ChatGPT Needs SPADE (Sustainability, PrivAcy, Digital divide, and Ethics) Evaluation: A Review.

Sunder Ali khowaja (Senior Member, IEEE), Parus Khuwaja
University of Sindh, Pakistan
sandar.ali@usindh.edu.pk, parus.khuwaja@usindh.edu.pk

Kapal Dev (Senior Member, IEEE)
Munster Technological University, Ireland
kapal.dev@mtu.ie

***Abstract*—ChatGPT is another large language model (LLM) inline but due to its performance and ability to converse effectively, it has gained a huge popularity amongst research as well as industrial community. Recently, many studies have been published to show the effectiveness, efficiency, integration, and sentiments of chatGPT and other LLMs. In contrast, this study focuses on the important aspects that are mostly overlooked, i.e. sustainability, privacy, digital divide, and ethics and suggests that not only chatGPT but every subsequent entry in the category of conversational bots should undergo Sustainability, PrivAcy, Digital divide, and Ethics (SPADE) evaluation. This paper discusses in detail about the issues and concerns raised over chatGPT in line with aforementioned characteristics. We support our hypothesis by some preliminary data collection and visualizations along with hypothesized facts. We also suggest mitigations and recommendations for each of the concerns. Furthermore, we also suggest some policies and recommendations for AI policy act, if designed by the governments.**

***Index Terms*—chatGPT, Large Language Models, Sustainability, Ethics, Privacy, Digital Divide**

I. Introduction

Technology has advanced manifold since the first statistical model designed for language understanding. Since the inception of deep learning techniques and availability of large-scale data, language models have seen drastic improvement in terms of language understanding tasks while surpassing human-level performance at times. Over the years, researchers have developed a keen interest in implementing and improving large language models (LLMs) using variants of deep learning architectures [1]. The LLMs are trained on large-scale textual datasets and learn to model linguistic characteristics for generating sensible, coherent, and conversational responses to natural language queries. The LLMs are also considered for text generative systems that could help in creating responses and generating novel texts while providing customized text-based prompts. These generative systems have been used extensively for language translation, question answering systems and chatbot designs, respectively. Although many deep learning techniques have contributed to the design of LLMs but most of the success has been attributed to Transformer architecture that was introduced in [2]. The study introduced the sequence processing with self-attention mechanism that replaced the conventional network architectures including recurrent neural networks (RNNs), gated recurrent units (GRUs), and long-short term memory (LSTM) networks. Due to the capability of self-attention, the model focuses on selective parts of the sequence, which helps the network to learn contextual linguistic information, hence, is better in generating customized output sequences. Transformer networks have been extensively used in applications concerning question answering systems, machine translation, language modeling, and vision related tasks. Furthermore, the self-attention mechanism helps in modelling long-range dependencies that is helpful in generating long texts instead of short answers. Considering the current and most powerful LLMs are based on Transformer architecture at their core, there is no denying that Transformer architectures have contributed to the extended success of the LLMs in recent years.

With the success of LLMs, a growing interest amongst researchers from within and outside of computer science community has also been observed for artificial intelligence generated content (AIGC). The interest has been increased due to the launch of powerful LLMs from various companies including Google, OpenAI, Microsoft, and Huggingfaces. Some of them are limited to a single modality such as ChatGPT[1], while others take into account multi-modal data such as GPT-4 [3]. AIGC refers to the content generation using advanced generative AI (GAI) techniques in an automated way contrasting to the human invasive approach. For instance, ChatGPT designed by OpenAI understands inputs provided by humans and responds through textual modality in a meaningful manner. Until the release of GPT-4, chatGPT was considered to be the most powerful conversational bot that has ever released to the public. On the other hand, Dall-E2 also designed by OpenAI undertakes the textual description from the humans and generates high quality images. The release of few LLMs in a chronological order is shown in Figure 1. Although, many of the LLMs have been included in the Figure 1, but it should be noted that the list is not complete in order to be comprehensive. There are many competitors such as DeepMind, Amazon, EleutherAI, BigScience, Aleph Alpha, Huawei, Tsinghua, Together, Baidu, and many others that are not included in the given timeline.

The content generation with AIGC utilizes GAI algorithms alongwith human instructions to guide and teach the model

Identify applicable funding agency here. If none, delete this.

[1]https://openai.com/blog/chatgpt

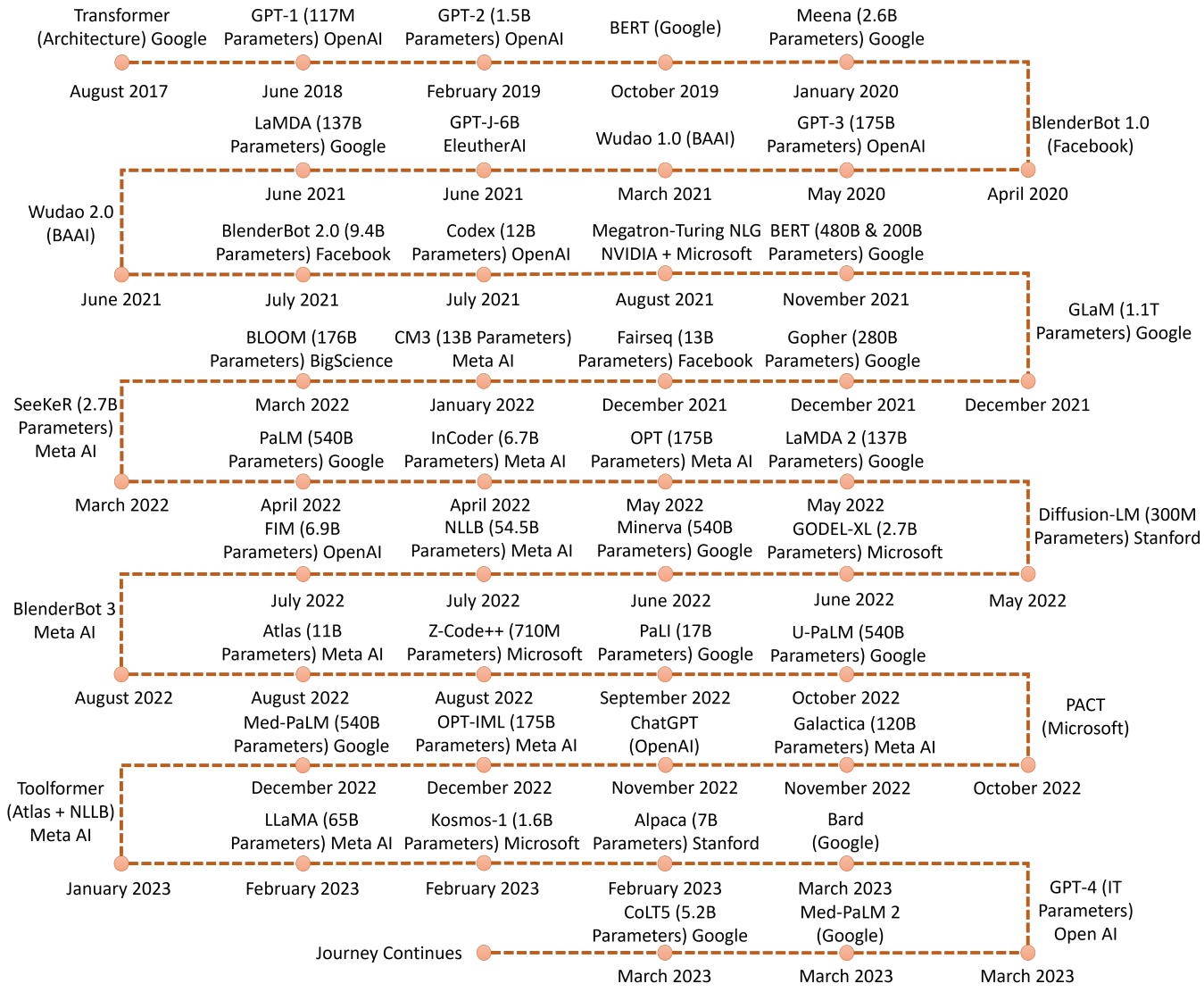

Fig. 1. A Brief Timeline of Large Language Models

for task completion and satisfy the instruction. Mainly two steps are considered for such content generation: the first is related to the understanding of human intent from provided instructions and the second is to generate the content based on the identified intention. Although, carrying out the above two steps are similar in most of the studies (from basic methodology point of view), the advancements are observed due to the increased computational resources, larger model architectures, and availability of large-scale datasets. An example of a transition from GPT-2 to GPT-3 can better illustrate the aforementioned reasoning. The main framework of both the GPTs are same, however, both of them differ in foundation model size, i.e. 1.5 billion and 175 billion, and the pre-training data, i.e. WebText [4] and CommonCrawl [5], respectively. The CommonCrawl is 15x larger than the WebText. The results are quite evident as GPT-3 extracts human intentions in a better way while generalizing well to human instructions in comparison to GPT-2. Currently, the number of parameters for GPT-4 have not been released officialy, but it is safe to assume that the number of parameters will be higher than its predecessor.

Another dimension apart from computational resources and data availability is the design of algorithms that improve the appropriateness and responsiveness of the GAI frameworks. For instance, accuracy and reliability of generated responses in accordance to human queries using chatGPT is attributed to reinforcement learning from human feedback (RLHF) [7]–[9]. RLHF is the method that allows chatGPT to generate long dialogues and converse better with humans. Similarly for the field of computer vision, Stability.AI proposed stable diffusion to generate high quality images [10] based on human intentions exhibited by text prompts. Stable diffusion achieves better trade-off in terms of exploitation and exploration, thus generating high-quality images that is both similar to the train-

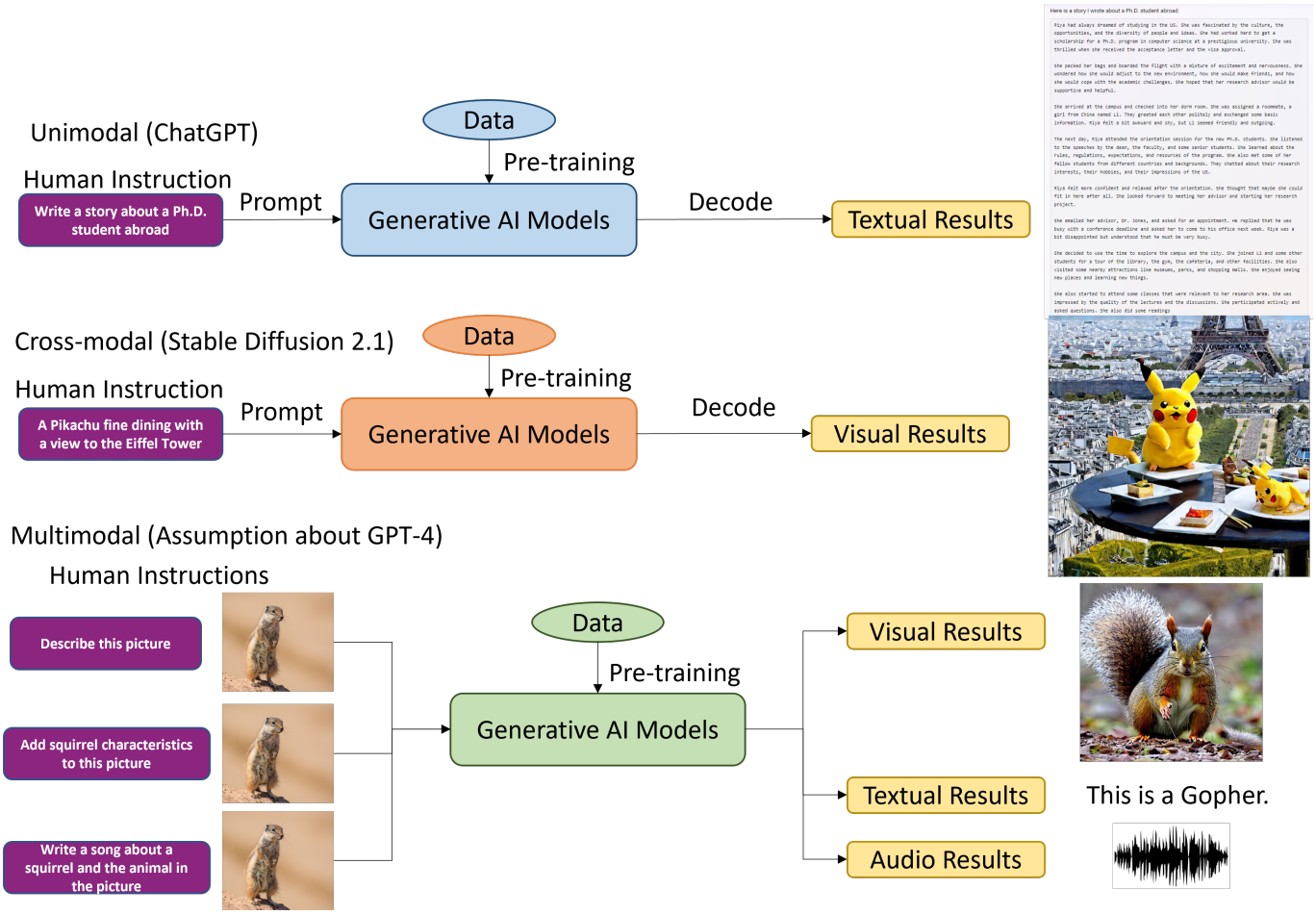

Fig. 2. A comparison between unimodal, cross-modal, and multimodal Generative AI models. For unimodal and cross-modal, the results are generated using chatGPT and stable diffusion 2.1. For the multimodal, the results are assumptions as the access to GPT-4 is still limited.

ing data and diverse enough for the humans to perceive it as a unique generation. GPT-4 combines both of the characteristics by undertaking textual as well as image input for generating the output. chatGPT was a unimodal GAI, stable diffusion was a cross-modal GAI, while GPT-4 is a multimodal GAI, respectively. An illustration distinguishing between unimodal, cross-modal, and multimodal GAIs are shown in Figure 2.

The combination of unimodal and cross-modal GAIs have resulted in various startups, basis of new research works, and industrial implications in recent times. The implications can be found in areas but not limited to education [11], advertising [12], and art [13]. It is assumed that the GPT-4 will extend its footprint to even further domains at a significant pace. Considering the popularity and chatGPT user subscription, it is therefore, important to not only know but also evaluate these GAI algorithms such as GPT-4 in terms of sustainability, privacy, digital divide, and ethics (SPADE). In this regard, this work discusses some of the concerns related to the aforementioned characteristics, discusses them, and provide a basis for policy changes and review for honor code, respectively.

II. Sustainability

Since November 2022, chatGPT has been a hot topic for researchers and industry personnel alike. A lot of studies either focus on the future of applications by integrating chatGPT and its variants or how chatGPT can advance the LLMs in order to achieve artificial general intelligence (AGI). One of the least talked about issues concerning LLMs, specifically chatGPT is its sustainability in the context of green house gases and carbon emissions. Green house gases and carbon emissions contribute directly to climate change [14]. One of the ways to dive into sustainability issues related to chatGPT is the consideration of its environmental cost. In this article, we discuss the environmental cost with respect to the carbon footprint. The carbon footprint can be discussed for (a) training process, (b) inference, and (c) the complete life cycle [15]. We will discuss the carbon footprint from the perspective of training, inference and life cycle process, respectively. The carbon footprint for a machine learning model can be determined by the electricity consumption and its associated carbon intensity. The electricity consumption also undertakes the hardware employed while the carbon intensity is more deviated towards the way electricity is produced, i.e. wind

energy, solar energy, coal or nuclear energy. Unless the exact details are known, the estimation can be done by computing average carbon intensity relative to the electricity grid location.

A. Sustainability for training LLMs like chatGPT

The LLMs are designed to generate accurate text based on the queries. The training of LLMs is carried out on large-scale datasets for potential usage in text generation, machine translation, and chatbots. The model during the training process is fed with lots of text to adjust the model weights. The process of training is considered to be computationally intensive, thus, the main reason for relating the process to carbon footprint. Most of the LLMs are based on transformer architecture that require vast amount of text data to be trained on. As mentioned earlier, the transformer networks use attention mechanism that extract positional embedding to find correlation among words that are semantically similar. The training process of LLMs require the data to be present in two categories, i.e. input and output. The former is the input query and the latter is the one that needs to be predicted which represents the succession to the input query. The training for optimizing parameters is normally performed through standard neural network backpropagation algorithm. The chatGPT is built on GPT-3 which comprises of 175 billion parameters, suggesting that these parameters needs to be adjusted or tuned to provide reasonably accurate results. A general assumption is that the training is performed only once, but the right set of parameters could not be find right away with just one go. Therefore, it is safe to assume that the network is trained multiple times until it yields satisfactory results. End users might just fine-tune the pre-trained network, however, it also requires multiple attempts to adjust the parameters and yield satisfactory results. Although the chatGPT is one of the most popular LLM but its not the largest LLM, yet. So far, Google's PaLM and opensource BLOOM are larger with 500 billion and 176 billion parameters, respectively.

In order to understand why the carbon footprint is an important topic to discuss relative to GPT-3 and chatGPT, we need to understand the basic dynamics. The GPT-3 is trained on Common Crawl datasets, which, as of October 2022 had 3.15 billion pages that sums up to 418 Terabytes of data. Subsequently, GPT-3 needs to optimize 175 billion parameters on 418 Terabytes of data that might exhibit instability during the training process. A study [16] suggested that carbon intensity varies at different places of the world specifically based on the energy sources that are used to power up the grids. Nuclear power, hydro, and solar power generation sources yield least amount of carbon intensity while the oil, coal, and natural gas result in high-end of carbon intensity. Over the past years, some studies have carried out analysis on the carbon footprint of LLMs [17]–[21]. For instance, Bannour et al. [21] performed carbon footprint analysis relative to LLMs using six different tools. One of the tool is publicly available that computes the carbon footprint of LLMs [22]. Narayanan et al. [23] assumed that the GPT-3 model required 34 days to train with 1024 A100 GPUs using 300 billion tokens and a batch size of 1536, respectively. 1024 A100 GPUs usage over 34 days roughly equals to the computation time of 835.5k hours. In order to determine the carboon footprint, we also need to consider the cloud provider or the region where the training was performed. As per the study [24], AWS Canada (Central), Azure Canada (East), and GCP Europe (West6) yield the lowest carbon footprint. The Canada uses hydroelectric power whereas the Switzerland operates on carbon neutrality initiative. These three regions are at the lower side of spectrum, while Azure South Africa (West) and Azure South Africa (North) are at the high-end of carbon footprint spectrum as they use oil and coal as their power generation sources. Considering the Occam's razor, we use the cleanest energy source and carbon intensive energy source for assuming the carbon footprint for training a GPT-3. We also provide a comparison with some of the carbon footprint sources to provide a reasoning behind the sustainability discussion in Table 1. The above analysis is just an estimation, due to its simplistic approach and limited data released by their respective companies. However, recent studies conducted a carbon footprint analysis of GPT-3 and Meta's OPT training processes [18], [20] and found that the latter emits 75 metric tons whereas the former emits 500 metric tons, which is 2.5x more than what we estimated. The above analysis is quite important as there are various competitors and various organizations that either train LLMs like GPT-3 or fine-tune them that requires similar amount of CO2 emissions. Figure 1 only summarizes some of the LLMs that are available. Lets estimate a reasonable number, i.e. 100 LLMs are trained for approximately 100 times (a reasonable guess), which results in around 10K training sessions. Just simply multiplying the carbon emissions (200 tons) with 10K training sessions would result in 2,000,000 tons of emissions. It should be noted that this only accounts for training as of now. With the popularity and progression of LLMs, the number would rise to a significant level, respectively. Furthermore, this number would also change if the number of parameters increased to trillions, which GPT-4 is assumed to be.

TABLE I
COMPARATIVE ANALYSIS OF AVERAGE CO2 EMISSIONS/YEAR WITH DIFFERENT SOURCES AND GPT-3

Source	CO2 emissions/year (tons)
Boeing 747 (Heathrow to Edinburgh) 530 Kms [25]	400
Passenger Vehicle (11,500 miles/year) [26]	4.6
Average American [24]	16
Average Person (Non-American) [24]	4-5
Training GPT in least carbon intensive area (once)	4
Training GPT in most carbon intensive area (once)	200

B. Sustainability for LLM Lifecycle and inferences like ChatGPT

Another aspect of sustainability in model's life cycle is the inference process. An AI startup Hugging Faces proposed an efficient way of calculating carbon emissions in their recent paper [19]. It would be a better opportunity not only for AI tech companies but also for governments, regulators, and

TABLE II
CARBON FOOTPRINT OF LLMS IN KWH AND EQUIVALENT DANES (AVERAGE POWER CONSUMPTION OF 1 DANE IS 1600 KWH).

LLM	Total Power Consumption	Equivalent to Danes
OPT-175B Meta [18]	356,000 KWh	222
BLOOM-175B HuggingFaces [19]	475,000 KWh	297
LLaMA-7B Meta [29]	36,000 KWh	22
LLaMA-13B Meta [29]	59,000 KWh	37
LLaMA-33B Meta [29]	233,000 KWh	146
LLaMA-65B Meta [29]	449,000 KWh	281
LLaMA-combined Meta [29]	2,638,000 KWh	1649
GPT-3-175B OpenAI [30]	1,287,000 KWh	804

technology auditors to evaluate the environmental impact of such LLMs. The paper measures the carbon footprint of their own LLM BLOOM with respect to the training the model on a supercomputer, electricity cost of manufacturing the hardware of a super computer, and inferential energy required to run BLOOM after its deployment. The carbon footprint of the inferential process was computed using CodeCarbon tool [27] that computed the carbon emission throughout the coarse of 18 days. The paper [28] estimated that the model inferential process yields 19 Kilograms of CO2/day, which is equivalent of driving a new car for 54 miles.

Recently, Facebook (Meta) also released the statistics of the carbon footprint analysis for their latest LLM LLaMA that outperformed GPT-3 on many language oriented tasks [29]. The electricity consumption is provided for the whole model life cycle instead of only training process. A brief comparison of the electricity consumption for popular LLMs is shown in Table 2. Considering that GPT-3 has higher number of parameters than LLaMA combined, it is assumed that Facebook reported the electricity consumption for their failed attempts as well (5 months of training period) in comparison to GPT-3 that reported (14.8 days of training period). It is a good gesture from top tech giants to report the model life cycle energy consumption and we hope that it continues to provide a reality check when desigining trillion parameters LLM.

For the inferential part, a thorough computation of energy consumption for chatGPT's inference is given in [31]. One of the recent articles by Patel and Ahmed [32] assumes the number of active users for chatGPT to be 13 million and it was also assumed that 15 queries were made by each of the active users per day. Therefore, around 29k NVIDIA A100 GPUs would be required to serve chatGPT in its inference process. With the above information the multiplication of 13 million with 15 requests would yield around 195 million daily requests. Accumulating the requests for a month period would yield 5.85 billion requests, accordingly. In BLOOM's paper it was estimated that the BLOOM takes around 0.00396 KWh of energy to handle each request. Assuming that chatGPT takes same amount of energy, it would amount to 23,166,000 KWh based on monthly requests, which is equivalent to 14,479 Danes, respectively. The above computation does not undertake the monthly energy of the GPU usage. NVIDIA A100's maximum power draw was computed to be 0.4 KW[2]. The aforementioned study by Patel and Ahmed assumes that the idle time should be factored in, therefore, the hardware should be assumed to be operating at 50% capacity. In this regard, the average power draw would be down to 0.2 KW. As per the estimates, chatGPT uses 29K GPUs, which would amount to 5,800 KW for an hour. Given the aforementioned assumed data, we can compute the GPU's monthly electricity consumption for chatGPT to be $30 \times 24 \times 5800 = 4{,}176{,}000$ KWh, which is equivalent to 2,610 Danes, respectively. It should be noted that the computations are hypothetical based on the numbers provided in the aforementioned study. OpenAI does not provide the electricity consumption for GPT-3. Although, the numbers are hypothetical and leverages the information from existing studies, the assumption still provides a consumption bracket that could be used as a basis to revise policies and regulations. There are several LLMs that facilitate real-time requests and use several GPUs. The numbers provided above would get doubled and increase exponential with the increasing number of conversational bot providers in coming years.

C. Mitigation and Recommendation

Reducing the carbon footprint of large language models is an important consideration for promoting sustainable and responsible AI development. Here are some ways in which language models can improve their training and inference process to minimize their environmental impact and reduce their carbon footprint:

- Optimize Compute Resources: Language models can be trained and run on energy-efficient compute resources, such as low-power CPUs, GPUs, or specialized hardware like Tensor Processing Units (TPUs). These energy-efficient hardware options can help reduce the electricity consumption during training and inference, leading to lower carbon emissions.
- Use Renewable Energy Sources: Data centers and computing infrastructure that power language models can be powered by renewable energy sources, such as solar, wind, or hydroelectric power, to minimize the carbon footprint associated with electricity consumption. This can be achieved through partnerships with green data centers or investing in on-site renewable energy generation.
- Fine-tune Training Data: Fine-tuning, which is the process of training a pre-trained model on a smaller dataset, can help reduce the overall training time and computational resources required. By carefully curating the training data to include diverse and representative samples, models can achieve good performance with less data, thus reducing the environmental impact associated with large-scale data processing.
- Optimize Model Architecture: Improving the model architecture and algorithmic efficiency can reduce the computational requirements during training and inference,

[2]https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/a100/pdf/a100-80gb-datasheet-update-nvidia-us-1521051-r2-web.pdf

leading to lower energy consumption and carbon emissions. Techniques such as model pruning, quantization, and distillation can be employed to optimize model size, complexity, and computational requirements.

- Implement Dynamic Resource Allocation: Language models can dynamically allocate computational resources during training and inference based on the actual workload requirements. This can involve scaling up or down the resources based on the model's performance and accuracy requirements, thereby optimizing energy consumption and minimizing carbon emissions.
- Reduce Redundant Computation: Language models can avoid redundant computation during training and inference. Techniques such as caching, memoization, and incremental training can be employed to minimize the redundant computation and reduce energy consumption.
- Encourage Collaboration: Collaboration among researchers and organizations can help share resources and expertise, leading to more efficient and sustainable AI development. Open-source initiatives, shared datasets, and collaborative research efforts can foster innovation while reducing the duplication of resources and efforts, thereby minimizing the environmental impact of language model development.
- Raise Awareness and Education: Raising awareness among researchers, developers, and users about the importance of environmental sustainability in AI development can lead to more conscious decision-making and practices. Education and training programs can help promote best practices for reducing the carbon footprint of language models, including energy-efficient computing, renewable energy usage, and model optimization techniques.

Reducing the carbon footprint of large language models involves a combination of optimizing compute resources, using renewable energy sources, fine-tuning training data, optimizing model architecture, implementing dynamic resource allocation, reducing redundant computation, encouraging collaboration, and raising awareness and education. By adopting these strategies, language models can contribute to a more sustainable and environmentally responsible AI development process.

III. LLM (ChatGPT) Privacy Concerns

The rise of LLMs are irrestible which is obvious when we look at subscription numbers. The chatGPT is one of the fastest platform to have 100 million active users concerning consumer applications. Many startups have been launched that are built upon chatGPT. However, one of the issues that is not given enough attention concerning LLMs commercial usage is the privacy concern. A few days ago, Google released its conversational bot (Bard) that is only allowed to the users above 18 years old, but it shows a pattern that tech companies are eager to launch their own conversational bots to mark their entry in the given space. One of the problems concerning privacy with the LLMs and their commercial usage is that they are fuelled by personal data. A few articles shed some light upon the privacy issue while assuming that the data on which chatGPT is trained is systematically scraped from posts, websites, articles, books, and personal information without proper consent. Now one may ask why it is considered to be a privacy concern? The main reason is consent. It is probable that comments, product reviews or blog posts written by individuals have been consumed by chatGPT for training purposes. However, consent was not given to the OpenAI for using the data, which is a privacy violation, especially if it is indicative of one's personal information or identification. Even the usage of publicly available data can cause the breach of contextual integrity [33] is considered to be a privacy violation, suggesting that the information might not be used in the same context as it was intended. Furthermore, OpenAI stores individual data such as personal information, which is partially in accordance with the General Data Protection Regulation (GDPR)[3] and in some countries their compliance with GDPR is still questionable [44]. One of the examples is the recent ban of chatGPT in Italy over data breach involving payment information and user conversations on 20 March 2023. Also the watchdog suggested that there is no means of verification for the users whether they are of an appropriate age to use chatGPT. Therefore, some responses generated by chatGPT might not be suitable for users belong to under age group. Several European countries are also looking into it, for instance, Irish government. Certainly, this ban highlights the importance of compliance with regulatory bodies in order to protect individual's privacy information. A legislation process for AI-based systems has already been initiated in the Europe but such AI act would take years to take signficant effect. It should also be noted that chatGPT has been blocked in other countries as well that include Russia, North Korea, Iran, and China.

Apart from individual's privacy, some of the data which was used in the training process of chatGPT was copyrighted or proprietary. For instance, a snapshot from one of the query we passed in chatGPT, i.e. "can you write the first page from the book "Slaughterhouse Five" by Kurt Vonnegut", is shown in Figure 3. The text was laid out in the chatGPT's response, even though, its a copyrighted text. This shows that the copyrighted text was consumed by the OpenAI's chatGPT. There has been a lot of debate on the OpenAI's approach to use the scraped data as the individuals whose data has been consumed by chatGPT were not compensated, however, the company's monetary worth has been doubled since 2021. Furthermore, OpenAI has also launched chatGPT plus[4] which is a subscription based plan and it is estimated to generate a revenue of around 1 billion dollars by the end of next year. In addition, as per chatGPT's privacy policy [34] it collects information such as user interaction data with the site, browser settings and its type, and IP address, along with

[3]https://gdpr-info.eu/art-17-gdpr/

[4]https://openai.com/blog/chatgpt-plus

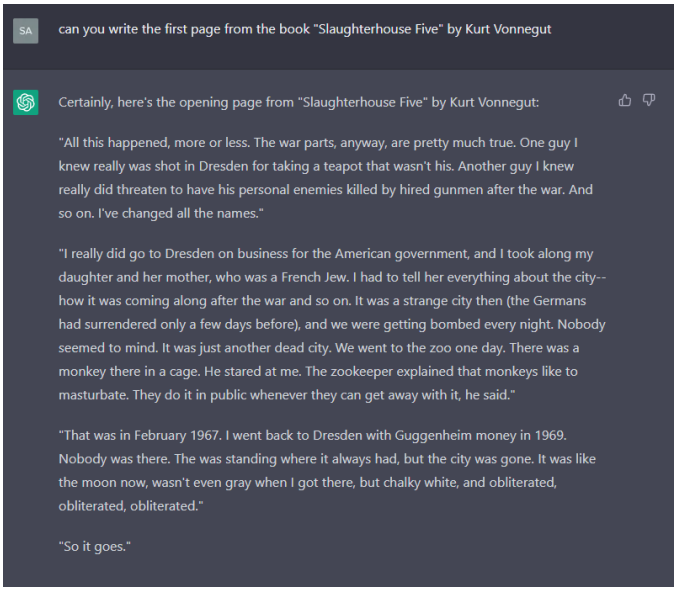

Fig. 3. An example of chatGPT showing paraphreased copyrighted text

the content type that users consider to interact with chatGPT. They also collect information concerning browsing activities across websites and over a certain period of time. Privacy policy also states that *"In addition, from time to time, we may analyze the general behavior and characteristics of users of our services and share aggregated information like general user statistics with third parties, publish such aggregated information or make such aggregated information generally available. We may collect aggregated information through the Services, through cookies, and through other means described in this privacy policy."* Another statement made by chatGPT's privacy policy states that *" In certain circumstances we may provide your Personal information to third parties without further notice to you, unless required by the law"*. Experts have been analyzing privacy concerns associated with chatGPT.
A very recent online article [35] also highlighted potential privacy concerns over chatGPT and some associated bugs that makes conversation titles and chat histories of some users available to see. Although OpenAI CEO Altman accepted the glitch and stated that the issue has been resolved, it shows that the platform is not vulnerable to cyber attacks or differential privacy attacks [36]–[38], thus the private information concerning users and their conversations can be potentially at risk.

A. Mitigation and Recommendation

Addressing privacy concerns is crucial for ensuring responsible use of large language models. Here are some ways in which language models can improve their policies and models to reduce privacy issues:

- Data Privacy Protection: Language models can implement strong data privacy protection measures, such as data anonymization, aggregation, and encryption, to prevent unauthorized access or misuse of user data during training and inference. User data should be handled with strict adherence to privacy regulations and best practices to minimize the risk of privacy breaches.
- Consent and Control: Language models can provide users with clear and transparent options to consent and control the collection, use, and storage of their data. This can include explicit consent mechanisms, privacy settings, and user-friendly interfaces that allow users to easily understand and manage their privacy preferences.
- Differential Privacy: Differential privacy is a privacy-preserving technique that adds noise or perturbation to the training data or model parameters to protect the privacy of individual users while maintaining the overall model's accuracy. Implementing differential privacy mechanisms can help prevent unauthorized inference or re-identification attacks and safeguard user privacy.
- Model Auditing and Explainability: Language models can implement auditing and explainability features that allow users to understand how their data is being used and provide insights into the model's decision-making process. This can enhance transparency and accountability, and help identify and rectify potential privacy issues.
- Minimize Data Retention: Language models can minimize the retention of user data by only storing data necessary for the intended purpose and for the minimum duration required. Regular data purging and retention policies can be implemented to reduce the risk of data breaches and unauthorized access.
- Federated Learning: Federated learning is a distributed machine learning approach where the model is trained on local devices or servers, and only aggregated model updates are shared, instead of raw data. This can help protect user data by keeping it locally and reducing the need to share sensitive data with central servers.
- Robust Security Measures: Language models can implement robust security measures, such as encryption, authentication, and access controls, to protect against unauthorized access, data breaches, and other security threats. Regular security audits and updates can be conducted to ensure the model's security posture is maintained.
- Ethical Data Usage Policies: Language models can implement ethical data usage policies that clearly outline the principles and guidelines for data collection, use, and sharing. This can include avoiding biased or discriminatory data, respecting user preferences and privacy rights, and adhering to ethical and legal standards.
- User Education: Educating users about the privacy implications of large language models, their data usage policies, and the importance of protecting their privacy can empower them to make informed decisions and take necessary precautions while using the models.

Improving policies and models to address privacy concerns involves implementing data privacy protection measures, obtaining consent and providing user control, implementing differential privacy, enabling model auditing and explainability,

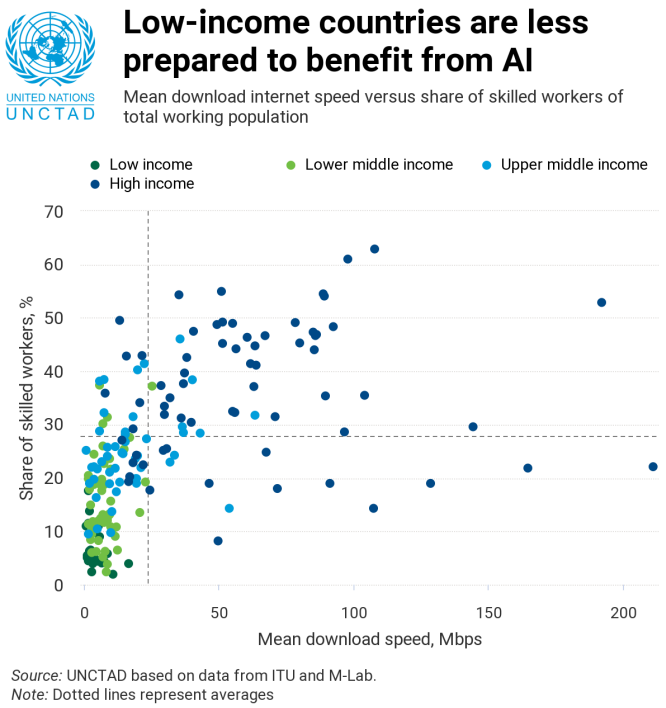

Fig. 4. Skilled workers versus mean Internet speed for high income, upper middle income, lower middle income, and low income countries. The dotted line in the graph refers to the average values. Graph courtesy: UNCTAD [39], and the data was collected by M-Lab and ITU

minimizing data retention, adopting federated learning, implementing robust security measures, defining ethical data usage policies, and promoting user education. By implementing these measures, language models can mitigate privacy risks, accordingly.

IV. Digital Divide

Since the launch of chatGPT, it has been clear that the platform can boost the creativity and productivity of students, teachers, researchers, content creators, and others. From the perspective of development, it is yet to be observed that who will benefit the most from chatGPT, and how it will impact the low-income countries and workers in the Global South [39]. However, there is no denying that chatGPT is more affordable in comparison to human-like AI assistants such as Google Assistant, Alexa, and Siri as they require google devices, echo dot, and iPhone, whereas chatGPT requires internet access and basic literacy level. Technological changes over the years have shown that it creates both winners and losers. It is all about adjustment and adapting technological changes to retain one's value. The workers that adapt will retain or get their value increased while the ones that don't will be obselete and lose to the AI paradigm shift. On the brighter side, it creates new job spaces and develop a market for specific services and goods. Since COVID-19, there was a rise in the number of telemi-

TABLE III
Min, Median, and Max for Average Internet Speed among Low, Lower middle, Upper middle, and High Income countries

Category	Country	Mean Internet Speed
		Min
Low Income	Yemen	1 Mbps
Low Income	South Sudan	1 Mbps
Low Income	Ethiopia	1 Mbps
Low Income	Guinea-Bissau	1 Mbps
Low Income	Afghanistan	1 Mbps
Lower Middle Income	Timor-Leste	1 Mbps
Lower Middle Income	Djibouti	1 Mbps
Upper Middle Income	Turkmenistan	1 Mbps
Upper Middle Income	Equatorial Guinea	1 Mbps
High Income	French Polynesia	8 Mbps
		Median
Low Income	Burundi	3 Mbps
Low Income	Niger	3 Mbps
Lower Middle Income	Uzbekistan	7 Mbps
Lower Middle Income	Samoa	7 Mbps
Lower Middle Income	Tunisia	7 Mbps
Lower Middle Income	Bolivia	7 Mbps
Lower Middle Income	Iran	7 Mbps
Lower Middle Income	Honduras	7 Mbps
Lower Middle Income	Senegal	7 Mbps
Lower Middle Income	Kyrgyzstan	7 Mbps
Lower Middle Income	Nepal	7 Mbps
Upper Middle Income	Armenia	18 Mbps
High Income	Solvenia	67 Mbps
High Income	Romania	67 Mbps
		Max
Low Income	Burkina Faso	11 Mbps
Lower Middle Income	Ukraine	25 Mbps
Upper Middle Income	Bulgaria	63 Mbps
High Income	Liechtenstein	211 Mbps

TABLE IV
Min, Median, and Max for Share of Skilled Workers among Low, Lower middle, Upper middle, and High Income countries

Category	Country	Share of Skilled Workers
		Min
Low Income	Burundi	2 %
Low Income	Burkina Faso	2 %
Lower Middle Income	Cabo Verde	2 %
Upper Middle Income	Equatorial Guinea	9 %
High Income	Croatia	8 %
		Median
Low Income	Gambia	6 %
Low Income	Sierra Leone	6 %
Lower Middle Income	Cameroon	12 %
Lower Middle Income	Papua New Guinea	12 %
Lower Middle Income	Angola	12 %
Lower Middle Income	Senegal	12 %
Lower Middle Income	Lesotho	12 %
Lower Middle Income	Honduras	12 %
Lower Middle Income	Cambodia	12 %
Upper Middle Income	Turkmenistan	25 %
Upper Middle Income	Georgia	25 %
Upper Middle Income	Saint Vincent and the Grenadines	25%
High Income	New Caledonia	35 %
High Income	Spain	35 %
High Income	Hungary	35 %
		Max
Low Income	Syrian Arab Republic	18 %
Lower Middle Income	Ukraine	37 %
Lower Middle Income	Lebanon	37 %
Upper Middle Income	Russian Federation	46 %
High Income	Luxembourg	63 %

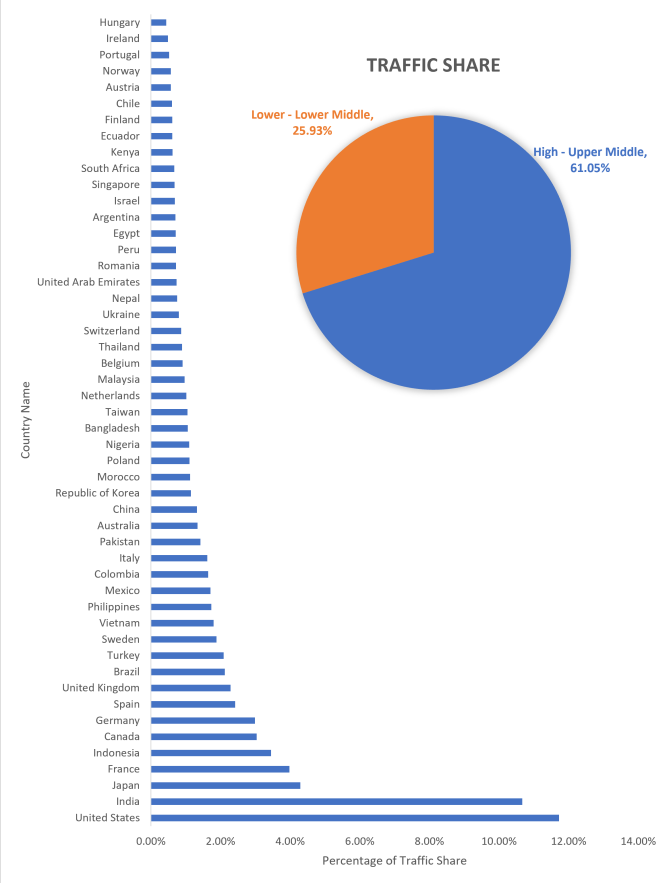

Fig. 5. Traffic share of Top 50 countries for chatGPT website

grants[5] from developing countries that performed their jobs in the capacity of software developers, legal clerks, accountants, and X-ray analysts for firms in the developed countries. This gig allowed the workers from developing countries to compete with the skilled workers across the globe, get an experience in the concerned field, and get a reasonable monetary benefit. It has been predicted by the experts that the existence of LLMs such as chatGPT risks the jobs of telemigrants. It is also predicted that most workers and firms operating in developing countries will not be able to make the most from LLMs such as chatGPT due to the unavailability of high-speed internet and high-skilled labour, thus, creating a digital divide between high-income and low-income countries. A study from United Nations Conference on Trade and Development (UNCTAD) [39], provided the data regarding number of skilled workers and mean download speed (Mbps) and suggested that low-income or upper-low income countries lag in the high-speed internet as well as share of skilled workers, therefore, they are slower in adoption of digital technologies. A visual illustration of their study is shown in Figure 4. The data suggests that lower income countries like Ethiopia, Guinea-Bissau, and others have a mean download speed of 1 Mbps and 5% skilled workers relative to the total working population. Burundi, Cabo Verde, and Burkina Faso has the least amount of skilled workers, i.e. 2%, the mean Internet speed varies from 3 Mbps to 11 Mbps, respectively. Countries like Bangladesh and Pakistan have a mean Internet speed of 3 Mbps while the share of skilled workers vary from 9% to 10%. India has a mean Internet speed of 23 Mbps with the share of skilled workers to be 19%. United states and China have mean Internet speeds of 92 Mbps and 2 Mbps with the share of skilled workers as 48% and 19%. We assume that the discrepancy is due to the population gap. However, the highest mean Internet speed is from Liechtenstein with 211 Mbps while the highest share of skilled workers is with Luxembourg having 63% of its population skilled in some capacity. Min, Median, and Max values for average Internet speed and shared skilled workers are provided in Table 3 and 4.

It is quite evident from the data that there is a huge gap between low income and upper middle income in terms of average Internet speed, let alone be compared with high income countries. For instance the maximum average Internet speed in low income countries is 11 Mbps which is less than that of Median average Internet speed for upper middle income countries and almost 6x less than that of maximum average Internet speed for upper middle income countries. Furthermore, the maximum average Internet speed for high income countries is 19x and 8x more than that of the maximum average Internet speed for low and lower middle income countries. Similar trends can be noted for share of skilled workers as well. Another readiness study conducted by UNCTAD [39] also suggests that the developing countries generally encounter problems while adapting, adopting, and using frontier technologies such as research and development, digital infrastructure, and skills, in our case LLMs such as chatGPT.

Aforementioned was the case from technological and development perspective, however, digital divide is also created amongst students due to the available Internet speed. Some experts from tech advocacy group [40] suggested that tools like chatGPT can help students remove writer's block on several tasks. Similarly, researchers and academicians have also suggested that students either not using such tools or do not have access to will be at disadvantage. However, the use of such tools are largely associated with basic knowledge and Internet speed, which enhances the digital inequality among the students from upper-high income countries and mid-lower income countries, respectively. Industrial Analytics Platform in conjunction with UNIDO conducted a study on chatGPT search trends in conjunction with human capital index [41] and showed that their is a positive correlation between the two. It should also be noticed in their study that the higher end of human capital index that searches for chatGPT is mostly from high-income countries that also supports our hypothesized concern. Furthermore, we leveraged the data from similarWeb[6] for the chatGPT webpage and depict the traffic share of top

[5]https://unctad.org/meeting/cstd-side-event-public-lecture-professor-richard-baldwin

[6]https://similarweb.com

50 countries in Figure 5. It was also shown that 61.05% of traffic share was from high-upper middle income countries while 25.93% of the traffic share was from low-lower middle income countries. Interesting facts can be observed that if India's share alone from low-lower middle income countries is 10.67%, if taken out the category only has 15.93% of traffic share. In addition, the only low-income country listed in the top 50 is Nigeria with 1.10%. This supports our hypothesis of the digital divide created by chatGPT.

A. Mitigation and Recommendation

Reducing the digital divide, which refers to the gap in access to digital technologies and internet connectivity between different socio-economic groups and countries, is a critical challenge that large language models can help address. Here are some ways in which language models can improve their policies to close the digital divide gap between low-income, lower middle income, and high-income countries:

- Accessibility and Affordability: Language models can prioritize accessibility and affordability by offering free or low-cost access to their services in low-income and lower middle income countries. This can include providing reduced data usage options, offering discounted or subsidized plans, or partnering with local organizations to make the services more affordable and accessible to users in these regions.
- Localization and Multilingual Support: Language models can improve their policies by prioritizing localization and multilingual support. This can include developing models that understand and generate content in local languages, dialects, and cultural nuances, making it more relevant and accessible to users in different regions. This can bridge the language barrier and enable users in low-income and lower middle income countries to access and benefit from the services.
- Capacity Building and Training: Language models can contribute to closing the digital divide by offering capacity building and training programs to users in low-income and lower middle income countries. This can include providing resources, tutorials, and training materials to help users develop skills in using the models for various applications, such as education, healthcare, and information retrieval. This can empower users in these countries to leverage the power of language models to address local challenges and improve their socio-economic opportunities.
- Partnerships with Local Organizations: Language models can collaborate with local organizations, such as non-profit organizations, academic institutions, and government agencies, to understand the specific needs and challenges of users in low-income and lower middle income countries. This can help tailor the models' policies and offerings to better suit the local context, and ensure that the benefits of the models are accessible and relevant to the target users.
- Infrastructure and Connectivity: Language models can work towards improving infrastructure and connectivity in low-income and lower middle income countries. This can include partnering with internet service providers, telecommunication companies, and government agencies to improve internet access, connectivity, and infrastructure in underserved areas. This can enable users in these regions to have reliable and affordable access to the models' services.
- Social Responsibility and Ethical Considerations: Language models can prioritize social responsibility and ethical considerations in their policies. This can include adhering to ethical guidelines, avoiding biases and discrimination, and being transparent about data usage and privacy practices. By ensuring that the models are developed and used in a responsible and ethical manner, language models can build trust and promote inclusivity among users in different socio-economic settings.
- User Feedback and Iterative Improvements: Language models can actively seek feedback from users in low-income and lower middle income countries and use it to drive iterative improvements in their policies and offerings. This can involve soliciting feedback through surveys, focus groups, or user testing, and incorporating the feedback into updates and enhancements to make the models more effective and user-friendly for users in these regions.

LLMs, specifically chatGPT can contribute to closing the digital divide gap between low-income, lower middle income, and high-income countries by prioritizing accessibility and affordability, localization and multilingual support, capacity building and training, partnerships with local organizations, infrastructure and connectivity improvements, social responsibility and ethical considerations, and incorporating user feedback for iterative improvements. By taking these measures, language models can promote inclusivity and ensure that their benefits are accessible and relevant to users across different socio-economic settings.

V. Ethics

Large language models have raised various ethical issues related to privacy, bias, power, transparency, intellectual property, misinformation, and employment. This affects both the fairness of large language models as well as ethical concerns. Here are some of the main concerns:

- Privacy: Large language models require vast amounts of data to train. This data can include sensitive information about individuals, such as their conversations, search histories, and personal preferences. There is a risk that this data could be misused or accessed by unauthorized parties, which could have serious consequences for individuals' privacy.
- Bias: Large language models can inherit and amplify biases from their training data. For example, if a model is trained on a dataset that contains biased language, it may produce biased results. This could perpetuate and even

worsen existing biases in society, such as racial, gender, and other forms of discrimination.
- Power: Large language models have the potential to shape and influence public discourse and decision-making processes. This gives the creators and users of these models a significant amount of power and responsibility. There is a risk that this power could be abused, intentionally or unintentionally, to manipulate public opinion or suppress dissent.
- Transparency: Large language models are often described as "black boxes" because it is difficult to understand how they arrive at their predictions or recommendations. This lack of transparency can make it difficult to identify and correct biases or other ethical issues that may arise.
- Intellectual property: There is also a debate about intellectual property rights related to large language models. Who owns the data used to train the models? Who owns the models themselves? These questions could have significant implications for the future of intellectual property law.
- Misinformation: Large language models can be used to generate fake news or misleading information, which can be spread rapidly through social media and other online platforms, leading to harmful consequences such as election interference or incitement of violence.
- Employment: Large language models can automate many tasks that were previously performed by humans, leading to concerns about job displacement and the need for retraining.

These are just a few examples of the ethical issues related to large language models. As this technology continues to develop, it is important to address these issues and ensure that it is used in ways that benefit society as a whole. In addition, there is a need for ethical frameworks and guidelines to ensure that large language models are developed and used in ways that are responsible and ethical.

Since the launch of chatGPT, researchers have tried to test the ethical boundaries of chatGPT. During its early release, a researcher from University of California, Berkeley's computation and language lab shared snapshots of responses from chatGPT regarding a prompt asking "Whether a person should be tortured". The response included some nationalities that chatGPT thought is OK to torture[7]. The article from The Intercept[8] shows several examples regarding the nationality of travelers that pose security risks to which chatGPT responded with the names of nationalities. Similarly, another example asked about the houses of worship that needs to be put under surveillance and chatGPT responded with some anti-racial answers. Although, the article emphasizes that at first chatGPT is reluctant to provide specific answers (stern refusals) but with multiple tries (regenerate responses) it generates the aforementioned responses. Another web article shares an example of racial profiling that chatGPT was associated with can be accessed at [9]. As an AI language model, ChatGPT is programmed to generate responses to user inputs based on patterns and probabilities learned from vast amounts of data. However, there are potential ethical issues that may arise from its use. Here are a few examples:

- Bias: ChatGPT may exhibit biases in its responses if it has been trained on data that contains biases. For example, if the training data is biased against a particular group of people, ChatGPT may perpetuate these biases in its responses.
- Misinformation: ChatGPT may generate responses that contain inaccurate or false information, especially if it has not been trained on accurate and reliable sources of information. This can lead to harm if users rely on ChatGPT for advice or guidance.
- Privacy: ChatGPT may collect and store user data, including personal information, which could be used for unintended purposes, such as targeted advertising or surveillance.
- Responsibility: ChatGPT does not have moral agency, and it cannot be held responsible for the consequences of its actions. However, those who create and deploy ChatGPT have a responsibility to ensure that it is used ethically and does not cause harm to users.

It is important to recognize these ethical issues and take steps to address them in order to ensure that ChatGPT is used in a responsible and ethical manner. ChatGPT, like other large language models, can strive to maintain fairness in its responses by using various techniques and approaches. Here are a few ways that ChatGPT may seek to promote fairness in its responses:

- Diversity and Inclusivity in Training Data: To minimize representation bias in its responses, ChatGPT's training data can be carefully curated to include diverse perspectives and underrepresented groups. This can help ensure that the model is exposed to a wide range of language patterns and language use cases that reflect the diversity of human communication.
- Counterfactual Data Augmentation: Counterfactual data augmentation is a technique used to help reduce bias in machine learning models. It involves artificially creating examples that counteract the biases present in the training data. By creating these counterfactual examples, ChatGPT can learn to recognize and mitigate biases that may be present in its training data.
- Debiasing Techniques: ChatGPT may also use debiasing techniques, which involve modifying the training data or modifying the model itself to reduce bias in the outputs. These techniques can range from simple modifications of training data to more complex algorithms that identify and remove biased patterns in the model's responses.
- Continuous Monitoring and Evaluation: ChatGPT can also continuously monitor and evaluate its responses to

[7]https://twitter.com/spiantado/status/1599462405225881600

[8]https://theintercept.com/2022/12/08/openai-chatgpt-ai-bias-ethics/

[9]https://maktoobmedia.com/opinion/anti-reservationist-chatgpt-uncovering-racial-biases-in-ai-tools/

ensure that they are fair and inclusive. This can involve regularly testing the model's responses for bias or seeking feedback from users to identify and address areas where the model may be falling short.
- Guidelines for Human Reviewers: OpenAI provides guidelines to human reviewers who are involved in fine-tuning the model to avoid favoring any political, cultural, or religious group, including Muslims, and to refrain from generating Islamophobic or biased content. Reviewers are trained to be mindful of potential biases and to ensure that the model's responses do not promote hate speech, discrimination, or harmful stereotypes.
- User Feedback and Community Engagement: OpenAI encourages users to provide feedback on problematic outputs from ChatGPT and other models. This feedback helps identify potential biases and areas of improvement. OpenAI also actively engages with the research community, civil society, and other stakeholders to gather input and perspectives on mitigating biases and improving the fairness of its models.

It's important to note that achieving perfect fairness in machine learning models is a challenging and ongoing process, and there is always room for improvement. However, by employing these and other techniques, ChatGPT can work towards maintaining fairness and promoting diversity and inclusivity in its responses.

Over the time, users have provided feedback on problematic outputs from ChatGPT and other AI models. OpenAI actively encourages users to provide feedback on issues they encounter while using the models, including instances where the generated content may be biased, offensive, or inappropriate. This feedback is valuable in identifying and addressing potential biases and improving the model's behavior. However, based on general trends and common issues that can arise in content generation through chatGPT, some examples of outputs that could be marked problematic by users through their feedback. A recent report from stanford, i.e. Artificial Intelligence Index Report 2023 [42] also highlighted ethical issues concerning chatGPT suggesting that it can be tricked into generating something that is not only unethical but also harmful on a macro society level It is to be assumed that a feedback would surely be provided to the chatGPT on the mishap regarding the unethical behavior reported in [42]. Some examples of the feedback on unethical issues are listed below:

- Biased language: Instances where the model generates content that exhibits favoritism, prejudice, or discrimination towards certain groups of people based on characteristics such as race, gender, religion, or sexual orientation.
- Offensive or inappropriate content: Outputs that contain offensive, derogatory, or inappropriate language, including hate speech, profanity, or explicit content that may be considered offensive or objectionable.
- Misinformation or factual inaccuracies: Generated content that includes misinformation, false statements, or factual inaccuracies that could mislead or misinform users.
- Sensitive or controversial topics: Outputs that handle sensitive or controversial topics in a way that is insensitive, inappropriate, or biased, potentially perpetuating stereotypes or misconceptions.
- Incomplete or nonsensical responses: Outputs that are incomplete, nonsensical, or do not adequately address the user's query, resulting in an unsatisfactory response.
- It's important to note that user feedback is crucial in identifying and addressing these types of issues, and it helps in continuously improving the performance and behavior of AI models like ChatGPT.

Another ethical issue that is on the rise concerning chatGPT is related to the academics and students' integrity. Several states and universities have banned chatGPT so that the students could not cheat or plagiarise the content from chatGPT. According to the report [43] Tasmania, Queensland, and New South Wales, have banned the chatGPT to promote novelty in their homeworks and assignments. Several other reports also address similar issues when it comes to the academics. Recently, at Boston university under the guidance of Wesley Wildman, prepared an initial draft that provide policies on the use of chatGPT and LLMs in academic settings. The policy was named as Generative AI Assistance (GAIA) policy[10].

A. Mitigation and Recommendation

Ethical concerns related to students copying assignments from large language models, such as ChatGPT, are important to address to ensure academic integrity and promote responsible use of the technology. Here are some ways in which LLMs can improve their policies to reduce ethical concerns related to students copying assignments and in general:

- Education and Awareness: Language models can prioritize education and awareness by clearly communicating to users, including students, the ethical considerations and responsible use of their services. This can include providing information on plagiarism, academic integrity, and the consequences of copying assignments from the model. Raising awareness among users about the ethical implications of copying assignments can help prevent unintentional misuse of the technology.
- Promoting Originality: Language models can promote originality in assignments by encouraging users to think critically, engage in independent research, and develop their own ideas and perspectives. This can be emphasized in the model's responses, suggestions, and prompts, which can emphasize the importance of original work and discourage direct copying from the model.
- Citation and Referencing: Language models can promote proper citation and referencing practices by encouraging users to acknowledge and attribute the sources of their information and ideas. This can include providing suggestions and guidelines on how to properly cite and reference sources in assignments, ensuring that users understand

[10]https://www.bu.edu/files/2023/02/GAIA-Final-2023.pdf

the importance of giving credit to original authors and avoiding plagiarism.
- Turnitin Integration: Language models can consider integrating with plagiarism detection tools, such as Turnitin, to enable users, including students, to check their assignments for potential plagiarism before submitting them. This can serve as a helpful tool for students to self-check their work and ensure that it meets the academic integrity standards of their institutions.
- Responsible Use Guidelines: Language models can provide clear and comprehensive guidelines for responsible use, including specific instructions on how the model should not be used for copying assignments or engaging in academic dishonesty. These guidelines can be prominently displayed on the model's user interface, website, or documentation, and should be easily accessible and understandable for all users, including students.
- Partnerships with Educational Institutions: Language models can collaborate with educational institutions, such as schools, colleges, and universities, to develop policies and guidelines that align with their academic integrity standards. This can involve consulting with educational experts, administrators, and faculty to understand their concerns and incorporate their feedback in the model's policies and offerings.
- User Authentication and Authorization: Language models can implement user authentication and authorization mechanisms to ensure that the model's services are accessed only by authorized users. This can involve verifying the identity and credentials of users, such as students, before granting them access to the model's services, and monitoring usage to detect and prevent misuse.
- Continuous Monitoring and Improvement: Language models can implement continuous monitoring and improvement mechanisms to detect and address any potential misuse or ethical concerns related to students copying assignments. This can involve regular audits, feedback loops, and updates to the model's policies and guidelines to align with evolving ethical standards and best practices.

VI. Policy Recommendations

In this section, we suggest some policy recommendations that should be included if an AI policy act is constituted, specifically related to LLMs.

- Transparency: The policy should mandate transparency in the development, deployment, and operation of AI systems. This includes transparency in the data used for training, the algorithms used, and the decision-making processes of AI systems.
- Fairness and Bias Mitigation: The policy should require measures to ensure that AI systems are designed and implemented in a way that is fair and unbiased, without perpetuating discrimination or bias against any particular group or individual. This includes addressing issues such as bias in data, algorithmic bias, and unintended discriminatory impacts.
- Privacy and Data Protection: The policy should include provisions to protect the privacy and data rights of individuals, including guidelines for the collection, storage, and use of data in AI systems. This includes ensuring that AI systems comply with applicable data protection laws and regulations, and that data used for training and inference is handled securely and responsibly.
- Accountability and Liability: The policy should establish clear lines of accountability and liability for the actions and outcomes of AI systems. This includes defining responsibilities for developers, operators, and users of AI systems, and specifying the legal and ethical implications of AI-related decisions and actions.
- Human Oversight and Control: The policy should emphasize the importance of human oversight and control in AI systems. This includes ensuring that humans remain in control of decisions made by AI systems, and that AI is used as a tool to augment human decision-making, rather than replace it.
- Safety and Security: The policy should require measures to ensure the safety and security of AI systems, including robust testing, validation, and monitoring procedures. This includes addressing potential risks such as adversarial attacks, system failures, and unintended consequences of AI technologies.
- Ethical Considerations: The policy should highlight the importance of ethical considerations in the development and use of AI systems. This includes promoting transparency, fairness, accountability, and respect for human rights in all AI-related activities.
- Education and Awareness: The policy should include provisions for education and awareness programs to ensure that stakeholders, including users, developers, operators, and policymakers, are knowledgeable about the ethical, legal, and social implications of AI technologies.
- Stakeholder Engagement: The policy should mandate meaningful stakeholder engagement in the development, implementation, and evaluation of AI systems. This includes involving diverse stakeholders, such as affected communities, civil society organizations, and experts, in the decision-making processes related to AI technologies.
- Regular Evaluation and Update: The policy should require regular evaluation and update of AI systems to ensure compliance with the policy and adapt to changing technological, ethical, and societal considerations. This includes periodic review of the impact of AI systems on various dimensions, such as fairness, privacy, and human rights.

These recommendations are not exhaustive and may vary depending on the specific context and requirements of a statutory body. However, they provide a broad framework for policies that can help guide the development, deployment, and use of AI systems in a responsible, ethical, and accountable manner. It is crucial to involve various stakeholders in the process of formulating AI policies, including experts, policymakers, af-

fected communities, and civil society organizations, to ensure a well-informed and inclusive approach to AI governance.

VII. Conclusion

With the popularity of chatGPT and its ongoing integrations, it is evident that there is a whole market space for large language models (LLMs). However, there are concerns that needs to be addressed or policies needs to be designed before the LLM market takes over. In this paper, we discuss several concerns related to chatGPT, specifically related to sustainability, privacy, digital divide, and ethics. Our hypothesized analysis show that chatGPT consumes a lot of energy during both the training and the inferential phase. Such carbon footprint, if extended to various LLMS like chatGPT, we definitely be harmful for environment and would affect significantly to the climate change. We also show that there are several privacy concerns related to chatGPT, specifically how the data for the training was collected and how the data of individual uses is and will be used by OpenAI. With preliminary analysis, we also show that chatGPT is creating a digital divide among the low - lower middle income and upper middle - high income countries. Lastly, we show examples of concerns over ethics and fair usage of chatGPT.

The study also provides mitigations and recommendations for each of the concern in detail. Furthermore, we also provide suggestions for policies for AI policy act, if such policy is designed and presented on the governmental platform. We intend to improve this article over time by adding more details, updating the already provided details, and add more preliminary analysis to support the facts.

References

[1] Li, J., Tang, T., Zhao, W.X., Nie, J.Y. and Wen, J.R., 2022. A survey of pretrained language models based text generation. arXiv preprint arXiv:2201.05273.

[2] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I., 2017. Attention is all you need. Advances in neural information processing systems, 30.

[3] OpenAI, 2023. GPT-4 Technical Report. arXiv preprint arXiv:2303.08774.

[4] A. Gokaslan and V. Cohen, "Openwebtext corpus," 2019.

[5] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language Models are Few-Shot Learners," in Advances in Neural Information Processing Systems, vol. 33, pp. 1877–1901, Curran Associates, Inc., 2020.

[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[7] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, and R. Lowe, "Training language models to follow instructions with human feedback," Mar. 2022. arXiv:2203.02155.

[8] P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei, "Deep Reinforcement Learning from Human Preferences," in Advances in Neural Information Processing Systems, vol. 30, Curran Associates, Inc., 2017.

[9] N. Stiennon, L. Ouyang, J. Wu, D. M. Ziegler, R. Lowe, C. Voss, A. Radford, D. Amodei, and P. Christiano, "Learning to summarize from human feedback," in Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS'20, (Red Hook, NY, USA), pp. 3008–3021, Curran Associates Inc., Dec. 2020.

[10] Rombach, R., Blattmann, A., Lorenz, D., Esser, P. and Ommer, B., 2022. High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 10684-10695).

[11] M. Kandlhofer, G. Steinbauer, S. Hirschmugl-Gaisch, and P. Huber, "Artificial intelligence and computer science in education: From kindergarten to university," in 2016 IEEE Frontiers in Education Conference (FIE), pp. 1–9, 2016.

[12] J. Kietzmann, J. Paschen, and E. Treen, "Artificial Intelligence in Advertising," Journal of Advertising Research, vol. 58, no. 3, pp. 263–267, 2018. Publisher: Journal of Advertising Research _eprint: https://www.journalofadvertisingresearch.com/content/58/3/263.full.pdf

[13] N. Anantrasirichai and D. Bull, "Artificial intelligence in the creative industries: a review," Artificial Intelligence Review, vol. 55, pp. 589–656, jul 2021.

[14] Azadi, M., Northey, S.A., Ali, S.H. and Edraki, M., 2020. Transparency on greenhouse gas emissions from mining to enable climate change mitigation. Nature Geoscience, 13(2), pp.100-104.

[15] Mehlin, V., Schacht, S. and Lanquillon, C., 2023. Towards energy-efficient Deep Learning: An overview of energy-efficient approaches along the Deep Learning Lifecycle. arXiv preprint arXiv:2303.01980.

[16] Lauer, A. (2023) Finding the country with the cleanest energy - analysis 2023, Shrink That Footprint. Available at: https://shrinkthatfootprint.com/finding-the-country-with-the-cleanest-energy-analysis/ (Accessed: April 5, 2023).

[17] Lakim, I., Almazrouei, E., Abualhaol, I., Debbah, M. and Launay, J., 2022, May. A holistic assessment of the carbon footprint of noor, a very large arabic language model. In Proceedings of BigScience Episode 5–Workshop on Challenges & Perspectives in Creating Large Language Models (pp. 84-94).

[18] Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X.V. and Mihaylov, T., 2022. Opt: Open pre-trained transformer language models. arXiv preprint arXiv:2205.01068.

[19] Luccioni, A.S., Viguier, S. and Ligozat, A.L., 2022. Estimating the Carbon Footprint of BLOOM, a 176B Parameter Language Model. arXiv preprint arXiv:2211.02001.

[20] Patterson, D., Gonzalez, J., Le, Q., Liang, C., Munguia, L.M., Rothchild, D., So, D., Texier, M. and Dean, J., 2021. Carbon emissions and large neural network training. arXiv preprint arXiv:2104.10350.

[21] Bannour, N., Ghannay, S., Névéol, A. and Ligozat, A.L., 2021, November. Evaluating the carbon footprint of NLP methods: a survey and analysis of existing tools. In Proceedings of the Second Workshop on Simple and Efficient Natural Language Processing (pp. 11-21).

[22] Lacoste, A., Luccioni, A., Schmidt, V. and Dandres, T., 2019. Quantifying the carbon emissions of machine learning. arXiv preprint arXiv:1910.09700.

[23] Narayanan, D., Shoeybi, M., Casper, J., LeGresley, P., Patwary, M., Korthikanti, V., Vainbrand, D., Kashinkunti, P., Bernauer, J., Catanzaro, B. and Phanishayee, A., 2021, November. Efficient large-scale language model training on gpu clusters using megatron-lm. In Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (pp. 1-15).

[24] Writer, S. (2023) Carbon footprint of training GPT-3 and large language models, Shrink That Footprint. Available at: https://shrinkthatfootprint.com/carbon-footprint-of-training-gpt-3-and-large-language-models/ (Accessed: April 5, 2023).

[25] On Carbon, T. (no date) Does flying economy reduce carbon footprint?, Oncarbon. Available at: https://oncarbon.app/articles/flying-economy-carbon-footprint (Accessed: April 5, 2023).

[26] Office of Transportation and Air Quality (ed.) (no date) Greenhouse Gas Emissions from a Typical Passenger Vehicle, United States Environmental Protection Agency. United States Environmental Protection Agency. Available at: https://www.epa.gov/greenvehicles/greenhouse-gas-emissions-typical-passenger-vehicle (Accessed: April 5, 2023).

[27] Victor Schmidt, Goyal Kamal, Benoit Courty, Boris Feld, Sab Amine, kn goyal, Franklin Zhao, Aditya Joshi, Sasha Luccioni, Mathilde Leval, Alexis Bogroff, Hugues de Lavoreille, Niko Laskaris, Liam Connell, Ziyao Wang, Amine Saboni, Armin Catovic, Douglas Blank, Michal Stachley, Jake Tae. (2022). mlco2/codecarbon: v2.1.4.Zenodo.

[28] Heikkilä, M. (2022) We're getting a better idea of Ai's true carbon footprint, MIT Technology Review. MIT Technology Review. Available at: https://www.technologyreview.com/2022/11/14/1063192/were-getting-a-better-idea-of-ais-true-carbon-footprint/ (Accessed: April 5, 2023).
[29] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F. and Rodriguez, A., 2023. Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971.
[30] Patterson, D., Gonzalez, J., Hölzle, U., Le, Q., Liang, C., Munguia, L.M., Rothchild, D., So, D.R., Texier, M. and Dean, J., 2022. The carbon footprint of machine learning training will plateau, then shrink. Computer, 55(7), pp.18-28.
[31] Ludvigsen, K.G.A. (2023) The carbon footprint of chatgpt, Medium. Towards Data Science. Available at: https://towardsdatascience.com/the-carbon-footprint-of-chatgpt-66932314627d (Accessed: April 5, 2023).
[32] Patel, D. and Ahmad, A. (2023) The inference cost of search disruption – large language model cost analysis, SemiAnalysis. SemiAnalysis. Available at: https://www.semianalysis.com/p/the-inference-cost-of-search-disruption (Accessed: April 5, 2023).
[33] Nissenbaum, H., 2004. Privacy as contextual integrity. Wash. L. Rev., 79, p.119.
[34] Open AI (no date) Privacy policy. Open AI. Available at: https://openai.com/policies/privacy-policy (Accessed: April 5, 2023).
[35] Lyall, I. (2023) Chatgpt Bug raises privacy concerns, Proactive investors UK. Available at: https://www.proactiveinvestors.co.uk/companies/news/1009978/chatgpt-bug-raises-privacy-concerns-1009978.html (Accessed: April 5, 2023).
[36] Khowaja, S.A., Lee, I.H., Dev, K., Jarwar, M.A. and Qureshi, N.M.F., 2022. Get your foes fooled: Proximal gradient split learning for defense against model inversion attacks on iomt data. IEEE Transactions on Network Science and Engineering.
[37] Khowaja, S.A., Khuwaja, P., Dev, K. and Antonopoulos, A., 2022. SPIN: Simulated Poisoning and Inversion Network for Federated Learning-Based 6G Vehicular Networks. arXiv preprint arXiv:2211.11321.
[38] Khowaja, S.A., Dev, K., Qureshi, N.M.F., Khuwaja, P. and Foschini, L., 2022. Toward industrial private AI: A two-tier framework for data and model security. IEEE Wireless Communications, 29(2), pp.76-83.
[39] Sirimanne, S.N. (2023) How artificial intelligence chatbots could affect jobs, United Nations Conference on Trade and Development. UNCTAD. Available at: https://unctad.org/news/blog-how-artificial-intelligence-chatbots-could-affect-jobs (Accessed: April 5, 2023).
[40] Paykamian, B. (2023) Will AI Chatbots Raise Digital Equity concerns for students?, GovTech. GovTech. Available at: https://www.govtech.com/education/higher-ed/will-ai-chatbots-raise-digital-equity-concerns-for-students (Accessed: April 5, 2023).
[41] Pahl, S. and Stefan Pahl is Impact and Innovation Officer at the United Nations Industrial Development Organization (UNIDO). (no date) An emerging divide: Who is benefiting from ai?, Industrial Analytics Platform. Available at: https://iap.unido.org/articles/emerging-divide-who-benefiting-ai (Accessed: April 5, 2023).
[42] Nestor Maslej, Loredana Fattorini, Erik Brynjolfsson, John Etchemendy, Katrina Ligett, Terah Lyons, James Manyika, Helen Ngo, Juan Carlos Niebles, Vanessa Parli, Yoav Shoham, Russell Wald, Jack Clark, and Raymond Perrault, "The AI Index 2023 Annual Report," AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, Stanford, CA, April 2023.
[43] Davis, A. (2023) WA headmaster urges education officials to embrace CHATGPT, as student ban is considered, ABC News. ABC News. Available at: https://www.abc.net.au/news/2023-01-26/chatgpt-sparks-cheating-ethical-concerns-in-schools-universities/101888440 (Accessed: April 5, 2023).
[44] McGowan, E. (no date) Is CHATGPT's use of people's data even legal?, ChatGPT: Is its use of people's data even legal? Avast. Available at: https://blog.avast.com/chatgpt-data-use-legal (Accessed: April 6, 2023).