499 Proposal
Harshavardhan Reddipalli

Introduction

After conducting extensive research in artificial intelligence and large language models, I have chosen this area as the topic for my honors research. The focal point of my honors thesis revolves around a comprehensive investigation into the impact of Artificial Intelligence and Large Language Models on education. I've chosen this topic as LLMs have entirely changed the way students learn nowadays; about a year ago, students would have to visit several websites and take down information to understand a topic, but with the introduction of these models, you can get your answer with a quick search. The decision to delve into this topic also stems from my genuine passion for the transformative potential of AI and LLMs in reshaping the education sector. In recent years, technologies like ChatGPT, Bard, and Bing AI have already made substantial inroads into various industries, and academia is no exception, where they are altering how students learn and acquire knowledge. AI and LLM technologies offer a personalized learning environment by analyzing extensive datasets to tailor educational content to the person's need, which has been incredibly beneficial. However, I've observed peculiarities with these technologies, and that's another thing that piqued my interest. Whenever I ask mathematical questions involving numbers, it generates the correct equations, but struggles with basic arithmetic calculations. For instance, when I asked, "What is 1234 times 5678," it provided an incorrect answer, whereas a basic calculator on your phone would have given the correct result. Moreover, with this discovery, my research aims to unravel the nuances of LLM's limitations in specific subject areas, shedding light on the intricacies that can impact their effectiveness and prompting a deeper exploration into refining these technologies for a more comprehensive educational support system. The rise of these technologies has also raised many concerns regarding ethical considerations, and I will talk more about this later. My thesis aims to comprehensively understand the potential, challenges, and implications of integrating these AI and LLMs in education using the methods explained later. I aim to provide valuable insights that can drive positive change in educational and technology practices.

Now, let me introduce the scientific question or the creative endeavor. As mentioned, my research topic is "Investigate the impact of AI and LLMs on education ."I have not stated a set scientific question for my research as the subject is intricate and involves various factors that need careful consideration. I have decided to take a creative endeavor approach. My research aims to understand the impact these AI and LLMs have on learning, the potential, challenges, and implications of implementing these in the educational space for the future. I also plan to determine what subject areas seem more practical and if they offer any benefit over the traditional learning method. As mentioned above, the primary objective of my research would be to assess the effectiveness of AI and LLMs in improving student engagement and academic performance. This includes evaluating the accuracy of these technologies in providing educational information, identifying potential mistakes or false statements made by these models, and investigating the role of these technologies in enhancing the overall academic abilities of students. I have about three hypotheses based on my time and exposure to these technologies, and they would be, I have come up with a null hypothesis and an alternate hypothesis as well. The first null hypothesis (H0) is that there is no significant difference in academic performance between using AI and LLMs and traditional methods. The alternate hypothesis (H1) would be that using AI and LLMs significantly improves academic performance. The second null hypothesis (H0) is that the accuracy of AI and LLMs in providing educational information is consistent across different subjects. The alternate hypothesis would be that AI and LLMs show accuracy variations depending on the complexity of the subject matter.

Lastly, the third null hypothesis (H0) is that using these technologies in education does not decrease critical thinking and problem-solving skills. The alternate hypothesis (H1) would be that using these technologies in education reduces critical thinking and problem-solving skills.

Coming to other discipline-specific inquiries, I have also decided to figure something out with the help of my research, I plan to examine educators' perspectives regarding integrating AI and LLMs in the classroom, as this aspect plays a pivotal role in shaping the overall success and effectiveness of these technologies in educational settings. Understanding how educators perceive and engage with AI and LLMs will provide crucial insights into potential barriers, challenges, and opportunities associated with their implementation.

Now, let me review the significance of the research question or the creative endeavor. As mentioned above, this is personally significant for me. The choice to explore this topic stems from my genuine interest in the transformative potential of AI and LLMs and their growing presence in education. This research topic also holds significant importance in the current technological landscape and its influence on educational practices, as numerous studies highlight the benefit of technology-enhanced learning. Though the rise of these technologies has only boomed in the recent year or so, their influence will likely endure for a considerable amount of time. The integration of these technologies is a recent phenomenon, and there is much need for extensive research to understand its implications in various sectors of life, but more importantly for my research, education. The impact of these technologies on the traditional learning process is evident, with students experiencing a shift from conventional methods to quick and readily available answers. This shift also raises questions about the accuracy and reliability of these models, especially in disciplines involving complex mathematical calculations. As said before, this research is crucial, as I aim to provide valuable insights that can drive positive change in educational and technology practices. My creative endeavor relates to key literature in AI, LLMs, and education. A lot of literature discusses the potential of these technologies in various sectors, including education *(Analytics Vidya, 2023)*. However, there needs to be more literature regarding the accuracy of these models and the potential mistakes or false statements they might make *(Nature, 2023)*. A report from MIT's Computer Science and Artificial Intelligence Laboratory also highlights the inconsistencies in the responses generated by LLMs, leading to flawed reasoning. Addressing ethical and bias concerns related to AI and LLMs further enhances the significance of the research, as with the use of these technologies in education, an extensive amount of personal information could potentially be exposed on online platforms and the other hand, there are several concerns about the potential of these technologies to perpetuate existing systemic bias and discrimination *(Springe Link, 2021)*. My research aims to gain a more comprehensive and in-depth understanding of these issues. My research is essential for advancing knowledge in the discipline, as by investigating the impact of AI and LLMs on education, I aim to contribute to our understanding of how these technologies can be effectively used in an educational setting. It will also provide insights into the benefits and limitations of these models, as well as ethical and bias concerns. My research will also offer practical implications for teachers, policymakers, and tech developers by providing them with strategies to seamlessly integrate these models into the educational landscape, reshaping it for a brighter student future. My research will contribute to the ongoing discourse on the role of these technologies in the educational landscape. In conclusion, my creative endeavor is significant as it addresses a contemporary issue, relates to the key literature, and contributes to advancing knowledge in the discipline. As mentioned before, my research is timely and necessary in the current educational landscape, where these technologies have become increasingly prevalent.

Literature Review

Integrating Artificial Intelligence and Large Language Models into educational environments has sparked considerable interest and debate. This literature review delves into key discussions, findings, and perspectives surrounding using LLMs in classrooms.

Although I have gone over a few key pieces of literature in the section above, I will delve more into literature regarding the objectives and issues of my research. Firstly, regarding LLMs like ChatGPT in the classroom, researchers, educators, and companies are experimenting with ways to turn flawed but famous large language models into trustworthy, accurate 'thought partners' for learning *(Andy Extance, Nature, 2023)*. According to this article, Ronald Beghetto, an educational psychologist based out of Arizona State University, had his graduate students discuss their work in unusual ways, this being a creativity-focused chatbot that he had designed that is based on the same AI-powered technology that ChatGPT runs on. The feedback received from the students was overwhelmingly positive, and these bots helped generate more possibilities than they would have considered otherwise. Hence, with the knowledge of this, we could hypothesize that if the LLM is tweaked right, it would be greatly beneficial in the educational landscape, but we won't know for sure until more extensive research has been done. While many fear that the rise of these LLMs will increase the likelihood of students cheating on their assignments, Beghetto argues it could instead enhance the way education is learned. Some educators also see them as potential 'thought partners' that might cost less than human tutoring, and this is crucial as it will make education more accessible to those who require it. According to Theodore Gray, the co-founder of Wolfram Research, while one-on-one tutoring is the most effective, it is expensive and not scalable. However other software has been used to fix this, but it is not as effective, so with the help of LLMs, there could be a real possibility of making educational software work.

The potential of ChatGPT is humongous, and according to Jules White, director of the Initiative on the Future of Learning and Generative AI at Vanderbilt University in Nashville, Tennessee, ChatGPT is "an exoskeleton for the mind." However, like any other software, there are risks with LLMs, too. According to the article, the primary concern is that students might be too reliant on LLMs to do quick work without understanding the rationale. Another concern is that ChatGPT can lead students astray as the bot is notoriously brittle, gets things wrong if the question is phrased differently, and even makes things up. According to Wei Wang, a computer scientist at the University of California, Los Angeles, ChatGPT had got a lot of questions wrong when tested on questions in physics, chemistry, computer science, and mathematics taken from University-level textbooks and exams. Wang and her colleagues experimented and found that even with the right phrases, ChatGPT had gotten only one-third of the questions correct. This is also a crucial concern to my research, as LLMs have been known to perform better in subject areas with fewer arithmetic calculations. If LLMs are implemented into students' learning environments, this gap has to be addressed. Another concern is privacy and how OpenAI stores the prompts to train its models. According to this article, despite the risks, educators see huge potential in using artificial intelligence chatbots to enhance teaching and learning. According to Collin Lynch, a computer scientist at the University of North Carolina State, although there are risks associated with these technologies, we have to learn to mitigate them and embrace LLMs. My understanding so far from this article has been that, like any other technology in its early phases, there are a few hiccups while using it in our daily lives, but with the increased amount of time, data collection, and training, I do believe that these tools could one day have a major impact on the educational landscape for the better. Even UNESCO recommends that educational institutes validate tools such as ChatGPT before using them to support learning, and according to my knowledge and the current trend of where this is headed, I believe that one day, they will be a vital tool in education. The article mentioned above also discusses another approach to implementing these technologies into education, and that is with the help of AI tutors. The tool, Khanmigo, a result of the partnership between Khan Academy and ChatGPT, takes a different approach to implementing AI in education, offering students tips as they work through the exercise. This tool appears as a pop-up chatbot on a student's computer screen as they discuss the problem that they are working on with it. The tool automatically adds a prompt before it sends the student's query to GPT-4, instructing the bot not to give away answers and instead to ask lots of questions. This is an approach where the student doesn't completely rely on technology but receives help when required. However, according to Kristen DiCerbo, this is a 'productive struggle' as there is a fine line between a question that aids learning and one that's so difficult that it makes students give up. However, unlike regular ChatGPT, there aren't a lot of privacy concerns as OpenAI has agreed not to use the data from Khanmigo for training purposes. There are still mistakes in the answer as it works out of the same database as ChatGPT, but to improve its accuracy, Khanmigo sends the correct answer before sending its prompt to GPT-4. As mentioned, there are still concerns, as with any other chatbot; it should be checked for its tone not to belittle the students. To minimize mistakes, another approach that has been discussed is RAG, which integrates LLMs with textbooks that have been rigorously verified for accuracy. This is a dedicated educational LLM rather than an adaptation of an existing, general-purpose model, such as ChatGPT or Bard. This is already being used at ASU, and it has been said to have a lot of positive impacts. This would be the approach for the growth of LLMs, as initial research has proven that a tweaked LLM for education has more accurate responses. After reading this article and understanding the literature, it is clear that a lot of people in the education sector are pushing for the implementation of LLMs in education but still have to figure out the right sort of approach, as there are a lot of risks and concerns regarding various factors of it. To study the impacts of generative AI on teaching and learning, on December 8th, 2022, researchers from OpenAI, KhanAcademy, the Berkman Klein Center for Internet & Society at Harvard University, and other invited experts gathered to discuss the impacts of ChatGPT, and generative AI more broadly, on the future of teaching and learning. Participants included a cross-section of educators, including high school teachers, university administrators, professors, deans, and experts in computer science, law, public policy, philosophy, and other fields. The discussion focused on four areas: how skill and knowledge assessments might change; potential educational use-cases for generative AI and how they might integrate into the classroom; the effect of LLMs on cultivating creativity, skepticism, and novelty among students; and the practical, immediate issues teachers will face as a result of this technology and how might we mitigate these concerns. Through these discussions, many inquiries were raised; firstly, although LLMs offer opportunities to enhance education, they raise concerns about inequalities and changing educational dynamics. Students grapple with AI's impact on critical thinking and writing tasks, prompting a need for balance in skill development and knowledge acquisition. Privacy, academic integrity, and biases in AI outputs also require attention. Policy considerations are essential for defining roles and responsibilities in the evolving educational landscape. From these discussions, there were a lot of privacy concerns as well, related to the things I've discussed before, and to implement these LLMs into education, these have to be addressed. Talking about cognitive bias in AI and LLMS, according to a study titled "Challenging the appearance of machine intelligence: Cognitive bias in LLMs and Best Practices for Adoption" by Alaina N. Talboy and Elizabeth Fuller, there is a presence of cognitive biases in LLMs. The authors argue that while assessments of algorithmic bias in LLMs generally focus on systemic discrimination based on protected characteristics such as sex and ethnicity, over 180 documented cognitive biases pervading human reasoning and decision-making are often overlooked when discussing the ethical complexities of AI. Although this is not directly in correlation to what my research will be based upon, it is good to know other concerns related to the field of AI and LLMs that have negative implications. As mentioned, teachers seem to have an overall positive opinion of AI and LLMs. According to a conference paper titled "How useful are educational questions generated by Large Language Models?" they conducted a human evaluation with teachers to assess the quality and usefulness of questions generated by LLMs. This evaluation showed that teachers positively approve of the widespread use of LLMs in a classroom setting. This is crucial information to know and a hypothesis. Although this evaluation doesn't go in-depth, for example, classifying the results by teachers in different subject areas, it showcases the teacher's interest in implementing these technologies in the classroom. The accuracy of LLMs in different subject areas has also been discussed in recent research. An article titled "Evaluating Large Language Models: Methods, Best Practices & Tools" discusses the importance of evaluating LLMs across various criteria, which include accuracy, safety, and fairness. This article highlights the importance of evaluating LLMs for criteria that might often be overlooked. The article states that some critical reasons to evaluate LLMs include Ensuring Optimal Quality and Bias and Ethical Oversight. This has also occurred to me, which is why my research will comprise these elements. The article also emphasizes the need for human evaluation for criteria such as reliability, safety, and fairness, and that is another reason why this is something that I will cover in my research. Talking more about the accuracy of LLMs, which will be the principal element of my research, according to a report by the Bundesamt für Sicherheit in der Informationstechnik, the best LLMs in spring 2019 answered 32% of the questions correctly, which was only slightly above the value of 25% for pure guessing among the four multiple-choice answers. Although this research was done over four years ago and LLMs have come a long way since then, it gives us a good benchmark for the situation four years ago and if there has been any change. According to my hypothesis, which relates this article to the one I've mentioned previously, both researchers only get about 30% of the correct answers using LLMs. However, it is important to note that the test on LLMs at the University of California study focused on much harder STEM subject areas, whereas this research done in 2019 was more general. Nonetheless, these two articles give us an idea of what we might expect when I conduct my research. After doing more literature reviews, I found some data on the accuracy of LLMs in different subject areas. According to the Association for Computational Linguistics, LLMs have been shown to achieve good zero-shot accuracy on many numeric Math Word Problems. However, they often provide incorrect answers when solving arithmetic reasoning tasks. For example, a study found that LLMs can sometimes provide correct solutions to difficult problems while failing at trivial ones. Another study found that LLMs perform below an average mathematics graduate student and fail on simple mathematical problems in natural language. This is an area of concern as with the articles above, we have found out that the responses by LLMs change by the way the prompt is phrased, thus affecting its reliability and validity. Regarding chemistry, LLMs have been found to write code across various areas of chemistry, but they may provide unreliable results to each standalone question. Coming to physics, LLMs don't seem to be performing any better, as a study by Tomohiro Sawada found that current LLMs score well below 50% on demanding tasks. However, LLMs are not all accurate as, according to various articles, their accuracy is often on par with neurotypical adults in subject areas such as cognitive psychology and theoretical sciences. Hence, we can hypothesize that they struggle on topics that require more numerical calculations but are often accurate on more theoretically based topics. To conclude, positive outcomes of LLMs, such as enhanced creativity and cost-effective tutoring, are noted, but there are various concerns about reliability, privacy, and bias. Alternative approaches aim to solve these challenges but are not there yet. The main concern regarding these technologies is the accuracy and validity of content. However, preliminary research finds that they are more likely to perform better in subject areas that are more theoretical. There is also a lot of discussion among experts about the need to evaluate the impacts and mitigate practical concerns.

Methods

Let me review the method I will use for my creative endeavor. Firstly, I would choose three topics in mathematics and three in psychology and later take a test on the same platform, Khan Academy, for each. Let's call these M1, M2, M3, P1, P2, P3 respectively. For M1, I will learn the topic using traditional learning methods, including online searches, YouTube, and textbooks. For M2, I will learn the topic completely using an LLM like ChatGPT. For M3, I will not learn the topic but will take the test using the help of these LLMs. I would follow the same concept for P1, P2, and P3 as I did for M1, M2, and M3. The first topics, M1 and P1, are a benchmark for what we can expect to learn using the traditional learning method. They would then be compared with the results of M2 and P2 to compare the learning results using traditional methods and LLMs. Lastly, M3 and P3 would be used to figure out another objective we are trying to figure out: the accuracy of these LLMs in different subject areas and if that method of taking tests results in better results than learning the topics. To ensure the comprehension of information learned using these methods, I would retake the tests for each subject area without studying again to see how much information I have retained while learning using different methods. Using these methods, I can figure out a few objectives; firstly, I can figure out in what subject areas LLMs are more accurate, and if they are in any, this is done by comparing the mean results of M and P. Secondly, I can figure out if learning using AI and LLMs are beneficial comparing the results from M1, P1 and M2, P2 to see if there is a difference in the respective scores. Lastly, I can figure out the accuracy of these LLMs by comparing the results of M3 and P3 with a pre-existing answer sheet to calculate the accuracy in two different subject areas. Now, to figure out another objective, the opinion on using AI and LLMs in the classroom, I will be surveying about 20 - 30 professors, asking them about their preferences, concerns, and what level of inclusion they are willing to see if any. Coming to the actual procedure and techniques that I would be using for my research, there are various versions of LLMs and AI technologies, but to get the most accurate and possible valid results, I will be using GPT-4 from OpenAI as it has been known to receive more accurate results compared to other LLMs like Bard and BingAI. However, as I write this proposal, a breakthrough has occurred. Google has developed another AI/LLM technology, Google Gemini, that would power its Bard chatbot. From the initial reactions, it might be more powerful than GPT-4, so taking this into account, I will choose my LLM as soon as possible. Regarding professors' opinions, I will conduct a Google Forms survey as this seems to be the most efficient method to get results. Unlike other survey methods, the professors usually know how it works and require no pre-existing knowledge. Lastly, I plan to determine the bias and ethical concerns surrounding AI and LLMs. To go about this, I will mostly focus on a theoretical aspect but will do some testing of my own. Bias usually stems from the data the model is trained on; however, due to privacy concerns, it will be hard for me to access this data, so that I will do bias testing on these LLMs instead. To go about this, I will create specific inputs that check how these models respond to different demographic groups or controversial topics. For example, I would measure the sentiment of the model's responses to identical prompts, only changing the associated gender or race. I would also focus my research on this part of the literature review, as it gives me a general idea of what to expect and what other researchers have already discovered in this field. I also plan to interview experts and gain their valuable opinions and insights into this issue. Lastly, I also plan to find out if using these LLMs over time has a decrease in one's critical thinking. So, to go about this, I will be using 2 participants who are not me; let's call these P1 and P2. P1 will be exposed to LLMs for a specific amount of time, I think about a month, whereas P2 will not be exposed to LLMs at all during this experiment. I would ask P1 to continue with his usual college life but incorporate LLMs more and P2 to refrain from using any LLMs for this month. Periodically, maybe twice a week, I will conduct 'critical thinking' tests for these two participants and take down the scores. These tests will be of various subjects, but both take the same test each time. At the end of the month, I will compare the scores of the control group and the group that used LLMs; based on the results, I will conclude how 'critical thinking' is affected by LLMs, if any. As mentioned before, the materials I will use in my research are LLMs and literature reviews while conducting tests and surveys. No specialized training will be required for my research. In conclusion, this comprehensive research plan I have made aims to assess the effectiveness of learning and testing with LLMs in different subject areas, explore professors' opinions on AI/LLMs in education, investigate biases and potential impacts on critical thinking, and hopefully provide valuable insights into the integration of these technologies in educational settings.