Large Language Model - Ethics Paper

Zahir Choudhry
zchoudhry@oxy.edu
Occidental College

I. Introduction

Large Language Models have exploded in popularity since OpenAI released a ChatGPT to the public. Reaching one million users in just five days, ChatGPT is recognized for bringing Artificial Intelligence to the forefront of global conversations as well as providing the public with a free tool that has passed the bar exam, software engineering interviews, and other impressive accomplishments. While OpenAI and other Artificial Intelligence / Large Language Model companies have contributed so much to the world, there are other morally ambiguous aspects to Large Language Models that have yet to be addressed, neither by society nor the government. For my senior comprehensive project, I am looking to create a Large Language Model trained and fine-tuned on Occidental College data to create a singular place for students to learn about the school and find helpful school resources all in one location. While I have hope for the use of my project to improve the college experience of current and prospective students, I feel it is necessary to address the current pitfalls and ethical compromises of Large Language Models as well as look into potential problems that have yet to be addressed. The paper will proceed as follows, first I will address data bias and the dangers it presents to ethical machine learning projects as well as the work being done to reduce data bias. Next, I will go into the importance of algorithmic transparency and how it proves challenging in the case of language models. After transparency I will cover the issues of privacy, both in how data collection can invade a person's privacy while also addressing how a large language model can leak data. In acknowledging these claims, I hope to be able to address the potential ethical concerns of my project as well as inform the reader of how other platforms have approached these issues in the past.

II. Data Bias

In order to train a Neural Network, there is a need to collect vast amounts of data, whether that be for Large Language Models, Computer Vision, Natural Language Processing, or other machine learning fields. While data collection itself is not inherently a bad thing, there can be instances of biases in data sets that can drastically affect the performance of machine learning models like Large Language Models. A famous instance of biased statistics is evident in the crime data collected by the United States police force, which has been influenced by racial prejudice, particularly towards African Americans[**levin˙california˙2024**]. If the police were to use such a dataset on a computer vision model that identifies potential criminals, it is likely the program would unfairly discriminate against African Americans, profiling innocent people as criminals. A more concrete example of data bias is when Google's facial recognition program falsely classified African American as Gorillas on Google Images[**buranyi˙rise˙2017**], something which came about from the algorithm being racist data that relates African Americans to gorillas, among other animals. While the bias in the data is apparent in these examples, there is potential for other less studied or unused data sets in machine learning projects that may result in training a neural network to reinforce and perpetuate societal injustices and further cement these stereotypes into the zeitgeist of the world. With this in mind, it is extremely important that the data that any machine learning project is trained on is very broad, diverse, and uninhibited by social inequalities and systemic injustice. At this stage in the drafting process, I am uncertain whether I want to train a model from scratch or use a pre-trained model and fine tune it, so I will be tackling the issue with both of these methods. If I choose to go with a completely untrained model, my current plan is to train it using articles from The Occidental, the official newspaper of Occidental College. It is possible that within these articles there is bias that may feed my neural network false information. For example, if there were an article where the author wrote about CalTech and said something slanderous or untrue, my machine would accept that as fact and perpetuate this idea without further thought. The solution to this would be to use a pre-trained model, one that has been trained on large amounts of text data, however this too is questionable. In using a pre-trained model, I am placing trust in the group which created this pre-trained model that has an ethical and diverse dataset. Considering that ChatGPT, as the face of Large Language Models, had its own issues with both using and sourcing ethical datasets[**noauthor˙openai˙2023**], I find it very difficult to believe that other groups have been

able to build an effective Large Language Model trained entirely on ethical and diverse data. Whether it be training a model from scratch or utilizing a pre-trained model, both present the possibility of falling victim to data bias, perpetuating the spread of stereotypes and misinformation. [Another concern, could be who controls the data after I leave, in maintenance section]

III. Algorithmic Transparency

Algorithmic Transparency is an essential part of creating a more equitable technological world. Most companies do not participate in this practice, however considering that my project is for academia and not for profit, I have no issue sharing my algorithm as well as the datasets it was trained and fine-tuned on, but that still does not solve this issue. As it stands Large Language Models are a "black box", with experts having no idea how a neural network makes its decisions. While it is possible to track what stimuli used to train a model and how it is corrected when a model makes erroneous predictions, after a certain point the algorithm builds its own unique "trend-finding inclinations", stating that because the machine itself does not keep track of the inputs it uses to make its decisions, it is near impossible to understand how the algorithm got to this decision[**noauthor˙ais˙nodate**]. The two solutions to this problem at present are to either stop the use of deep learning algorithms for "high-stakes applications", for example the European Union is looking to create AI regulations for European businesses, or by creating ways to peer into the so-called black box, for which a concrete method has not been established. While it may be more ethical to remove deep learning algorithms from important projects, these guidelines and regulations assume that it is possible to detect Artificial Intelligence in results and situations. These laws are only as strong as they can be enforced, and if governments are unable to detect the usage of Artificial Intelligence[**manageengine˙university˙2023**] then it is borderline impossible to enforce these idealistic laws on the public. On the other front, there is currently no real way to dissect a Large Language Model. While there is the idea of tracking every single input used to make a decision starting from a model's inception, this solution seems extremely resource intensive. If a company decided to track and store every input for a Large Language Model, they would need to expend vast resources to store and analyze all of that information. On a daily basis, ChatGPT receives over ten million queries [**lammertyn˙60˙nodate**] leading to the processing of hundreds of millions of tokens within twenty four hours. Tracking these tokens would lead to storing at least a trillion total tokens over a year, and over time the maintenance and analysis for this data would not only be extremely resource intensive, but could also lead to security issues, making it possible for hackers to breach the model and potentially train the model on dangerous and incorrect information, further spreading stereotypes and misinformation. Therefore, there is currently no feasible way to create a fully interpretable Large Language Model that does not impose other serious ethical dilemmas. Let us hypothetically say I was able to create a totally transparent Large Language Model. Even though the transparency would allow for the models' hallucinations and erroneous responses to be diagnosed, it would only be understood and effectively accessible to people with knowledge about machine learning and large language models, placing a large amount of power into the hands of people who receive a computer science education. These computer science educated people would have full control over this program and could train it to push any agenda they desire upon the users. While algorithmic transparency is important for creating equity in the tech landscape, there are currently no realistic solutions available to solve the black box predicament. Even if there were solutions, creating fully transparent Language Models gives an unfair amount of power to those knowledgeable about Language Models to control their responses to push their own agendas and spread misinformation.

IV. Privacy

Data privacy is becoming an increasingly large issue as technology continues to grow. As I discussed in the data bias section, large language models require vast amounts of data to be trained and fine-tuned. While data bias is a concern when training machine learning models, another dilemma is the nature of the collected data, whether or not it is legal for a model to be trained on said data. In 2023 OpenAI was cited in a Lawsuit for training models like ChatGPT on illegal data, the contents of this data includes personally identifiable information from hundreds of millions of internet users, including children. This data was collected without the informed consent or knowledge of the users [**noauthor˙openai˙2023**]. Even though the collection of this data was unethical, the culmination of training on said data resulted in the creation of the most powerful Large Language Model to date. While it may have been possible to train ChatGPT on entirely ethically sourced data, the reality is that the amount of data would have been entirely too small. ChatGPT 3 has 175 billion parameters [**noauthor˙openai˙2023-1**], with machine learning projects, like Baidu's Ernie, having parameters reaching into the trillions [**noauthor˙18˙nodate**]. Considering how new the interest in datasets is, it is hard to believe that there are even a billion parameters out there to train a model on. When I searched "companies that create ethical datasets for machine learning" I found that not a single article focused on which companies were ethical, but rather

which ones created datasets that produced the best results in machine learning projects. This google search is indicative of the current climate around data collection. What is the point of creating a dataset ethically if it takes longer to make and could even produce worse than data that already exists? For example, if you were looking to train a language model on how to create an email, it would be both more time and resource efficient to train that model on the billions of already existing emails in the world rather than creating your own faux email which emulates the various natures of emailing.There is currently no incentive to not use private data to train a large language model aside from the legal aspect, and as of writing this paper OpenAI has yet to face any severe repercussions. Aside from the training aspect, there are other privacy concerns involving Large Language Models. ChatGPT has been subject to data leaks and hacks ranging from the individual level to corporate affairs. In 2023 ChatGPT played a role in the leak of sensitive device solutions data from Samsung, resulting in the company completely banning the use of generative AI in the workplace [**park˙samsung˙2023**] . I am extremely concerned about how relaxed the security around such an extremely dominant internet service is. If a multi-billion dollar organization struggles to properly protect its own data, I find it near impossible to provide assurances that my own language model would be able to withstand a malicious cyberattack. Due to the sheer power that large language models hold in the world, a security breach could result in the loss of extremely sensitive data that could have lasting effects on both the victims and those responsible for creating the model. While data privacy is a very important aspect of technology that needs to be addressed, it is clear to me that effective models like ChatGPT opt for results over ethics to achieve infinitely better outcomes than they would have gotten on other data. While I both do not have the option to nor would utilize Occidental College's private student and faculty data, the methods used by companies like OpenAI indicate that my pursuit of ethical data may result in a model that may be lacking in parameters and or performance, making me question the worth of pursuing a project which cannot function properly without violating the private data of every person of Occidental College, as well as any other people whose data may have been violated when Mistral was initially trained assuming I use a pre-trained model.

V. Maintenance

For as expansive and deep the knowledge of ChatGPT 3 is, its knowledge ends before the start of the COVID-19 Pandemic [**noauthor˙can˙2021**], a massive moment that altered the mechanisms of the world as we know it. It has now been four years since COVID and ChatGPT has yet to be updated, likely due to the sheer volume of data it would have to be trained on to be up-to-date. GPT-4 would be released three years later, being updated only until September 2021 [**david˙openai˙2023**]. While training a model on new incoming data itself is not ethically concerning, the issue of finding ethical clean non-biased data sets comes back into question. For my own project, I currently aim to fine tune my model monthly to capture the important changes that have occurred at Occidental College, however I recognize how idealistic this goal seems. While tackling data bias initially is already a daunting task, trying to maintain the integrity of my data by creating a regular re-training pattern creates the potential for even more bias seeping into my neural network even higher. In order to combat this I would need to work with a team that is actively collecting, cleaning and testing the data for bias and then training the machine on said data. Considering how entities much larger than myself like OpenAI took three years to update their model by one, I would say this task is herculean to say the least. Were I to try to single-handedly retrain my model, I have no doubt that it would lead to hallucinations and the spread of misinformation about Occidental College. Frequently updating the model could also result in increased susceptibility to API Injections, which would further spread misinformation by making the model misrepresent itself, its creators as well as Occidental College. If it was possible to safely update a large language model's data so frequently, I have no doubt that OpenAI or other large language model companies would have already done it.

VI. Conclusion

In writing this paper I have determined that it is simply not possible to create a large language model without having to make serious moral compromises. When it comes to training a language model, it is impossible to navigate around all of the bias that is inherent to the data that the machine is trained on. Given the nature of humans and our inherent biases, all data is skewed in some aspect and that makes it very difficult to create a model that has little to no bias. With the current nature of how language models work, it will also be extremely difficult to make them more transparent, as two of the same neural networks trained on the same datasets can come out with varying results and attempting to track inputs could lead to analyzing thousands if not millions of tokens a day. In terms of privacy, the best models were illegally trained on real life data, meaning that creating a more ethical model would not only take more time, but also produce less accurate results, as it was trained on imitations rather than actual raw data. Lastly, trying to maintain an up-to-date language model is not only incredibly resource intensive but also leaves the system vulnerable to the spread of misinformation either through hallucination or API injection. Creating a language model is something

that cannot achieve results without crossing grave ethical boundaries.