# Artificial Divides: Global AI Access Disparities and Constructions of New Digital Realities

**Lizhi Peng**

**A thesis**

**submitted in partial fulfillment of the**

**requirements for the degree of**

**Master of Arts**

**University of Washington**

**2024**

**Committee:**

**Bo Zhao**

**Mia Bennett**

**Program Authorized to Offer Degree:**

**Geography**

University of Washington

## Abstract

Artificial Divides: Global AI Access Disparities and Constructions of New Digital Realities

Lizhi Peng

Chair of the Supervisory Committee:

Bo Zhao

Department of Geography

Following the wake of ChatGPT's release in late 2022, we have witnessed the launch of an "arms race" of generative AI technology as large language models (LLMs) entered a phase of rapid development and advancement, with promises of revolutionary transformations of work spaces and everyday life from major tech companies. As contestants and major power players like OpenAI, Google, Anthropic, and Meta enter the game, many ethical concerns have been raised regarding whether this technology will truly be the beginning of the next technological revolution, and if so, whether it will be beneficial to human society as a whole or only serve in the interest of a few. At the same time. Many AI scientists, researchers and industry leaders have come forth with claims that how we handle this new technology will be critical to the wellbeing or even

survival of humanity in the future. As the industry chases after the promise of the sparks of artificial general intelligence (AGI), that one day truly autonomous super AIs capable of outsmarting the human brain in generalized tasks can be achieved, discussions of AI alignment, the checks and balances that will keep AI acting and behaving according to human values and principles continue. Yet, even as we enthuse over the potential transformations this technology brings to our society, it is important to point out that true alignment requires the input from people across all backgrounds and walks of life. It would be concerning to leave the definition of “human values” in the hands of a few leaders behind closed doors.

This thesis consists of two papers exploring the potential ethical risks and concerns AI technology raises regarding accessibility and equity issues both in the AI industry and in the broader society centering two major questions: “who has access to AIs?” and “who builds the AIs?”. In the first paper, I will examine current tangible and intangible barriers to accessing AI subscriptions, and how performance of an AI differs across linguistic and geographical contexts, as a way of painting a bigger picture of the network of unfair representations behind the training, deployment and access of commercial AIs. In the second paper, I build on top of the previous theoretical foundations to connect current discussions in AI training and alignment ethics with interdisciplinary views and critiques on technology and society. I propose a framework that conceptualizes large language models as models of our society, interpreting AI as the reproduction and embodiment of the intricate power dynamics and inequalities fueled by social discourses and media representation. In this model I refer to as “layers of realities”, I explore the complex

relationship between the physical world, the internet and digital media, and the world of large language models, in order to highlight the urgency to not see AI biases as a standalone issue within the industry, but a sign that alerts us to address issues in the physical and digital world such as unequal technology access and unfair mis/under-representation of marginalized identities in this increasingly digitized world.

# The New Digital Divide:

# Global Disparities in the Generative AI Industry


## Abstract

In this paper I explore the disparities in accessibility and quality of premium artificial intelligence (AI) services, using OpenAI's ChatGPT Plus subscription as a case study to analyze global inequalities in the distribution of AI technologies. The focus for this paper is on economic viability and linguistic proficiency of current AI models, while acknowledging the complex network of unfair underrepresentation and underperformance issues beneath the surface. The research highlights how premium AI subscriptions like ChatGPT Plus exacerbate existing disparities by privileging certain geographies and demographics while excluding others, creating feedback loops enhancing the biases in AI development and reinforcing the digital divide. With the employment of a novel metric, the National Artificial Intelligence Cost Effectiveness (NAICE) index, this study puts the relative affordability and performance of AI services across different regions into perspective, revealing significant disadvantages faced by countries in Sub-Saharan Africa, South Asia, the Middle East, and North Africa compared to North American and European countries. It shows that while AI has the potential to enhance societal functions, its current trajectory could deepen global inequalities. My findings emphasize the need for more inclusive AI development practices and broader accessibility to ensure that the benefits of AI technologies are equitably shared. Lastly, I highlight the importance of interdisciplinary efforts to address the sociopolitical implications of AI and call on the academic community to lend your expertise to this pressing issue.

## Introduction

In this paper I will be examining the global disparity in accessibility and quality of premium AI services, to conceptualize the question of “who” have access to AIs, highlighting the multidimensional nature of the AI equity dilemma being more than a matter of economic feasibility, instead it involves digital representation of demographics deeply shaped by the complex power dynamics of geopolitics and control of global discourses. I will be examining ChatGPT’s premium subscription, which gives access to their latest model ChatGPT-4, as an entry point, due to its popularity and its current status as the catalyst that launched this current AI “arms race”. Not only is ChatGPT often used as a comparison for benchmarks for new model research, the price point set for its premium subscription was followed by many of its major competitors on the market. In many senses, OpenAI’s viral model has set the standard for the generative AI industry to follow. Through analyzing the different obstacles users may face in accessing and utilizing this service, I hope to build an understanding of how this emerging technology can potentially enhance global disparities from two main perspectives: the economic viability to afford the subscription, and the actual performance of the model when taking into account its proficiency to interact with different demographics and communities, which in the scope of this study is limited to linguistic difference. While ideally, a more comprehensive evaluation would also consider factors such as the model’s understanding of different cultural and religious knowledge, of different racialized and gendered experiences, such a task requires data not readily available and necessitates a large team of experts from various backgrounds

and disciplines. And I hope that this paper can pave the way for future research to dig deeper into this new perspective on AI alignment.

**Background**

In early 2023, global attention has become increasingly fixated on the emerging applications of generative artificial intelligence (AI) technologies following the rise to popularity of ChatGPT, a chatbot powered by a large language model developed by American technology company OpenAI. While there have been many ethical issues surrounding the training of these models and their potential for misuse, such as spreading misinformation or plagiarism of intellectual properties.

ChatGPT is an AI Chatbot based on the GPT-3.5 large language model, a natural language processing model trained with dataset and calculating power on a massive scale to iteratively predict texts to achieve the effect of autocompletion, and, in its particular use case, generate almost human-like responses to user input. It is worth noting that the term ChatGPT can also be used to refer to the GPT-4 model, which is a more advanced version of the AI with multi-modal support. Since its launch in November 2022, it quickly became viral and broke all previous records to become the world's fastest app to reach 100 million users in just two months after the launch, a feat that took the previous record holder TikTok nine months and Instagram more than two years to achieve (Milmo, 2023).

Following its rise to popularity in early 2023, the world seemingly entered an era of rapid evolution in the AI industry. Microsoft, who has been a major investor and partner of OpenAI, with the latter's models trained and deployed on its cloud computing platforms, took advantage of the momentum with its GPT-4 powered new Bing search engine and Edge browser, forcing its competitor and long time monopoly of the search engine business Google into a rushed and unpolished demo of their own AI Bard based on their LaMDA model (Metz and Grant, 2023, Coulter & Bensinger (2023)). It has since announced an ambitious plan to incorporate GPT-4 support into other products from Office 365 to Github and even Windows itself, aiming to have the AI-powered copilot to revolutionize the idea of productivity in work places (Dohmke, 2023; Spataro 2023). These are but a few examples of practical applications being enhanced with generative AIs like ChatGPT. Outside of the big tech, ChatGPT is also gaining attention in its potential as a platform connecting the user to other tools and services with third party plugins and even showed potentials for identifying other specialized AI models to "outsource" its work (Shen et al, 2023), all through its natural language interface. ChatGPT may well serve as the "fundamental model" that many have speculated will be key in powering professional and enterprise level AI tools such as GeoAI for spatial modeling (Janowicz, 2019). While these are in theory possible with any large language models, due to ChatGPT being the most popular and accessible one to be released so far, it is likely it will continue to be regarded as an industrial standard for some time to come, especially following the release of its app store-like custom GPT update..

ChatGPT, especially following the release of its GPT-4 Model, has so far been one of the global leaders in generative AI models available on the market, both in terms of the high-quality of its responses and low bar of entry. Until recently in April 2024, the basic ChatGPT model required a valid phone number verification from a supported country or region in order to participate in the public preview, but is free to use otherwise. While others, like Google's Bard (now replaced by Gemini) or Microsoft new Bing (now renamed as Copilot), had a history of originally requiring users to sign up for a waitlist that can go from days to months before they are able to get access. What they have in common is that initially these supported countries are restricted to a small number of countries starting with the United States, and often the United Kingdoms and other English Speaking countries. In terms of its high-quality responses, a research team from Microsoft given early access to the GPT-4 model went as far as boldly claiming that this may be the "sparks of artificial general intelligence", showing potential for a "true" intelligence that can achieve and exceed human level intelligence in generalized tasks the future, after a thorough research of the model finding many unexpected and unexplainable signs showing the model's apparent emerging logical capabilities. To quote the paper published by Microsoft Research: "Why does it exhibit such general and flexible intelligence when it is at its core merely the combination of simple algorithmic components—gradient descent and large-scale transformers with extremely large amounts of data?" (Bubeck et al, 2023). Not even its developers could yet answer to this. Some scholars are theorizing that the model may already possess a "theory of mind" under certain definitions and suggest it may be beneficial to start looking at these neural networks from a psychological perspective (Kosinski, 2023). Perhaps one thing

that can be agreed on is that future research into how AIs understands and interacts with the world and the implications of AI technologies will, and should be increasingly relevant to scholars across all disciplines. As part of this effort to make AI a topic of interdisciplinary focus, this paper focuses on the technology's sociopolitical implications as it threatens to magnify the existing technological inequality on a global scale.

While the question of how generative AI like ChatGPT may be used for our society, as well as their potential pros and cons are still being widely discussed, consensus from scholars across various fields could agree that it has immense potential to improve our society, as much as we need to watch out for the possible harms it could cost us (Dwivedi et al, 2023). There is little doubt that ChatGPT has the potential to become an integral element of many people's daily lives. ChatGPT has the potential to reduce technical barriers and help small businesses and individual entrepreneurs turn their ideas into products with the upgrades it received alongside the GPT-4 updates, including web browsing, image generation and coding capabilities. But such power is not equality distributed among populations across different geographies and communities as it is barred behind paid access through the ChatGPT Plus subscription. In this paper I will mainly focus on the economic and linguistic aspects of this gap, but I want to make a quick note that there exist many other forms of disparities beyond the ones we are able to study and present, including race, gender, sexualities that may exhibit different patterns when examined at a different scale.

## Theoretical Perspectives

### *The New Digital Divide*

The term digital divide itself can be seen as an umbrella term that encompasses a variety of inequalities brought forth by the advent of technologies. These inequalities can span across different industries such as healthcare, communication and transportation, it can refer to the gap in physical access to smart devices, the internet, or digital media, or it can refer to the gap in one's familiarity and skill with technologies. What's more, it continues to take on new forms of division as new technology emerges (Van Dijk, 2020, Lythreatis et al, 2022). Ever since the beginning of the Covid-19 Pandemic, it has become increasingly clear that access and experience with digital technologies is becoming more vital than ever, with the likes of digital payments, digital healthcare, remote classes and remote working becoming more popular and even pseudo-mandatory. In China, for example, research found that the age-induced digital divide has expanded substantially during the pandemic as the country picked up speed in digitization at the height of the pandemic, which caused inconvenience to the elderly population as they struggled with accessing healthcare and making transactions over smart devices (Song et al., 2021). This increasing digitization of work and life continues to further marginalize already technologically disadvantaged populations as part of the non-sanitary risks brought forth by the pandemic (Nguyen et al, 2021). And now, with the growing popularity of generative AI technologies, another form of digital divide may have already emerged.

Characteristics of this new AI-led digital divide can be summarized as follows: the lack of education and literacy on the technology for the public, and the underrepresentation of communities within the models. With the AI's convincing human-like mannerisms and overly enthusiastic news boasting the capabilities and intelligence of this new wave of AI models, it is easy for unsuspecting person to believe that the AI we have today have achieved the level of human intelligence or sentience of a similar kind as in science fiction, when that is still far from the truth. There have been warnings against the ramifications of anthropomorphizing AIs, which are fundamentally different from human consciousness (Salles, 2020). But the market for AI companions and assistants continues to grow, events such as Microsoft's preview launch of new Bing assistant "Sydney" had shown, the public can be susceptible to developing feelings for AI assistants when they exhibit certain unhinged human-like traits and behaviors, and are prone to the illusions that there is a spark of sentience underneath, to the degree of (Edwards, 2023, Peng, 2024).

The underrepresentation issue within models, on the other hand, is a distinctive trait of this new digital divide. Unlike the traditional digital divide that is centered around the access to digital media, which is something visible and quantifiable, the unequal representation of different identities and experiences within language models is an invisible gap. People can have access to the same AI service in principle, but the actual performance and relevancy of a model's output is highly variable depending on the user's background and identity, their native language, for example. This is part of a much larger subject that I will only be examining one aspect of in the scope of this

paper. In the technical report published for GPT-4, researchers recorded the GPT 4.0 model's performance across 27 different languages using a system of multiple choice questions spanning 57 subjects in both academic and professional settings. They found that the model is much more proficient when interacted with in English, followed by many other major European languages. While less-spoken languages such as Marathi and Telugu that were tested performed considerably worse, scoring below even GPT 3.5's performance in English in the evaluation(OpenAI, 2023). It is worth noting that the test focused on the model's performance in academic and professional settings, and does not take into account more qualitative metrics such as the model's capacity in understanding specific cultural contexts, social cues or localized knowledge in regional dialects of creole languages. When in real life these nuanced factors all play a part in determining the usefulness of an AI. A model's understanding of different languages is highly dependent on the quality of the training data. In the case of GPT-3, ChatGPT's predecessor, nearly 93% of its training data was in English by word count, whereas the aforementioned Marathi and Telugu only consist of 0.002% and 0.001% of the data respectively. Likewise, many widely spoken languages, such as Chinese, only made up 0.1% of the training data, showing a significant underrepresentation of Chinese speaking communities in the training data (Brown et al., 2020). This discrepancy likely arises from the greater availability of English content in publicly available datasets commonly used for AI training, which is beyond the control of developers and reflects wider global power dynamics behind the dominance of English in research and industry. It is a glimpse into a broader, intersectional web of global inequalities that yet remains obscured without readily available data to support them (Peng, 2024). But that is not to

say OpenAI did not introduce their own biases in the process of developing their models. For example, the Common Crawl is one of the largest open-source datasets of aggregated web crawler data, commonly used in training of generative AI and a major source of data for OpenAI in training GPT-3, taking up 60% of the training data (Brown et al., 2020). The statistics on linguistic diversity of the dataset reported that while still disproportionately dominant, English data contributes to just under 50% of the entire dataset. To put into perspective, after OpenAI's curation and assessment to filter out "quality data", English data increased from a level of ~45% in one of its major sources to 93% in the final training data, while the likes of Chinese went from ~5% down to ~0.1%.

***The multidimensionality of AI Inequality***

The factors shaping inequalities during the training of LLMs is multidimensional and spans the aforementioned linguistic disparities, political, economical, cultural, and more. Given the substantial investments needed to develop proficient AI systems, not all countries, particularly smaller and developing ones, will have the resources to create their own equivalent of ChatGPT, nor would it be economically feasible or environmentally responsible to do so. Although there are an increasing number of alternative LLMs to ChatGPT to have been announced or in development, to name a few of its major competitors, Claude from Anthropic, Bard(now upgraded into Gemini) by Google and META's Llama, which is open sourced and powers a substantial amount of the current line of fine-tuned models developed by smaller companies and academic researchers. With the majority of companies having the requisite resources to develop and train their own models based in the United States and answering to U.S.

regulations, U.S. legislations hold unrivaled political power and influence over the global AI industry, at least for the time being. Which raises concerns such as security risks of having sensitive data being held and processed by foreign companies as well as potential biases in the models in different cultural, political and linguistic contexts. Therefore it is crucial that we examine and try to understand the disparities among countries in terms of affordability and accessibility of these technologies.

So why should we care about ChatGPT Plus, when the base model is currently free for (almost) all to use? And what does the premium subscription do, exactly? For now it translates to priority access to the ChatGPT platform during peak hours whereas free users may be refused service, in addition to limited usage of early access models yet to be released to the public. Among those models are the latest GPT-4 which has been drawing attention to its much improved performance, not only in terms of quantitative benchmarks it was tested against, but also its unexpected emergence of apparent “intelligence”. Since its release, the GPT-4 model has been upgraded to support visual input, image generation, web browsing, built-in code execution environment, and more, all of which can be game-changing in their own sense. Take for example the custom GPTs feature, which allows users to customize an AI assistant’s behaviors and grant it expanded API calling privileges, allows them to function as autonomous agents capable of independently executing tasks such as doing web research, performing data analysis, writing and sending emails, creating diagrams or even maps. While new features will be gradually rolled out to free users over time, months of priority access of increasingly powerful models and features in the future, like the potentially upcoming GPT-5 model,

will continue to create rifts between free and paid users. Whether the user is an entrepreneur looking to launch a new business idea, an academic doing research on AI, or just an average user trying to make their daily lives more convenient, being a few months ahead of peers give subscribers the first mover advantage in the rapidly evolving industry and a jumpstart in developing AI-enabled products, where ideas become more valuable as the AI itself in part contributes to the breaking down of technical barriers. While the majority of the public could only speculate and rely on second-hand information, sometimes outdated or mixed with misinformation from social media and content creators. Not only does this give certain groups the privilege to control the narrative on AI, it also limits which demographics are able to actively engage in the shaping of future models through feedback.

**Methodology**

In order to explore and visualize the tip of the iceberg on the intersections of affordability gap and unequal representation of communities, I focused on collecting available demographic and economic data from publicly accessible databases from reputable organizations under the United Nations Statistics Division and the World Bank, as well as previous published research and benchmark scores from first-party sources such as OpenAI. In order to analyze and compare the overlapping and intersecting dimensions of unequal access to AI, I developed a novel metric to represent the relative affordability and quality of AI services on a national scale: the National Artificial Intelligence Cost Effectiveness (NAICE) index. The NAICE index is calculated by integrating each country's relative GDP per capita, compared to the United States, with an estimated

weighted average of the model's performance in the population's first language. The formula for NAICE is presented as follows:

$$\text{NAICE} = P_r \times \left(\sum_{i=1}^{n} \omega_i \times S_i\right)$$

Here the $P_r$ represents the relative gross domestic product per capita in units of purchase power parity, expressed as a percentage normalized against the United States. The $S_i$ denotes the model's benchmark score for language $i$, and $\omega_i$ is the percentage of population speaking language $i$.This composite index encompasses two critical dimensions of global AI disparities: the economic and performance factors, underscoring the inequitable affordability and varying quality of AI services across different regions. While ideally, the index should consider additional performance factors such as the model's understanding of different cultures and granular local knowledge, this simplified formula is meant to highlight the multidimensional nature of these equity issues, especially to emphasize that there are many aspects of this inequality that can be difficult, if not outright impossible to quantify, rather than to be interpreted as a comprehensive representation of the current global disparity.

***Data Sources and Collection***

The data for this analysis were derived from multiple publicly available databases and government websites. Population percentages by language were obtained from the United Nations Statistics Division's Demographic Statistics Database. In instances where specific language data were unavailable, a country's primary language was assigned based on listings from the CIA World Factbook. Economic indicators and

population figures were sourced from the World Bank. AI performance benchmarks were adapted from OpenAI's GPT-4 Technical Paper, which detailed the outcomes of rigorous internal testing of the model through the MMLU (Multimodal Multi-Language Understanding) examination. Languages not covered in the benchmark received a default score of 75, acknowledging the limitations of available data.The data for these analyses come from a range of publicly available databases and government websites. Specific percentages of population by language is calculated using data from the United Nations Statistics Division's Demographic Statistics Database. It is important to note that the exact date of these data were updated varies by country. When the information is unavailable, a country is instead assigned their primary language based on the CIA World Factbook listings. The economic data and population data are sourced from the World Bank. I selected the year 2022 specifically due to it being the latest cutoff date before projections begin. The benchmark scores for each language are adopted from OpenAI's official *GPT-4 Technical Paper* (Bubeck et al, 2023), in which they disclosed the MMLU benchmark scores (MMLU stands for Massive Multitask Language Understanding, a series of comprehensive tests for measuring a language model's problem solving capabilities across a variety of subjects in both professional and academic settings) for their models during internal testing. Languages that were not tested on the benchmark were given a default score of 75 taking into account the overall scoring tendencies. In the year since the paper has been published, the model has been updated multiple times and due to how rapid advancements in the AI industry can be, these may not reflect the latest performance of the model.

### *Limitations and Future Improvements*

Due to the sheer scale of the topic and limited time and resources for research, many assumptions had to be made, for example it is difficult to take into account how people who are bilingual or multilingual experience this linguistic discrepancy. For countries like where multilingualism is common and social stratification is much more prevalent, a thorough research would necessitate a local scholar's expertise. So once again I would like to reiterate that I conduct this analysis with hopes that this work will pave the way for future research on a more extensive and expansive scale.

## Results and Discussion

My analysis of the AI cost-effectiveness index yielded a few significant findings: the majority of Sub-Saharan African, South Asian, Middle East and North African countries are in the bottom 50% of populations in terms of AI access. Which is clear to see in Figure 1, which shows a scatterplot of the NAICE index scores of 176 countries or areas, measured against their population size with a horizontal line indicating the estimated 50% division of the world's population by NAICE scores. East Asian and Pacific countries and Latin American and Caribbean countries have a larger spread near the center, while the top of the chart is dominated by North American and European countries.

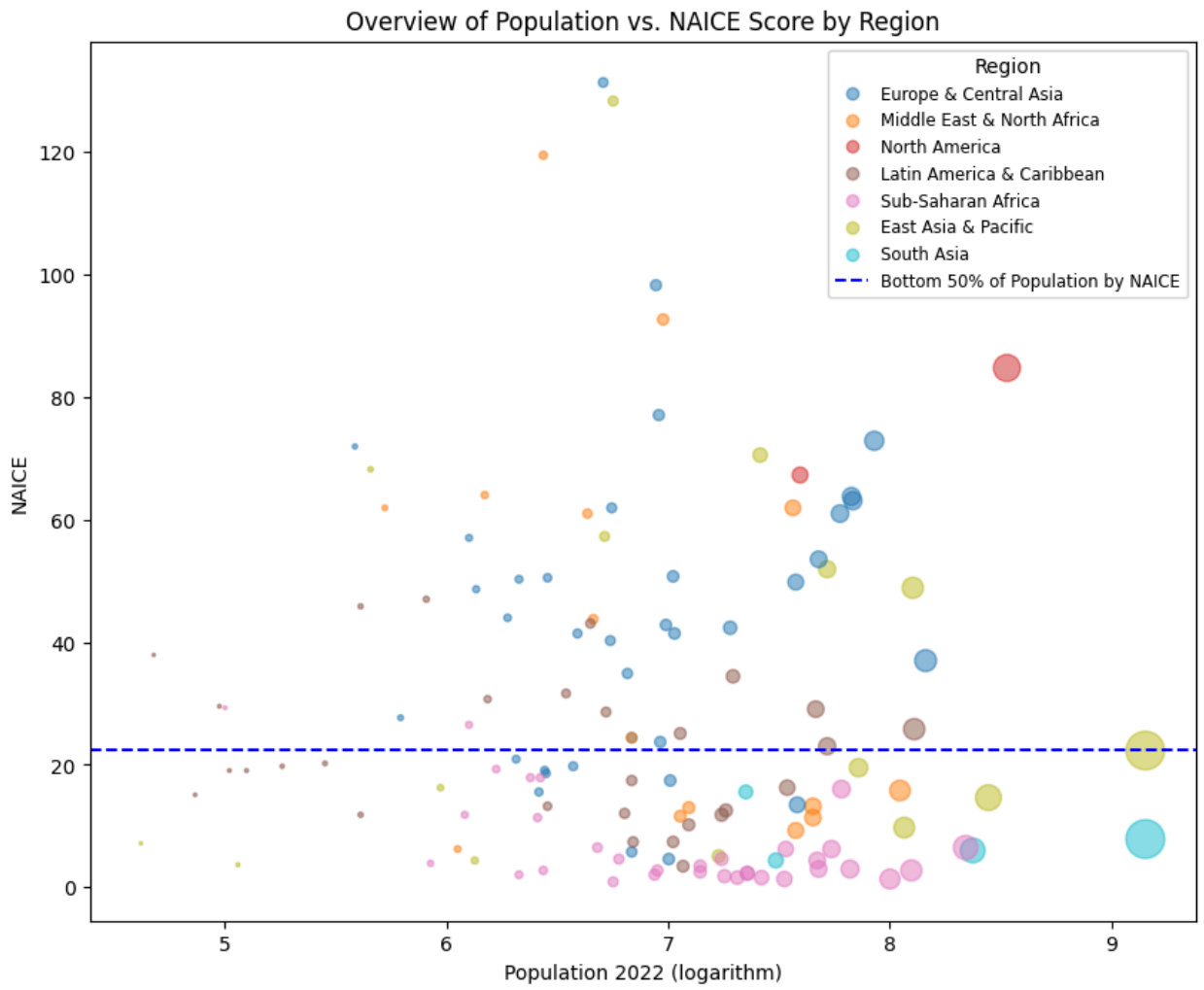
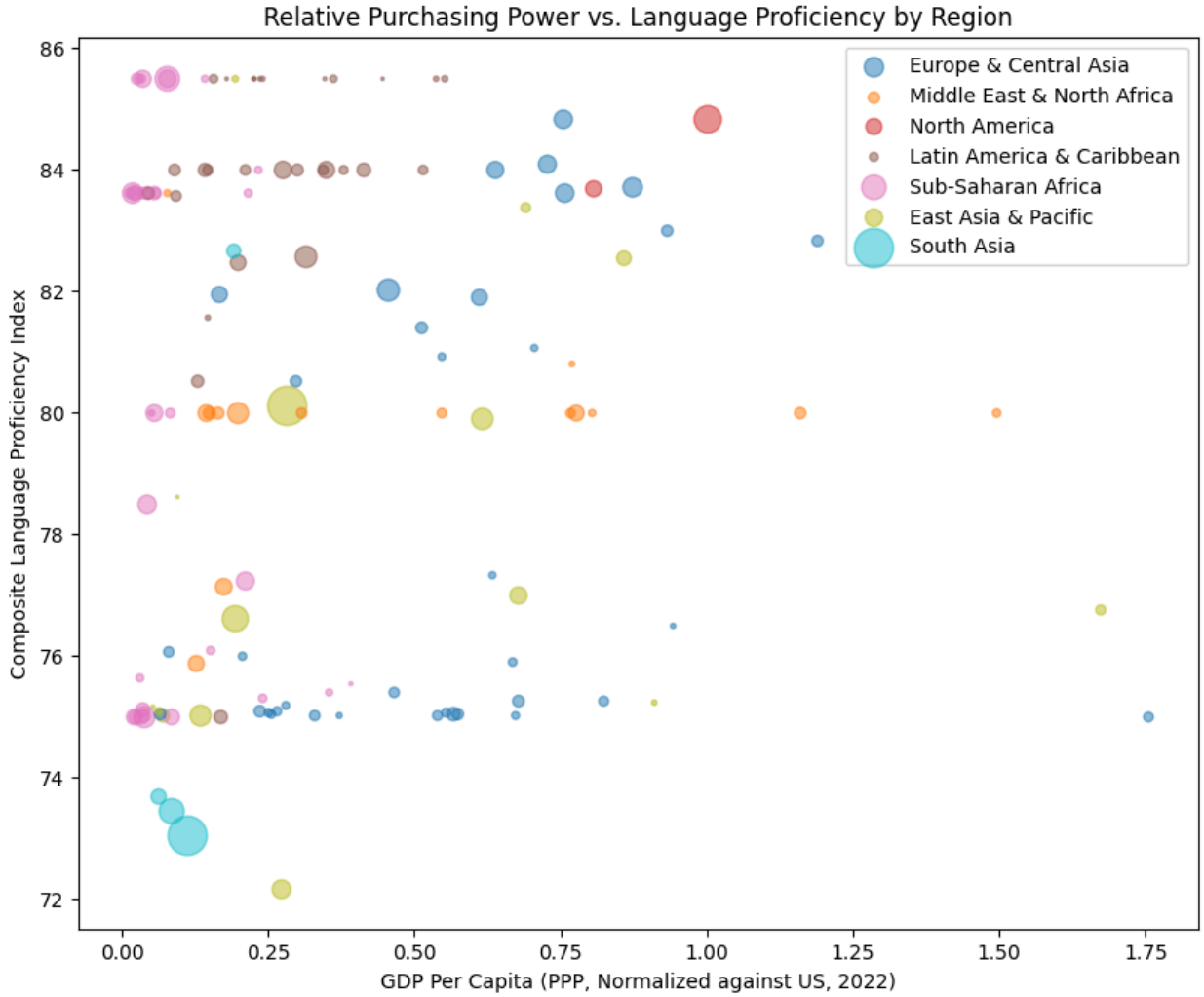



Figure 2 takes a closer examination at the intersections between the composite linguistic performance scores compared to the relative purchasing power in each country. Here we can see that major North American and European countries are at a relative advantage compared to much of the world, both in terms of economic status and linguistic dominance. It is interesting to see that many Latin American & Caribbean countries as well as Sub-Saharan African countries, despite being at an economic disadvantage, do not possess much of a linguistic barrier when using AIs, in part due to their colonial history and sharing the same language as their colonizers. (It is important to clarify that this assumes their languages are identical, when in reality, creolization can

result in terms and dialects that are far apart enough to confuse AIs). This is in contrast to East Asian & Pacific countries, notably Japan, Korea and Singapore, despite having robust economies, nonetheless suffer from less competent AI in their official languages. When considering disparities in affordability between the Global North and the Global South, as well as the linguistic and cultural hegemony of the white European-Americans, the majority of the world is placed at a disadvantage in one way or another, oftentimes in multiple overlapping fashion.



This situation raises several concerns, such as which societies truly benefit from AI-enhanced productivity. Notably, the underpaid Kenyan workers who contributed to the development of ChatGPT models may not be able to afford access to the latest model themselves(). I also want to acknowledge that numerous other barriers exist beyond the fI'm presenting in this paper. For instance, ChatGPT Plus was initially made available only to users in the United States before gradually rolling out to users around the world. Countries like China, Russia, and Iran were not supported by OpenAI and it had been briefly suspended in Italy out of privacy concerns (OpenAI, n.d.). While users from these regions could potentially bypass these restrictions through proxies and virtual private networks (VPNs), this workaround still presents a significant deterrent for the average citizen. The fact that even though Chinese as a language is supported, the people from China are denied its service serves as a political motivation for China to develop its own LLMs, which leaves its citizens with fewer options for their source of information. The perspectives I've presented are not intended to be comprehensive representations of the complex network of inequalities perpetuated by the AI industry. In fact, there may be no straightforward solution to accurately present this landscape.

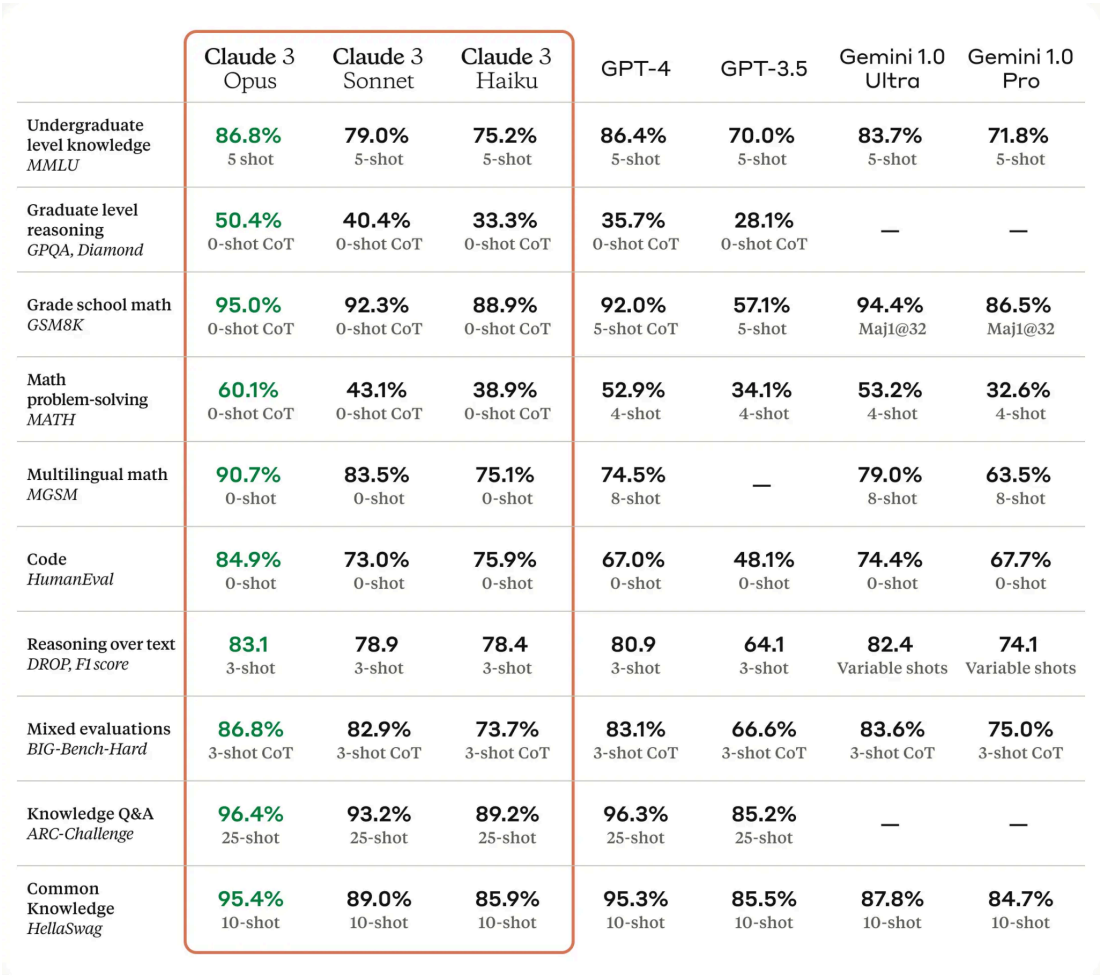
And to emphasize the importance of improving accessibility, companies like OpenAI rely on user feedback in a process called reinforced learning from human feedback (RLHF) to help correct the model's behaviors and address misinformation and biases, in addition to potentially using their conversations to train their future models. When certain demographics are barred from participating in this process, we are also excluding their input in future generations of AIs. In a sense, the training of AIs can be

seen as to some extent a limited democratic process. It goes to show that visibility in the digital world can, in a literal sense, affect whether and which communities benefit from AI technology, and how well they do, underscoring the need for more diverse data practices and alerts us to the implications of invisible digital exclusion and displacement.

For example, if ChatGPT makes mistakes or generates harmful content in English, it may be more likely to be identified and reported in a timely manner. But what if it misinterprets an expression in a more obscure language, like Icelandic, or Punjabi? Or if it has a tendency to provide inaccurate and discriminatory information about a less-known cultural practice in Southeast Asia? There would be far fewer users from these backgrounds who can catch these errors, even fewer willing to submit feedback to the developers. In fact, this has already been found to leave vulnerabilities and security risks, as researchers found it possible to exploit the model's lower "intelligence" by prompting in less common languages to bypass ethical protocols more easily and generate harmful outputs (Yong et al., 2024). The potential exploitation of AI vulnerabilities in lesser-spoken languages is not just a technical oversight but a reflection of deeper societal divides. These vulnerabilities not only leave marginalized communities more vulnerable to misinformation and biased outputs, but also risks perpetuating a cycle of exclusion. A cycle fueled by insufficient data and a lack of feedback from these communities, which in turn leads to continuous underperformance and bias in AI models tailored to these languages and cultures. As we enthuse ourselves with the newest capabilities and potentials of generative AIs, it is crucial to keep in mind that it may further exacerbate the current inequalities in our society.

The substantial costs associated with training, fine-tuning, and testing large language models—owing to the hardware, data, and human resources required—currently remains an insurmountable barrier for most developers, save for those backed up by tech giants, Google and Microsoft and the likes. However, recent findings, such as the Alpaca model from Stanford University's Center for Research on Foundations, suggest that it may be possible to create adequate models at significantly lower cost by fine-tuning open-sourced large models such as Llama from Meta, on AI generated instructions instead of real human feedback (Taori et al., 2023). This approach could enable smaller companies or even individuals to leverage open sourced models, albeit with limited performance compared to the most advanced models. However, there is still concern over whether models trained this way can truly understand nuanced human contexts, beyond the typical benchmarks that only examine a model’s logical reasoning capabilities. It risks creating feedback loops with AI feeding itself hidden biases from the AI that generated the instructions. This also leaves researchers reliant on the generosity of the industry regarding the open-sourcing of their models, if they choose to do so at all. Unfortunately, the direction which we are headed towards appears to give way to heightened competition over cooperation. To quote OpenAI's GPT-4 Technical Paper: "Given both the competitive landscape and the safety implications of large-scale models like GPT-4, this report contains no further details about the architecture (including model size), hardware, training compute, dataset construction, training method, or similar" (OpenAI, 2023). This sets a precedent for the industry where others may choose to follow. While these safety concerns are indeed valid due to high risks of AI misuse, it's

worth noting that this lack of transparency may hinder academics “outside the loop" from engaging with the conversation, at a time when critical interdisciplinary perspectives are much needed.



| | Claude 3 Opus | Claude 3 Sonnet | Claude 3 Haiku | GPT-4 | GPT-3.5 | Gemini 1.0 Ultra | Gemini 1.0 Pro |
|---|---|---|---|---|---|---|---|
| Undergraduate level knowledge *MMLU* | 86.8% 5 shot | 79.0% 5-shot | 75.2% 5-shot | 86.4% 5-shot | 70.0% 5-shot | 83.7% 5-shot | 71.8% 5-shot |
| Graduate level reasoning *GPQA, Diamond* | 50.4% 0-shot CoT | 40.4% 0-shot CoT | 33.3% 0-shot CoT | 35.7% 0-shot CoT | 28.1% 0-shot CoT | — | — |
| Grade school math *GSM8K* | 95.0% 0-shot CoT | 92.3% 0-shot CoT | 88.9% 0-shot CoT | 92.0% 5-shot CoT | 57.1% 5-shot | 94.4% Maj1@32 | 86.5% Maj1@32 |
| Math problem-solving *MATH* | 60.1% 0-shot CoT | 43.1% 0-shot CoT | 38.9% 0-shot CoT | 52.9% 4-shot | 34.1% 4-shot | 53.2% 4-shot | 32.6% 4-shot |
| Multilingual math *MGSM* | 90.7% 0-shot | 83.5% 0-shot | 75.1% 0-shot | 74.5% 8-shot | — | 79.0% 8-shot | 63.5% 8-shot |
| Code *HumanEval* | 84.9% 0-shot | 73.0% 0-shot | 75.9% 0-shot | 67.0% 0-shot | 48.1% 0-shot | 74.4% 0-shot | 67.7% 0-shot |
| Reasoning over text *DROP, F1 score* | 83.1 3-shot | 78.9 3-shot | 78.4 3-shot | 80.9 3-shot | 64.1 3-shot | 82.4 Variable shots | 74.1 Variable shots |
| Mixed evaluations *BIG-Bench-Hard* | 86.8% 3-shot CoT | 82.9% 3-shot CoT | 73.7% 3-shot CoT | 83.1% 3-shot CoT | 66.6% 3-shot CoT | 83.6% 3-shot CoT | 75.0% 3-shot CoT |
| Knowledge Q&A *ARC-Challenge* | 96.4% 25-shot | 93.2% 25-shot | 89.2% 25-shot | 96.3% 25-shot | 85.2% 25-shot | — | — |
| Common Knowledge *HellaSwag* | 95.4% 10-shot | 89.0% 10-shot | 85.9% 10-shot | 95.3% 10-shot | 85.5% 10-shot | 87.8% 10-shot | 84.7% 10-shot |

(Figure 3, an example benchmark score provided by Anthropic to promote their Claude 3 model)

**Conclusion**

In this paper, I have examined the disparities in accessibility to OpenAI's premium subscription service across different countries and explored the potential implications of this inequality. Although this case study focused on using ChatGPT to exemplify the

generative AI industry, the trends we have uncovered can often be seen in other AI assistant services such as NotionAI, Bard/Gemini or Claude. The Bard preview for example was initially only offered to users in the U.S. and the UK at its launch, with its conversations restricted to English, despite the model itself having multilingual capabilities (Hsiao and Collins, 2023). These practices, even if they are rooted in practical considerations for legal reasons or out of ethical considerations to prevent misuse of AI generated content, can still contribute to accelerating global inequality. These are the results of circumstances deeply rooted in global inequality and will persist or even risk intensifying without swift and dedicated efforts to address them. The same technology that divides us has ample potential to bridge inequalities and help marginalized communities instead, for example, the Icelandic government's partnership with OpenAI to preserve the Icelandic language using GPT-4 shows potential in leveraging large language models in preserving endangered cultures and languages. If AI is to benefit humanity as a whole, communities and populations from all backgrounds and communities should have an opportunity to participate in the production of knowledge.

At the end of this paper, I want to reiterate my stance that the topic of accessible, fair AI is a pressing concern that calls for our collective attention and action. What I have touched upon in this paper is only the tip of the iceberg, much more remains to be examined. With that said, I wish to call upon the wider academic community to lend their expertise and perspectives to this subject.

**References**

Beaunoyer, E., Dupéré, S., & Guitton, M. (2020). Covid-19 and digital inequalities: reciprocal impacts and mitigation strategies. Computers in Human Behavior, 111, 106424. https://doi.org/10.1016/j.chb.2020.106424

Benj Edwards. (2023, February 17). Microsoft “lobotomized” AI-powered bing chat, and its fans aren’t happy. Ars Technica. https://arstechnica.com/information-technology/2023/02/microsoft-lobotomized-ai-powered-bing-chat-and-its-fans-arent-happy

Bergvall‐Kåreborn, B., & Howcroft, D.. (2014). Amazon Mechanical Turk and the commodification of labour. New Technology, Work and Employment, 29(3), 213–223. https://doi.org/10.1111/ntwe.12038

Bhaduri (2019): GeoAI: spatially explicit artificial intelligence techniques for geographic knowledge discovery and beyond, International Journal of Geographical Information Science, DOI:10.1080/13658816.2019.1684500

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., & Zhang, Y. (2023).

Sparks of Artificial General Intelligence: Early experiments with GPT-4. arXiv. https://arxiv.org/abs/2303.12712

Central Intelligence Agency. (n.d.). Central Intelligence Agency. Retrieved April 4, 2023, from https://www.cia.gov/the-world-factbook/field/languages/

Coulter, M., & Bensinger, G. (2023, February 9). Alphabet shares dive after Google AI Chatbot Bard Flubs answer in ad. Reuters. https://www.reuters.com/technology/google-ai-chatbot-bard-offers-inaccurate-information-company-ad-2023-02-08/

Dohmke, T. (2023, March 22). GitHub Copilot X: The AI-powered developer experience. The GitHub Blog. https://github.blog/2023-03-22-github-copilot-x-the-ai-powered-developer-experience/

Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., Baabdullah, A. M., Koohang, A., Raghavan, V., Ahuja, M., Albanna, H., Albashrawi, M. A., Al-Busaidi, A. S., Balakrishnan, J., Barlette, Y., Basu, S., Bose, I., Brooks, L., Buhalis, D., ... Wright, R. (2023). 'So what if ChatGPT wrote it?' Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research,

practice and policy. International Journal of Information Management, 71. https://doi.org/10.1016/j.ijinfomgt.2023.102642

Hadeler, E., Prose, N., & Floyd, L. (2021). Teledermatology: how it is impacting the underserved. Pediatric Dermatology, 38(6), 1597-1600. https://doi.org/10.1111/pde.14838

Hsiao, S., & Collins, E. (2023, March 21). Try Bard and share your feedback. Google. https://blog.google/technology/ai/try-bard/

Jason Pontin, “Artificial Intelligence, With Help From the Humans,” The New York Times, March 25, 2007, sec. Business Day, https://www.nytimes.com/2007/03/25/business/yourmoney/25Stream.html.

Kate Crawford and Vladan Joler, “Anatomy of an AI System: The Amazon Echo As An Anatomical Map of Human Labor, Data and Planetary Resources,” AI Now Institute and Share Lab, (September 7, 2018) https://anatomyof.ai

Kosinski, M. (2023, February 4). Theory of Mind May Have Spontaneously Emerged in Large Language Models. arXiv. https://ar5iv.org/abs/2302.02083

Krzysztof Janowicz, Song Gao, Grant McKenzie, Yingjie Hu & Budhendra

Lythreatis, S., Singh, S. K., & El-Kassar, A.-N. (2022). The digital divide: A review and future research agenda. Technological Forecasting and Social Change., 175. https://doi.org/10.1016/j.techfore.2021.121359

Metz, C., & Grant, N. (2023, February 6). Racing to catch up with chatgpt, Google plans release of its own chatbot. The New York Times. https://www.nytimes.com/2023/02/06/technology/google-bard-ai-chatbot.html

Milmo, D. (2023, February 2). CHATGPT reaches 100 million users two months after launch. The Guardian. https://www.theguardian.com/technology/2023/feb/02/chatgpt-100-million-users-open-ai-fastest-growing-app

Norris, P. (2001). Digital Divide: Civic Engagement, Information Poverty, and the Internet Worldwide (Communication, Society and Politics). Cambridge: Cambridge University Press. doi:10.1017/CBO9781139164887

OpenAI. (n.d.). Retrieved April 4, 2023, from https://platform.openai.com/docs/supported-countries

OpenAI. (2023, March 15). GPT-4 Technical Report. arXiv. https://ar5iv.org/abs/2303.08774

Peng, L., & Zhao, B. (2024). Navigating the ethical landscape behind ChatGPT. Big Data & Society, 11(1). https://doi.org/10.1177/20539517241237488

Rao, D. (2023, March 21). Responsible innovation in the age of Generative AI. Adobe Blog. https://blog.adobe.com/en/publish/2023/03/21/responsible-innovation-age-of-generative-ai#grounded-in-ethics-and-responsibility

Salles, A., Evers, K., & Farisco, M. (2020). Anthropomorphism in ai. AJOB Neuroscience, 11(2), 88–95. https://doi.org/10.1080/21507740.2020.1740350

Shen, Y., Song, K., Tan, X., Li, D., Lu, W., & Zhuang, Y. (2023, March 30). HuggingGPT: Solving AI Tasks with ChatGPT and its Friends in HuggingFace. arXiv. https://ar5iv.org/abs/2303.17580

Spataro, J. (2023, March 20). Introducing Microsoft 365 copilot – your copilot for work. The Official Microsoft Blog. https://blogs.microsoft.com/blog/2023/03/16/introducing-microsoft-365-copilot-your-copilot-for-work/

Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P., & Hashimoto, T. B. (2023, March 13). Alpaca: A Strong, Replicable Instruction-Following Model. Stanford CRFM. https://crfm.stanford.edu/2023/03/13/alpaca.html

van Dis, E. A. M., Bollen, J., Zuidema, W., van Rooij, R., & Bockting, C. L. (2023, February 3). Chatgpt: Five priorities for Research. Nature News. https://www.nature.com/articles/d41586-023-00288-7

World Bank. (2022). World Development Indicators: GDP per capita (current US$). World Bank Group. Retrieved May 29, 2024, from https://data.worldbank.org/indicator/NY.GDP.PCAP.PP.CD

Yong, Z.-X., Menghini, C., & Bach, S. H. (2024, January 27). Low-resource languages Jailbreak GPT-4. arXiv. Preprint. https://arxiv.org/abs/2310.02446

# Reconstruction of Social Inequalities in a Digital Reality


## Abstract

This paper introduces a novel framework called Layers of Realities for analyzing inherited biases in large language models(LLMs) development as a window of insight into deeper societal inequalities issues. I combine lines of thoughts from technological insights in AI research and critical feminist social theories to interpreting LLMs not as mere tools of production, but as spatial embodiments reflecting our society's various inequalities, its stratification and structures of power that replicates themselves in the digital space. In the paper I discuss the intricate interactions and connections between three representations of the world- corresponding to the physical world, internet and LLMs on an increasingly distorted scale: the Physical Reality, the Digitized Reality and the Modeled Reality. By building a framework for analyzing these alternate interpretations of realities as a collective, I call for more comprehensive readings of equity issues in AI ethics, not as isolated issues but as part of larger systems of inequalities that have taken root across multiple layers of realities. Through this I hope to draw attention to the necessity of recognizing the voices of marginalized communities in an increasingly digitized society. In doing so, the paper is not only a theoretical exploration but also a call to action for conversations between AI research and social scientists to look beyond issues of algorithmic biases and work on interdisciplinary efforts to tackle the underlying causes embedded in our social structures.

## Introduction

With the generative AI “arms race” attracting attention across the globe, the idea of AI ethics and AI alignment has become an increasingly prominent topic as many dubbed the crisis of potential uncontrollable AI an “existential crisis” for humanity. In early 2023, following the rise to popularity of ChatGPT, a number of well known scientists, industry leaders, politicians and public figures have been active in alerting the public to the potential threat of these new actors of our society. The first major milestone was a petition calling for a 6 month pause of developing models more powerful than GPT-4, before human legislation and regulations can keep up with the rapid advancements in the field. Later a substantial number of public figures, including reputable scholars, government officials and top executives of the leading AI companies like OpenAI and Anthropic themselves collectively signed an ominous statement: “Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war.” (Center for AI Safety, 2023). OpenAI for example had committed itself to the idea of “AI superalignment”, promising to dedicate part of its resources to ensure that in the event artificial superintelligence that exceeds human intelligence is achieved, it will continue to operate in alignment with human values and welfare. The team, however, has since been disbanded following the resignation of several key scientists. This paper is in part a response to the ongoing discussions around AI ethics and asking the question: who are these “humans” in these statement? Here in this paper, I aim to synthesize the ideas behind AI ethics in large language model training, and ideas from critical social perspectives, to bridge the gap between social sciences critiques and cutting edge technologies, both offering critical

perspectives on the technology itself and abridging technical concepts for social scientists to understand. I create the “layers of realities” framework as a way of understanding the relationship between society, the internet and large language models. In this framework, I argue we should examine LLMs not as mere algorithms or assistants, but as an alternative spatial projection of the internet and by extension our society. This way of thinking about AI biases and ethics provide new perspectives on the visible or invisible inequalities in our society, and highlight struggles for fair digital representation on the internet and to be visible in public research.

**Large Language Models as Spatial Embodiments**

Instead of focusing on the more nuanced technical details, I conceptualize large language models (LLMs) as spatial social models based on foundational techniques behind models such as GPT. In the realm of natural language processing, a method known as word2vec was a popular approach for translating human-readable words into machine-legible vector data. This technique is the predecessor that laid the groundwork for modern large language models, including the now global phenomenon ChatGPT. The method transforms a vast amount of text data as tokens into vectors in a high dimensional vector space, thus the spatial relationships of those words come to represent the semantic relationships between them (Mikolov et al, 2013). In this vector space, tokens can be compared for their similarities and difference in spatial terms, in a similar fashion to Tobler’s first law of geography: closer things are more related to one another. One can anticipate the word “run” to be closer to “walk” than it is to “cook”. Another popular analogy is that taking the word “king”, subtracting “man” from it and

adding “woman” would result in “queen”. With words reduced to mathematical properties, computers and algorithms are thus able to work with and “understand" human languages by running calculations and analyzing the complex patterns and relationships between words.

While word2vec captured simple linear relationships among words, the evolution of deep learning brought about the Transformer architecture, a paradigm shift that significantly enhanced how models handle language. Unlike earlier models that processed words sequentially, Transformers employ a mechanism called "attention" that assesses all words in a text simultaneously. This comprehensive view enables the model to maintain context and produce more relevant outputs, revolutionizing the capabilities of LLMs (Vaswani et al, 2017). From these basic principles and vast amounts of training made possible in modern data centers, these language models are able to read over texts and predict the probability of the next word in a sentence.
The current large language models are not sentient, despite their seemingly intelligent behaviors at times. Even when they generate responses to the user’s input, they are not exactly “conversing” with the user, rather they are simply predicting what a most appropriate response would be. If we take a spatial metaphor, it would be similar to pathfinding in a high dimensional maze given a general direction by the user.
Naturally, this method of understanding the world is inherently imperfect, teeming with assumptions and biases embedded in the data in addition to sometimes the language itself (such as cases where words can have masculine or feminine properties, or grammatical shifts based on the speaker’s social status), despite the best efforts to

mitigate them. But from a different angle, it can also be argued that if we take into account the amount of data and context involved in the process, they can be considered (imperfect) miniature replicas of our society, at least the part of the world that has been digitized.

Advanced language models today have achieved exceptional performance beyond what even their developers have anticipated. Many models have exhibited unexpected capabilities that hints of possible logical capability. Some of the current flagship models, for example the GPT-4 developed by OpenAI, are considered by many researchers as showing potential that artificial super intelligence that exceeds human level intelligence is possible (Bubeck et al., 2023). With AI becoming more complex and human-like, it is becoming a trend for AI related research to be interdisciplinary. Some scholars are suggesting we can benefit from looking at AI from the perspectives of psychology and neural sciences (Kosinski, 2023). In contrast to those employing interdisciplinary approaches to understand the models themselves, I will explore how the models when viewed through interdisciplinary perspectives can help us understand our own society. Through viewing LLMs as spatial embodiments of our social constructions, we can reflect on how concepts such as gender, race, etc. are constructed and perceived by our society, and how we can rethink our relationship with our identities and the world.

**Digital Twins and Social Reproduction of Inequalities**

The concept of the digital twin has gained significant attention in recent years and has attracted extensive research and definitions. A digital twin can be defined as a comprehensive digital description of a product that can simulate the behavioral

characteristics of a realistic model (He, 2023). It involves the integration of data between a physical and virtual machine in either direction (Fuller et al., 2020). Furthermore, it can be described as a realistic digital representation of something physical (Cooper et al., 2022). The evolution of digital twin technology has been marked by the convergence of advanced machine-learning algorithms, data analytics, visualization, modeling, and simulation techniques. This convergence has paved the way for the development of sophisticated digital twin systems capable of bridging the physical and virtual realms seamlessly (Jeong et al., 2022).

However, here in this paper I would like to extend the definition of digital twins to represent something more than physical- just as digital twins can be used to simulate physical objects, for planning smart cities or modeling assembly lines, it has the potential to be used for modeling our society in an abstract sense. In this alternative interpretation of the digital twin, the average internet user is acting as unconscious sensors in their everyday life, sending a steady stream of data to the internet by means of social media updates or forum posts when they share their knowledge and experiences. This information, while qualitative and highly subjective, when aggregated on a massive scale like in the case of LLMs, can become a loose physical representation of the proceedings and history of our society in vector space. This digital representation of our society is not a digital twin in the sense of a simulation system for real time climate or traffic data, but one that simulates the consensus of mainstream society.

The projection of the real world into digital space is inherently a lossy conversion. At any given time, only a fraction of the entirety of the vast repository that is humanity's rich history and diverse cultures is available in the form of digital datasets, ready for use in the training. Even fewer data can make it through the selective and biased curation process to become training materials for the LLMs. What is considered valuable enough to become digitized by those with the power and privileges to do so, and what is then considered high quality data to be used in model training, every step along with the way is teeming with subjective decisions based on assumptions, biases and convenience of the decision maker's background and identity. I view this as an application of the social reproduction theory, often used by marxist feminist scholars in their analysis, which states that society perpetuates itself not only to ensure its continuity, but also reproduces its systemic inequalities and mechanisms of oppression (Rodríguez-Rocha, 2021). In the digital context, this means that biases, privileges and inequalities in the real world are mirrored and magnified in digital space. They seep into AI models in the form of bits and pieces of biased data as they get picked up by web scrapers, in a similar fashion to how microplastics infiltrate the ecosystem and find their way up the food chain, until they eventually turn up at our dining tables. As another example, developers do not need to deliberately train a LLM to understand who Harry Potter is or the plot of the series, when pieces of the information can be found throughout the internet and end up in the training data inevitably. Instead, given how popular the series is, researchers are finding that it is incredibly difficult to try and make a model "forget" about the series(Eldan & Russinovich, 2023). And thus, far from being a faithful replica of the real world, what this results in is a warped digital reflection of our world that

greatly distorts and overemphasizes certain experiences and perspectives over others, creating what might be considered a parallel or alternate reality. A reality as experienced and perceived by the privileged members of society.

**Data Feminism and Defining Humanity**

It is well known in the machine learning community that algorithmic bias is produced in a "garbage in, garbage out" fashion, meaning that training on biased data results in biased algorithms. LLMs are not immune to this law, as they can often make biased assumptions especially visible when producing images. This is also widely known in computer vision, as there is already extensive criticism both across mainstream media and within academia on computer vision algorithms' biases in identifying and categorizing people of color, especially for black women who suffers at the intersection of racial and gendered discriminations (Buolamwini & Gebru, 2018). These kinds of algorithmic biases have been extensively discussed in ideas like data feminism which challenges the binary nature of algorithmic classification and advocates for embracing pluralism in AI technology (D'ignazio and Klein, 2020). While for smaller models a seemingly obvious solution is to improve data quality and diversity by spending additional resources on curating the training data and ensuring a fair representation of different identities and communities, this becomes increasingly less practical when it comes to training LLMs due to their "hunger" for data. As the performance of LLMs are heavily reliant on their size and scale, the required datasets to train new models have grown to a scale that makes extensive human vetting and supervision unrealistic.

These nuanced and overlapping factors come together to paint a picture of “inevitable” bias in the training of LLMs. While that is no reason to discontinue the ongoing effort to minimize and mitigate their implications, it should challenge us to confront the historical biases and systemic inequalities that lead up their embedding in society. To be transparent and honest about them, rather than to “fix” the issue with performative actions, as Google have been criticized in recent times for Gemini’s image-generation tool producing historically inaccurate images such as “ethnically diverse” German soldiers during WW2, when they implemented low-effort performative measures to force the model to diversify gendered and racial expressions as an attempt to conceal the underlying biases in the algorithm (Milmo, 2024). In a sense, the internet and by extension, large language models are extensions of the social mechanisms that reproduce and amplify existing inequalities. Not only do they assist in perpetrating biased and discriminatory narratives, they also play a part in suppressing the voices of marginalized groups.

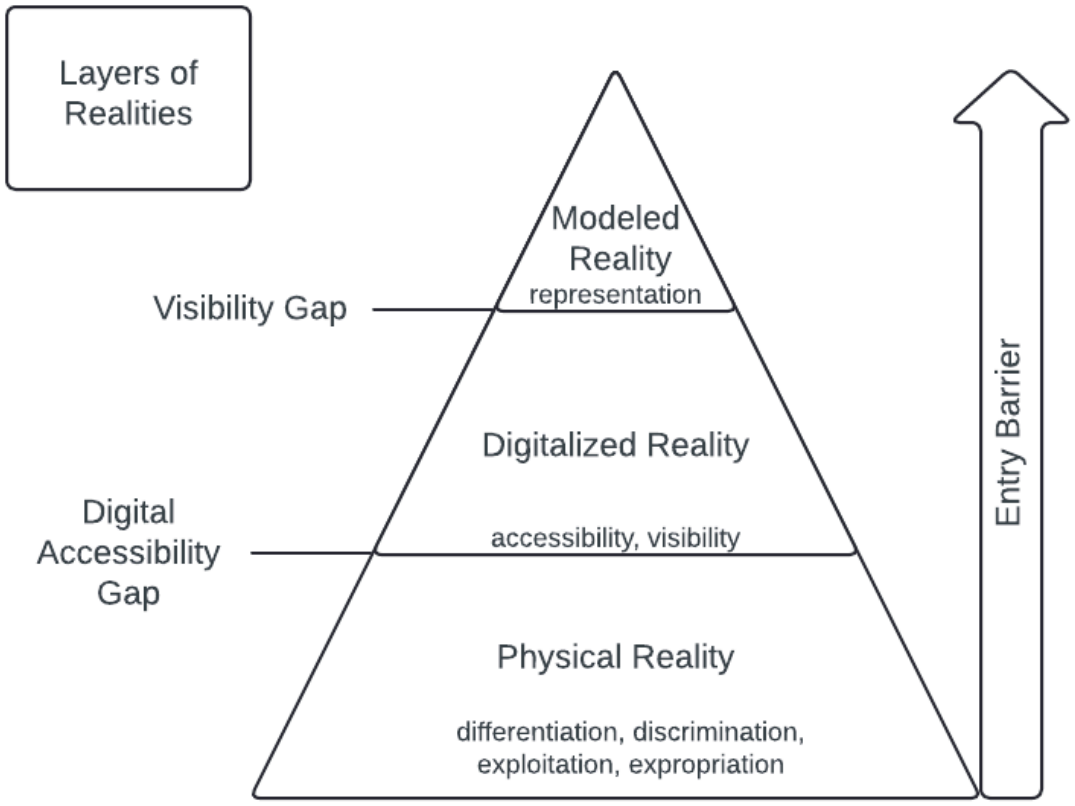
The reproduction of social stratification raises another alarming question: who is the “human” in the context of the AI community’s common goal of ensuring AI alignment with "human values". My thinking here is influenced by decolonial feminist scholarship and critiques of the idea of a universal image of humanity. When human values and experiences across cultures and societies are so diverse and often contradicting one another, we must be wary of potentially allowing the idea of being human to be defined and reduced to the experience of a small group of privileged demographics, as it had often been the case throughout history. A long line of critiques have Decolonial critiques

from scholars such as Sylvia Wynter point out the hypocrisy of the construction of the image of the “rational man” as a tool of European colonizers to differentiate themselves from the indigenous and African descent populations as victims of their colonization (Wynter, 1995). As pinned by works of Silvia Federici in her historical analysis of witch hunting and the stigmatization, exploitation and expropriation of women’s bodies and labor, the same logic of “difference making” creates conditions that divide, dehumanize and institutionalize the exclusion and discrimination against populations of marginalized identities. (Federici, 2004, Collard & Dempsey, 2018). If we allow the idea of human value to be reduced to a single, universal rule, then we risk the erasure of marginalized groups and identities. It is for this reason I raise the question of “what does it mean to be human” within the layers of realities framework, not only as a way of challenging the AI industry’s power in defining humanity, but also to invite reflections on our perception of social identities compared to the classification methods of an AI. inviting us to reflect and act on the disparities in our own society as we would intervene in the case of a biased AI.

**The Layers of Realities Framework**

The “Layers of Realities” theory is a framework I conceptualized building on top of the aforementioned theoretical perspectives framing LLMs as representations of alternate realities to the real world. Here I also introduce the internet as an intermediate layer between the digital twins of society and LLMs. The three coexist as mirrors of one another, in a pair of “digital triplets”. The three alternate perceptions of realities as represented by each of the triplet, which I name the Physical Reality, the Digitized

Reality and the Modeled Reality, exist as layers built on top of the one another, becoming increasingly warped and distorted in the process. Figure 4 below offers an overview of the structure of the layers of realities highlighting the key struggles and barriers between each reality.



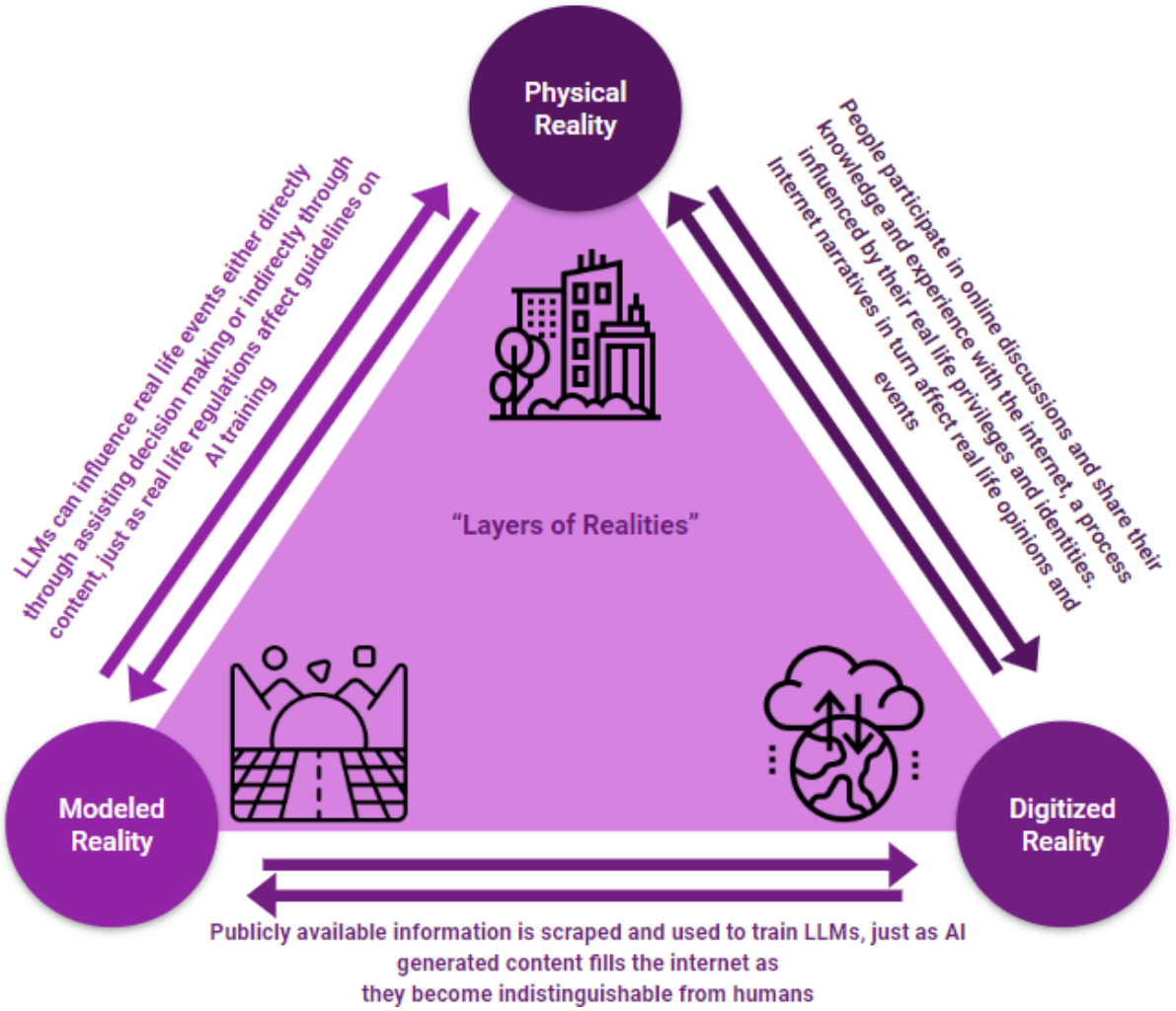

Physical Reality encompasses the tangible world where human interactions and experiences take place. This layer of reality needless to say has the lowest entry barrier, all people who exist and have existed have been part of the physical reality. It is also the most data-rich layer and the foundation of all other realities. Without our physical existence, our technologies behind the internet and LLMs would not exist. Yet while it is accessible to all, it is riddled with all manners of social inequalities, exploitations,

oppressions and strife. It is here that the politics and power dynamics shape the layout of layers above. For those privileged enough to be unaffected by the digital accessibility gap and have access to the internet, their devices act as portals to access the next layer of reality: Digitized Reality. Digitized Reality embodies the collective lived experiences of all those who have had access to the internet and spent time contributing their knowledge to it. While this realm resembles Physical Reality's complexities, Digitized Reality is not strictly a derivative of Physical Reality. It has its own set of rules to follow that can sometimes result in groups who are marginalized in Physical Reality to find empowerment instead. While in the Physical Reality, exploitations and oppressions may take on more explicit, physical form, here in the digital space, that theme becomes replaced by the power struggles of control over discourse and narratives, of fighting the recommendation algorithms for publicity and visibility. Not every community's voice can be heard on mainstream sites, be given room on the front page and trending page of people's social media feeds, or be portrayed in a fair light.

From Digitized Reality is born the Modeled Reality as a projection of a projection. I frame it as such because just like how the Digitized Reality is an oversimplification of Physical Reality, no matter how much data is aggregated for LLM training, compared to the entirety of the internet, it is still a very small amount of cherry-picked experiences and perspectives. In a similar fashion to how barriers and privileges in the Physical Reality affect who have access to the Digitized Reality, the stratification on the internet, the disproportionate visibility and representation of communities, and LLM developers'

conscious or unconscious assumptions in data curation procedures give rise to further distortions of this reality. A prime example of how discourse power shapes this landscape is that popular elements of western culture like Harry Potter do not need any explicit effort on the developers' end to be included in a model's knowledge base. They simply make their way into the Modeled Reality through sheer popularity and widespread presence in the Digitized Reality. In fact, research has shown that it is even a hassle to try and remove such knowledge from a trained model (Eldan & Russinovich, 2023).

While the layers of realities are built on top of one another, their dynamic is not strictly a hierarchical one as these realities exist not in parallel but in an intertwined and overlapping fashion. The influence of the power dynamics is not simply a one-way determination as we go from the physical to the virtual, but in a cyclical fashion. I argue that society, internet and LLMs should be viewed as a collective coexistence, critiques and observations over one should not take place without considering the others. In figure 5, I present examples of how the interactions and power exchanges might take place between these layers of realities.




The major contribution of the "Layers of Realities" framework lies in its capacity to provide a nuanced understanding of AI ethics and the broader social issues of exclusion and marginalization. It challenges us to recognize that these equity issues are not merely a product of any particular layer alone, but part of a larger construct that has taken root across all layers of reality at once, replicating itself akin to that of the mythical beast hydra, where one head gets cut off, new heads grow out. The layers of realities framework provides the theoretical foundation for looking beyond simply addressing issues like the struggles against unfair (under/mis)representation in our media, bridging

the digital accessibility gap or mitigating AI biases as isolated issues, it urges the combination of technological insight with critical social theories and calls for coordinated interdisciplinary actions. We must learn to take into account the context of all layers of realities at once, in a coincidental fashion to the attention mechanism that revolutionized the deep learning technique.

**Conclusion**

The main goal of this paper is to synthesize critical feminist theories and AI alignment ethics in producing a framework that helps us conceptualize and understand the implications of large language models on our society, as well as the society itself. In the Layers of Realities framework I have laid out the necessary foundations for assessing struggles of social equity and biases not just as isolated incidents, but taking into account broader contexts. In doing so, I hope to not only facilitate further conversations between AI ethics research and the more radical social theories, but also to inspire explorations in creative methodologies that utilize the unreliable and undesirable AI biases and turn them into useful analytical tools. As the stakes of digital inequality continue to rise in the aftermath of the Covid-19 pandemic, the necessity of acknowledging and confronting new forms of visible and invisible gaps and barriers brought forth by emerging new technologies will only become more pressing if our society continues down its current trajectory of digitization. As we move forward, the Layers of Realities framework offers a valuable foundation for ongoing discussions about AI ethics. It compels us to reflect on how AI shapes and is shaped by societal forces, and how it becomes an intricate force shaping our society in turn. I hope that in

the near future, the groundwork laid by paper will serve as the basis for me or other scholars to flesh out the framework and serve as guidance for responsible development and implementation of AI technologies.

**References**


Buolamwini, J. & Gebru, T.. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, in *Proceedings of Machine Learning Research* 81:77-91 Available from https://proceedings.mlr.press/v81/buolamwini18a.html.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. Science (New York, N.Y.), 356(6334), 183–186. https://doi.org/10.1126/science.aal4230

Center for AI Safety. (2023) Statement on AI Risk. Retrieved from https://www.safe.ai/statement-on-ai-risk

Collard, R.-C., & Dempsey, J. (2017). Politics of devaluation. Dialogues in Human Geography, 7(3), 314–318. https://doi.org/10.1177/2043820617736602

Collard, R.-C., & Dempsey, J. (2018). Accumulation by difference-making: An anthropocene story, starring witches. Gender, Place & Culture, 25(9), 1349–1364. https://doi.org/10.1080/0966369X.2018.15213

D’Ignazio, C. and Klein, L. (2020). Data feminism..
https://doi.org/10.7551/mitpress/11805.001.0001

Eldan, R., & Russinovich, M. (2023). Who’s Harry Potter? approximate unlearning in llms. arXiv.org. https://arxiv.org/abs/2310.02238

Federici, Sylvia. 2004. Caliban and the Witch: Women, the Body and Primitive Accumulation. New York: Autonomedia.

Federici, S. (2019). Social reproduction theory. Radical Philosophy, 2(04), 55–57.
Sartori, L., & Bocca, G. (2022). Minding the gap(s): public perceptions of AI and socio-technical imaginaries. AI & SOCIETY, 38, 443-458.

Fuller, A., Fan, Z., Day, C., & Barlow, C. (2020). Digital twin: enabling technologies, challenges and open research. Ieee Access, 8, 108952-108971. https://doi.org/10.1109/access.2020.2998358

Jones, D., Snider, C., Nassehi, A., Yon, J., & Hicks, B. (2020). Characterising the digital twin: a systematic literature review. Cirp Journal of Manufacturing Science and Technology, 29, 36-52. https://doi.org/10.1016/j.cirpj.2020.02.002

Mikolov, T., Chen, K., Corrado, G.S., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. International Conference on Learning Representations.

Milmo, D. (2024, February 22). Google pauses AI-generated images of people after ethnicity criticism. The Guardian. https://www.theguardian.com/technology/2024/feb/22/google-pauses-ai-generated-images-of-people-after-ethnicity-criticism

Nguyen, M., Hargittai, E., & Marler, W. (2021). Digital inequality in communication during a time of physical distancing: the case of covid-19. Computers in Human Behavior, 120, 106717. https://doi.org/10.1016/j.chb.2021.106717

Noble, S. U., & Tynes, B. M. (2016). The Intersectional Internet: Race, Sex, Class, and Culture Online (2nd ed.). Peter Lang International Academic Publishers.

Peng, L., & Zhao, B. (2024). Navigating the ethical landscape behind ChatGPT. Big Data & Society, 11(1). https://doi.org/10.1177/20539517241237488

Ragnedda, M., & Muschert, G.W. (Eds.). (2013). The Digital Divide: The Internet and Social Inequality in International Perspective (1st ed.). Routledge. https://doi.org/10.4324/9780203069769

Rodríguez-Rocha, V. (2021). Social Reproduction Theory: State of the field and new directions in geography. Geography Compass, 15(8), e12586. https://doi.org/10.1111/gec3.12586

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, California, USA. https://dl.acm.org/doi/10.5555/3295222.3295349

Yu, S., Qian, C., & Pickard, S. (2021). Age-related digital divide during the covid-19 pandemic in china. International Journal of Environmental Research and Public Health, 18(21), 11285. https://doi.org/10.3390/ijerph182111285

## Acknowledgements

I would like to express my gratitude for the many amazing people in my life who have helped and supported me along the way. Beginning with my mom, Lijuan Zeng, without whose support I would not have had the opportunity to pursue my education in this country. A strong, resourceful and empowering woman, she had overcome many unimaginable obstacles and made incredible sacrifices to make this a reality. I am truly grateful for the privilege of being here today. She has, and always will be an inspiration and role model to me.

I would also like to thank my committee members, Bo Zhao and Mia Bennett, who helped me develop and polish my ideas in addition to introducing me to new

concepts, theories and literature over the last two years. I am especially grateful for Bo as my advisor and committee chair, who had encouraged me to further my education towards the end of my undergraduate years, when at the height of the Covid-19 pandemic everything about the future seemed uncertain, and I felt very lost. I'm glad to be part of the UW Geography department. Speaking of which, I want to thank all the faculty, staff, and fellow graduate students of this department who have made it a supportive and welcoming community with your collective presence.

I want to say thank you to the Aragon family, to Diana, Cecilia, Dave, Jamie and Ken for being my family here, supporting me through some difficult times in my life where my biological family could not. For giving me a home away from home and a reason to stay in Seattle. It would not have been the same without you.

And last but not least, a tribute in memory of Hayley Summer Smith. I still can't believe how long it's been without you. Words alone are insufficient to describe how much I love and miss you. You have always inspired me to do more good and to be a better person, to love, to care. I hope I'm making you proud.