unite.ai/prompt-engineering-in-chatgpt/.

Mittelstadt, B., Wachter, S. and Russell, C. (2023). To protect science, we must use LLMs as zero-shot translators, *Nature Human Behaviour*, 7(11), pp.1830–1832.

Niemela, J.J. (2021). Physics for a better world, *Nature Physics*, 17(8), pp.871–872.

OECD. (2023a). Evaluation Criteria-OECD. www.oecd.org. [Online].Available from: https://www.oecd.org/dac/evaluation/daccriteriaforevaluatingdevelopmentassistance.htm.

OECD. (2023b). OECD.org - OECD. Oecd.org. [Online]. Available from: https://www.oecd.org/.

OpenAI. (2023). GPT-4 Technical Report. arXiv:2303.08774 [cs].

Redman, M., King, A., Watson, C. and King, D. (2016_. What Is CRISPR/Cas9? *Archives of Disease in Childhood-Education & Practice Edition*, 101(4), pp.213–215.

Renato. (2023). Enhancing Chemistry Learning with ChatGPT and Bing Chat as Agents to Think With: A Comparative Case Study. arXiv (Cornell University).

Roy, A. (2020). Introduction To Autoencoders, Medium, [Online].
Available from: https://towardsdatascience.com/introduction-to-autoencoders-7a47cf4ef14b.
Sahoo, P., Singh, A.K., Saha, S., Jain, V., Mondal, S. and Chadha, A. (2024). A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications, arXiv (Cornell University).

Safrai, M. and Azaria, A. (2023). Performance of ChatGPT-3.5 and GPT-4 on the United States Medical Licensing Examination with and without Distractions. arXiv.org. [Online]. [Accessed 6 March 2024]. Available from: https://arxiv.org/abs/2309.08625.

Sarma, S.D. (2023). How AI and ML Will Affect Physics, *Physics*, 16, p.166.

Walsh, M. 2023. ChatGPT Statistics. (2023). Essential Facts and Figures. Style Factory. [Online]. Available from: https://www.stylefactoryproductions.com/blog/chatgpt-statistics.

West, D. and Allen, J. (2018). How artificial intelligence is transforming the world. Brookings. [Online]. Available from: https://www.brookings.edu/articles/how-artificial-intelligence-is-[14]