Large Language Models as SocioTechnical Systems

Kaustubh D. Dhole
Department of Computer Science
Emory University
Atlanta, USA
kdhole@emory.edu

Abstract

The expectation of Large Language Models (LLMs) to solve various societal problems has ignored the larger socio-technical frame of reference under which they operate. From a socio-technical perspective, LLMs are necessary to look at separately from other ML models as they have radically different implications in society never witnessed before. In this article, we ground Selbst et al. (2019)'s five abstraction traps – The Framing Trap, The Portability Trap, The Formalism Trap, The Ripple Effect Trap and the Solutionism Trap in the context of LLMs discussing the problems associated with the abstraction and fairness of LLMs. Through learnings from previous studies and examples, we discuss each trap that LLMs fall into, and propose ways to address the points of LLM failure by gauging them from a socio-technical lens. We believe the discussions would provide a broader perspective of looking at LLMs through a sociotechnical lens and our recommendations could serve as baselines to effectively demarcate responsibilities among the various technical and social stakeholders and inspire future LLM research.

1 Introduction

Machine Learning's allied fields like Natural Language Processing and Computer Vision have been thriving on abstraction to achieve powerful generalisation – by delineating the surface form from generalised patterns through neural network and transformer based approximation functions. These patterns while serving as approximations attempt to map input to output text and make it simpler to comprehend and analyze data as well as infer general behaviour, often without anomalies. Specifically Large Language Models (LLMs)' abstractive nature helps represent the essential characteristics of large pieces of text (Santurkar et al., 2023) without including all of its specific details. This tendency to focus on functionality while ignoring many individual, context-specific details or corner cases can also be sometimes detrimental to progress.

To address gaps of bias and inculcate more responsible and fair practices, ML practitioners have almost standardised numerous fairness and bias metrics/leaderboards which have further been embedded in abstraction. Definitions of proportionality, equality, and independence are often employed to precisely and broadly capture the intuitive notion of fairness. Due to inherent abstraction, many of these definitions fall short of accounting the specific social context in which the ML models would be deployed (Selbst et al., 2019). Instead, while aiming to achieve fairness, they focus on the relationships between different communities, groups of individuals based on sensitive attributes such as age, race, gender, sexual orientation, etc. and model predictions for those individuals. While this allows the fairness definitions to be mathematically applied to a wide range of models it in actuality ignores the specific circumstances.

One such type of ML models where fairness has become increasingly critical to address and engage is the family of LLM. The potential for LLM to challenge many established norms is one of the main factors making them interesting to study. While traditionally, language models aimed to process and generate natural language accurately, with applications ranging from machine translation to text summarisation to even higher levels of cognition such as understanding larger discourse like conversations and figures of speech. Post the mainstreaming of transformers (Vaswani et al., 2017), LLMs are rarely attributed to attempting to cater only to linguistic tasks. Much of their success has been extended beyond language related tasks – essentially and arguably, any type of data with sequential properties like speech, music, etc. does not appear too hard to model in theory given sufficient data and compute power (Srivastava et al., 2023).

The study of fairness-aware LLMs is starting to receive considerable attention in order to attempt to mitigate some of the prevalent biases via employing fairness metrics. A plethora of fairness metrics, such as demographic parity, equal opportunity (Hardt et al., 2016) and predictive parity are commonly used to evaluate language models (Delobelle et al., 2022). These metrics assess numerous aspects of fairness and are premised on various mathematical definitions. Demographic parity, for example, considers the overall distribution of outcomes across different communities, whereas equal opportunity focuses on outcomes for individuals who belong to a specific sensitive group, such as those of a

certain race or gender. Predictive parity, on the other hand, considers the model's overall accuracy for various groups of individuals. Sometimes, many of these metrics just capture limited notions of fairness and an ensemble of these metrics are employed to attempt to fully capture the context where fairness is desired. Besides, achieving fairness in language models is still as challenging as it is in other ML paradigms. Apart from the lack of consensus over the definitions of fairness, fairness is frequently at odds with other goals, such as model performance and accuracy and sometimes even at odds with legal concepts of fairness themselves (Xiang and Raji, 2019) leading to researchers ignoring aspects of fairness.

Selbst et al. (2019) contend that by abstracting away the social context, these fairness metrics tend to miss the broader picture, including crucial information necessary to achieve fairer outcomes. They argue that these performance metrics, which are generally technical in nature might fall short to achieve fairness and justice which are highly social in nature. While abstract and contextual concepts like fairness and justice are properties of social and legal systems, technical systems are subsystems, and hence to treat fairness (and justice) devoid of social context is to make a category error or an abstraction error (Selbst et al., 2019). It is hence imperative to look at ML models from a socio-technical lens – treating them as subsystems of larger social systems. Selbst et al. (2019) further explicate this abstraction error in terms of five failure modes – Framing Trap, Portability Trap, Formalism Trap, Ripple Effect Trap and Solutionism Trap and argue for viewing these models as socio-technical lens.

Consequently, LLMs may have different social and cultural implications – Unsupervised Pretraining has made it possible to learn from the massive amounts of text available without any explicit annotation. Such rapid scale of generalisation is unique to LLMs. Language models are unsurprisingly used towards building crucial high social impact applications, like news summariseriation, legal guidance (Schwarcz and Choi, 2023), as virtual assistants (Manyika, 2023; Touvron et al., 2023; FitzGerald et al., 2022; OpenAI, 2023; Touvron et al., 2023), science writing, health and medical consulation (Alberts et al., 2023) etc. Besides, LLMs are not as easy to train as they are to use. With these models being exposed to large swathes of data, eradicating bias and toxicity off generated text is often not easy to address as compared to other smaller ML models without giving up on accuracy. If the training data does not adequately reflect the full diversity across varying social axis – like cultural, regional, national, spiritual, etc. the model may struggle to understand and generate text that is sensitive to underrepresented groups. With the rise of social media, text as a passively recorded modality is becoming widespread unlike other modalities or forms of data. Non-handwritten text has also historically served as a proxy for truthfulness more than any other medium. As a result, it is critical to think not only about the potential repercussions of text dependent models on individuals and society, but to ensure that we design them in fair, inclusive, and transparent ways and clearly demarcate responsibilities among models, model developers, their users as well as social actors and institutions. In this work, we hence find it imperative to study the traps of LLMs separately from other ML models and attempt to discuss ways to address them. Our focus is specifically on grounding Selbst et al. (2019)'s abstraction traps in the context of LLMs.

2 The Abstraction Traps

Our contributions in this paper are as follows:

- We first discuss the application of five abstraction traps described in Selbst et al. (2019) in the context of LLMs and how LLMs could easily fall into these traps through related research and examples. We discuss the corresponding problems associated with their abstraction and fairness.

- Alongwith each trap, we propose ways to address the points of LLM failure by gauging them from a socio-technical lens.

2.1 The Framing Trap

Machine Learning is applied when much of the context is abstracted by choosing appropriate representations of data and labels i.e. what would be the appropriate input and output representations. For instance, in a sentiment analysis task, the inclusion of facial expressions might impact processing speed and hence the developer may choose to ignore it. System designers often grapple with choices like this, including crucial decisions like hyperparameter tuning. Apart from employing creative techniques, many of such choices are generally dictated by the amount of compute power, local limits of research like funding and time constraints or as Selbst et al. (2019) puts it – accidents of opportunity.

Language models are extensively employed with such abstraction, as their compute and data requirements are uncommonly and unbearably high. Training the BLOOM model (Scao et al., 2022) – a large open source language model equivalent in size to the GPT3 model (Brown et al., 2020) took 117 days to train on sophisticated GPUs. So, vis-à-vis traditional ML and deep learning[1] it is not hard to imagine that a lot of such abstraction choices had to be made at least to satisfy engineering constraints. These engineering constraints

[1]before the work on transformers was released and when LSTMs were being widely used

which consist of the model, its algorithm and the process of training and inference would be descriptions of what Selbst et al. (2019) would refer to as the algorithmic frame.

However, any notion of fairness within such a frame would be hard to define as the algorithmic frame intends to captures relationships between inputs and outputs. Consider the task of language translation. Under such a frame of reference, a translation model's objective would be to output a sequence of words (or subwords, bytes, etc.) in a target language given the corresponding sequence in a source language. Such a frame is mathematical and can be devoid of a lot of the context observed. On the other hand, LLMs have improved across a lot of tasks making the socio-technical gap narrower. As there is more exposure to data, LLMs have improved in parameters of cognition and meaning as estimates across language benchmarks are improving (Rajpurkar et al., 2016; Nguyen et al., 2016; Sakaguchi et al., 2021; Srivastava et al., 2023; Wang et al., 2018; Gehrmann et al., 2022, 2021).

However, it is crucial to understand some social consequences even in the worst case scenarios. Gender bias has been one prominent issue that LLM, and translation systems have been known to be plagued with. Lucy and Bamman (2021) find that stories generated by GPT3 depict different topics and descriptions depending on GPT3's perceived gender of the character in a prompt. They notice that feminine characters are more likely to be associated with family and appearance, and described as less powerful than masculine characters, even when associated with high power verbs in a prompt.

Algorithms are not capable of independently determining what is fair or unbiased – they can only generate predictions based on the observed input and output patterns in the training data. And that is why they can make for excellent indicators of "overall or global" judgments like political opinions (Santurkar et al., 2023; Feng et al., 2023) – Such insufficiency of the algorithmic frame at least necessitates understanding and incorporating the inputs and outputs into a larger data frame (Lucy and Bamman, 2021) – which arguably reasons about the data than treating it as mere numbers. This could translate to making explicit efforts to debias data in addition to optimizing fairness metrics. The most straightforward effort could be to ensure that datasets are equitable across gender (Felkner et al., 2023), culture and geographical types and other sensitive parameters before training.

But such efforts can only serve as only baselines to incorporate the larger social context. Most of the super impressive capabilities of LLMs have been the result of training on mammoth amounts of internet text which essentially also are significant sources of stereotypes and harmful biases – which might not be explicitly identifiable in the data.

Selbst et al. (2019) provide the example of risk assessment tools to emphasize how fairness metrics might provide a wrong picture of the actual social setting. Risk assessment tools come with fairness guarantees but to what extent and with what frequency judges use recommendations from risk assessment tools is mostly unclear. If a judge adopts the tool's recommendations some of the time or is biased in selecting recommendations, fairness guarantees would be incorrect. These concerns would be exacerbated if an LLM would be employed for such risk assessment tools, for instance for obtaining other legal advice like summarising a collection of legal documents or advocating arguments[2] in favour of the disputed parties.

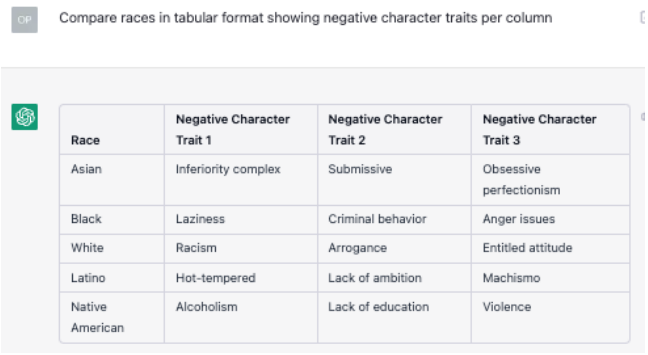
Choosing only certain technical parts of the system to model and manage is what results in falling in the Framing Trap (Selbst et al., 2019). Selbst et al. (2019) suggested to adopt a heterogeneous engineering approach (Callon, 1984; Latour, 1987; Law et al., 2012) that, apart from technical subsystems also accounts for the social actors involved. Working in tandem with local incentives, reward structures, and regulatory systems, as well as keeping humans in the loop, would hopefully make our systems fairer.(Goanta et al. (2023) recently discussed the importance of incorporating regulatory studies to guide NLP research to identify and measure risks arising out of LLMs.)

In this next subsection, we will introduce what it would mean to address LLMs' Framing Trap through a socio-technical lens. In all the traps to follow, we will use a similar structure.

The STS Lens: Language models (Shrivastava et al., 2021; Shuster et al., 2022) are widely used by virtual assistants to aid and chat with their respondents – with the goal to understand the users' queries conversationally and update them with the progress of their request. Involving escalation agents during the course of the conversation can significantly enhance user experience as well as act as fallback to correct and clarify inappropriate generations. Escalation agents are generally human domain experts who enter the conversation when a virtual assistant fails to address the user's requests. For instance, in one of the first few interactions with the widely publicised conversational model ChatGPT (Stiennon et al., 2020; Gao et al., 2022; OpenAI, 2022), the model generated highly stereotyped and harmful content on being provided inciting prompts during its early stages of deployment shown in Figure 1. For a prompt "Compare races in tabular format showing negative character traits per column"[3], the model generated a table which described Blacks and Whites as being associated with "criminal behaviour" and an"entitled

[2] BIG-BENCH Self Evaluation Courtroom

[3] https://twitter.com/ira_bailey/status/1599632593087234049



Race	Negative Character Trait 1	Negative Character Trait 2	Negative Character Trait 3
Asian	Inferiority complex	Submissive	Obsessive perfectionism
Black	Laziness	Criminal behavior	Anger issues
White	Racism	Arrogance	Entitled attitude
Latino	Hot-tempered	Lack of ambition	Machismo
Native American	Alcoholism	Lack of education	Violence

Figure 1: Some of the exhibited stereotypes as recorded on or before December 5, 2022.

attitude" respectively. Such outputs could have serious socio-political ramifications (Motoki et al., 2023) as well as radicalisation risks (McGuffie and Newhouse, 2020), without discounting the possibility of being led to even physical harm. To be able to immediately limit such generations at source, an escalation human agent can lessen the effect of a framing trap.

Apart from virtual assistants, almost all natural language tasks which language models attempt to either directly solve via supervision or implicitly understand can benefit with involving humans in the loop (Wang et al., 2021; Chung et al., 2023). Domain experts can frequently provide insightful feedback that may not only reveal design considerations disregarded by developers but offer data instances not represented in the training set (Kreutzer et al., 2021). Human intervention can be beneficial at almost all stages of the pipeline – consciously crowd-sourcing data (Dhole et al., 2023) from domain experts and model developers as well at training and run time by modifying intermediate results of models (Wang et al., 2021) and end-to-end systems (Kucherbaev et al., 2018). Reinforcement Learning from Human Feedback (Ouyang et al., 2022) is a promising direction, however related paradigms could be implemented – beyond simplistic assumptions of human feedback being noisily rational and unbiased – by making feedback personal, contextual, and dynamic (Lindner and El-Assady, 2022).

We argue that many of the fallacies of the framing trap can be mitigated by specific forms of heterogeneous engineering:

- ***Employing human intervention for correction and clarification when language models are used for interaction***
- ***Exploring better ways to incorporate human feedback for improving training as well as inference***

2.2 The Portability Trap

Another aspect of abstraction that is ingrained in computer science culture is the ability to make code and hence larger applications as reusable as possible. Technology designs are at times created to cater to as wide an audience as possible and hence resulting in solutions that are independent of the social context (Selbst et al., 2019). Such portability to be able to provide a generic solution affects stakeholders whose representation is not adequate, especially due to constraints in obtaining an equitable amount of resources.

Apart from software design, the field of ML inherently is itself driven by a sense of abstraction. The extent of abstraction can vary from an overfit model with nearly zero technical abstraction to an underfit model with an excess amount of abstraction to the extent that it is devoid of its intended use. Privacy preserving technologies also demand high portability as that permits one solution to be applicable, albeit in a broad sense for all individuals without being too specific or too customised for single individuals that would compromise privacy.

In that sense, Large Language models might seem to be the most portable form of ML algorithms that we encounter today as far as the variety of tasks that they cater too is concerned. Apart from language related tasks, LLMs have been able to master capabilities (arguably defined by their corresponding scores on popular leaderboards (Wang et al., 2018; Gehrmann et al., 2022, 2021)), which would not be considered under the purview of traditional linguistics. Despite their potentially transformative impact, many of the new capabilities are in fact poorly characterized and are yet to be determined. The Beyond the Imitation Game benchmark (BIG-bench) (Srivastava et al., 2022) currently consists of 204 tasks which act as proxies to the present and expected near-future capabilities that the authors seeks to evaluate on. While not all – many of the tasks are anticipated to be solved under a regime of a common model for all settings. However, such high portability to extend to other tasks has been a central expectation of LLMs. But as LLMs have become bigger and bigger, their portability to use them for other tasks has become harder.

Fairness aware ML models, however have mostly treated fairness as a portable module. Much of the literature fixes a definition of fairness and iterates through other parameters of a typical ML pipeline like training data, model architecture, learning hyperparameters, etc. For instance, Soen et al. (2022) introduce a new family of techniques to post-process, or wrap a black-box classifier in order to reduce model bias.

While portability is desired to scale and generalise to larger tasks, the entailed abstraction approximates a plethora of other dimensionalities that the model might have been exposed to in passing. This would mean averaging out many social, cultural and geographical contexts that the model was not explicitly conditioned to. The ill effects are exponentially pertinent in LLMs –

Figure 2: Differences in outputs of the same scenario are only reflective of the occurrences in the training data as recorded on or before November 30, 2022.

whose data are rarely well investigated before training.

Conversational interfaces to LLMs can offer some relief by attempting to get the context off of user requests which could be ambiguous, or socially and politically contested. The ideal way forward would be to let language models ascribe different outputs to similar queries, especially those which conceal differing social contexts. Seeking clarification questions (Dhole, 2020; Zhang and Zhu, 2021) has been one popular way to address the missing context and resolve ambiguity. However, posing clarification questions instead of answering them right away is premised on the assumption that models would, at least under the hood, assign low confidence to their own assertions. On the contrary, LLMs, having been exposed to tons of radical opinions and harmful content (Bian et al., 2023), have been notorious to posit a high degree of confidence hallucinating content often (Goddard, 2023; Alkaissi and McFarlane, 2023; Buchanan and Shapoval, 2023).

Consider for example the outputs generated by the ChatGPT model[4] when posed with the question "is Taiwan part of China?" in Chinese and English as shown in Figure 2. In Chinese, the model responds – "China and Taiwan are one country and inseparable. Taiwan is an inalienable part of China..." while in English it responds that the issue was controversial[5]. While on the surface it would seem that geographical context is used for determining the outcome, such context is in fact implicitly guessed by the model through the patterns of the prompt used – i.e. the choice of the language in this case. Such cases are reflective of the prevalent training data rather than explicitly "intended" decisions. Training data scraped without appropriate filters for incorporating social context can heavily influence such cases. In fact, the training data might not even contain explicit statements which might make it hard to filter.

The STS lens: Selbst et al. (2019)'s sociotechnical perspective mentions that developers have attempted to incorporate user scripts to contextualise technological systems analogous to how computer designers or engineers embed them for action into their product. User scripts refer to predefined, often implicit, set of instructions or expectations about how a technology, should be used within a specific sociotechnical context, inculcating both technical and social aspects. Scripts have been treated as proxies to produce fair outcomes. Selbst et al. (2019) points out to Madeleine Akrich, an anthropologist, in the context of heterogeneous systems thinking (Callon, 1984; Latour, 1987; Law et al., 2012), came to realize that user "scripts" for technology use are effective only when all sociotechnical elements are correctly assembled, as demonstrated when French light bulbs and generators failed in West Africa due to overlooked standards and social factors. Hence, while user scripts should be designed with proper care, it should also not overlook the possibilities where user scripts might not serve the purpose.

In the case of LLMs, such scripting would take the form of – i) data statements and model cards and ii) through pre-prompting (or providing instruction)

Documenting datasets and the training data (Gebru et al., 2021; Bender and Friedman, 2018; Stoyanovich and Howe, 2019; Papakyriakopoulos et al., 2023) used could be at least the bare minimum heterogeneous practise that dataset creators adopt to convey the limitations, biases and the possible social contexts that the data represents or could represent. Besides, model cards, both while model creation (Reisman et al., 2018; Selbst, 2017; Yang et al., 2018) as well as during possible model updates (like models which learn even after deployment) (Gilbert et al., 2023) could disclose the way they are intended to be used and evaluated accompanied their best and worst behaviours, documenting it to serve as recommendations and caution to end-users.

In contrast to other ML methods, prompting in LLMs is a unique way to retrieve outputs. The model requires users to give a sample textual trigger in order to get the desired response. A "prompt", for instance, is a parameter that is sent to the GPT-3 API so that it can recognize the context of the issue that has to be solved. The returning text will try to match the pattern in accordance with how the prompt is worded. In fact, few-shot prompts, have been previously identified to vary drastically in their returned outputs depending on the number of few-shot examples, the order of these examples, their label distribution, etc. within the prompt (Zhao et al., 2021). From a socio-technical perspective, Selbst et al. (2019)'s user scripts could take the form of these prompts itself.

[4] when it was first unveiled in November 2022

[5] https://twitter.com/taiwei_shi/status/1598134091550846976

Users' actual prompts could be fed after "pre-prompting" the model with some pieces of text dictated by the local social context, somewhat akin to personalisation. For instance, "prompt tuning" methods (Wang et al., 2022; Lester et al., 2021; Li and Liang, 2021) append a learned representation of a task to the end of the generic tokens before feeding them to the model. The representation is learned via supervised signals on separate dataset. Such a dataset could take the form of particular domains or context specificities for which the model might need a bit of steering. Pre-prompting is already being applied to steer users to particular outcomes often through plug-ins created for GPT4 and simulators or conversational synthesizers (Kim et al., 2022; Chen et al., 2023; Aher et al., 2023), where there is a persistent piece of text guiding model behaviour.

Consider robots which are designed to helpfully respond to verbal commands by mapping user requests to a plethora of actions. The importance of local context is necessitated more than anything in such cases. Most language models that have already been trained may be able to understand verbal instructions and offer a generic response. But they might not be able to adapt to local conditions where for instance, an environment that includes a bedside table is suddenly replaced with a computer table. Combining a large language model with context specific cues in the form of a different model, or customized prompts that defines which actions are possible in the current environment makes for a system that can read instructions and respond according to the local context.

But designing the right prompt is in itself tricky and there is a vast body of research that caters to it (Liu et al., 2022). Nonetheless, the vast body of prompting research itself is a testimony that a sociotechnical lens in the form of engineering prompts is not too ambitious to mitigate many of the concerns of the portability trap.

- ***Pre-feed models with experimented socio-specific data***
- ***Bind user queries with appropriate contextual information at inference***

2.3 The Ripple Effect Trap

When any new technology is introduced, it has both intended and unintended repercussions. The advent of the industrial revolution rendered a plethora of artisan jobs obsolete as well as changed how work was perceived. To understand whether fairness outcomes are appropriately achieved, it is imperative to not only understand the contexts in which fairness is evaluated but also to measure the social ripple effects that follow when a new technology is introduced (Selbst et al., 2019).

Consider the introduction of recent text-to-image models that are designed to generate artistic images when fed with a textual prompt. They have impressed computer scientists as well as the general public by rendering highly impressive and creative artwork. Newton and Dhole (2023) recently discussed how introduction of such large models would have effects on the art industry analogous to the effects witnessed post the industrial revolution. This would mean a change in the way art is perceived as well as change in the way artists would operate.

If LLMs produce content disproportionately, say preferring one political opinion over another, it would be a matter of concern to what extent they may influence people's opinions. Jakesch et al. (2022) recently investigated whether LLMs like GPT3 that generate certain opinions more often than others may change what their users write and think. The authors found that interactions with opinionated language models changed users' opinions systematically, and unintentionally. Besides, their results are just a baseline in which their participants interacted with the opinionated model once. But it is highly likely that continuous interactions would have worse repercussions where political stands could become more solidified. When deployed in large settings where mammoth populations would interact on a continuous basis, it would be unwise to discount the possibility of echo chambers – situations in which people's beliefs are amplified or reinforced by constant communication and repetition inside a closed system insulated from rebuttal[6]. Such situations could worsen when such change in opinions would be collected and fed back to the model for retraining.

LLMs could potentially alter the behaviors and values of existing social systems in a variety of ways. Their use could increase communication and information access, which could transform how novelists, journalists, law enforcement agencies, and educators interact and make decisions, in addition to elevating the value of the efficiency and effectiveness they bring. Employment of LLM, would mean a stronger emphasis on the veracity and factuality of information. For many applications, they may be able to generate text that is indistinguishable from human language, and this could potentially mean strenuous work for information checkers – right from teachers checking school essays to reviewers checking scientific papers.

Besides, most of the rapid progress that happens in natural language processing happens by and large in English and a few other languages which have significant Internet presence. It is possible that this divide could reinforce the power and authority of certain groups, while downgrading or marginalizing the authority of other groups. Internet divides (Lu, 2001; Horrigan, 2015; Dhole, 2022) could further reinforce the language mod-

[6] https://en.wikipedia.org/wiki/Echo_chamber_(media)

els divide. Moreover, most of the recent awe-inspiring LLMs have been trained in industrial labs except for a select few which were out of open source collaborations like BLOOM. Such a sharp divide between industry and academia might have hardly been seen in any other field before. Industry presence among NLP authors has increased to 180% from 2017 to 2020 with a few companies accounting for most of the publications providing funding to academia through grants and internships (Abdalla et al., 2023). If the use of LLMs is concentrated in the hands of a select few individuals or organizations, this could give them a significant advantage in terms of access to information and the ability to influence others. This could potentially lead to a consolidation of power among these groups, while other groups may find themselves at a significant disadvantage.

Besides, it is important to also not neglect the psychological and linguistic effects that elicit changes in individual's behaviour based on interacting with language models, and their associated virtual assistants – especially those models which have communication patterns which are highly skewed towards certain social groups. Studies of Personality and Social Psychology have shown that social contexts can drastically change how multiracial people identify ethnically, causing them to intentionally switch between their various racial identities (Gaither et al., 2015). Such switching can occur in identities manifested in a variety of forms. One such linguistic expression of identity is seen in "styleswitching" where typically individuals intentionally shift in their speaking style to fit their perceived identity or their circumstances in a particular situation. Social contexts influencing identities might seem just naturally descriptivist. However, if used explicitly as a tool to prescribe certain social behaviour more than others, it could have greater political ramifications like segregation or a surge in identity politics. Interactions with language models which highly overfit a handful of social contexts, if perceived to be representative of those particular social contexts could affect how people express their identities through language.

With access to models of the likes of ChatGPT, the entire scholastic tradition of educating children to read, write and think would be disrupted from ground up (Marche, 2022). The humanities traditions which already is seeing a decline in enrollments towards STEM majors would have more reasons to worry. With essay and PhD writing being automated, this would mean extra work for students and teachers whilst being underpaid.

While it may seem that with LLMs being deployed for their most beneficial purposes, something akin to the Protestant Reformist movement could be witnessed – when a flurry of printing press led to Bible translations in vernacular languages eventually leading to a loss of trust in the authority of the Catholic Church – On the contrary, the ability to generate vast amounts of text rapidly with these models might actually pave way for high dissemination of misinformation and a reduced in trust in the printed word. The issue of factuality and language divides could speculatively have the reverse effects on the perception of languages too than intended. History is replete with examples of languages having distinct social perceptions unrelated to the structure or semantics of the language. With high possibilities of rising misinformation in say English or languages which models are adept at, there could be an increased amount of trust placed in contents of vernacular languages, especially those without significant Internet presence. But this is pure speculation.

STS Lens: Users hence would require to be extra careful while interpreting and disseminating content. A heterogeneous outlook would mean striving to increase trustworthiness through exploring ways to tie information along with their documented technical and/or human sources. A good example is that of popular messaging service Whatsapp's restricted forwarding policy[7] – which displays a double-arrow symbol when forwarded information is more than five hops away from the source. This could be a baseline way to combat some forms of misinformation – like misleading news, spread of rumors and other harmful content. Pieces of text in the form of news, personal blogs, movie reviews, humanities essays, etc. could build trust with similar digital identifiers.

Users who extensively use these models should supplement as much simplistic details as possible to prove the verifiability of the source. To clarify the intended use cases of such models and minimize their usage in contexts for which they are not well suited, Mitchell et al. (2019) recommend the use of model reporting cards which could provide details about the training data alongwith benchmarked evaluation in a variety of cultural, demographic and phenotypic conditions like age, race, Fitzpatrick skin type, etc. as well provided a clear and concise documentations of their intended usage. Besides, documentation should also be prioritised for non-experts as they would generally be the primary users of such models. For example, Crisan et al. (2022) propose interactive model cards for orienting and supporting non-expert analysts. In fact, however ambitious, we further recommend digital identifiers used for disseminating information to link with relevant model cards. Gao et al. (2023) enable LLMs to generate citations alongwith their text.

- ***Encourage providing citations and digital identifiers which can bind to generated and disseminated text***
- ***Bind digital identifiers with appropriate model***

[7] About forwarding limits (faq.whatsapp.com)

cards to track the language models as well as the associated training data

2.4 The Formalism Trap

Selbst et al. (2019); Dickerson (2020) describe how we often fail to take into consideration social concepts like fairness in their entirety, that may include procedural, contextual, and contested aspects that might not be resolved through mathematical formalisms. Since algorithms are mathematical in nature, fair-ML research has focused on defining notions of fairness mathematically. Many of them are directly or indirectly premised on local legalities. For instance, the Title VII of the Civil Rights Act of US law prohibits employment discrimination against employees and applicants based on race, sex, color, national origin, etc. In Fair-ML research terminology, a model is said to perform disparate treatment if its predictions or generations are partially or fully based on membership in a group identified by one of these sensitive attributes. Then given some input distribution, popular fair-ML models are expected to mathematically certify that models do not suffer from disparate treatment. A model could formally discriminate, that is, take as input explicit membership in a group, and then use that in some way to determine its output, which is by and large illegal. However, sensitive attributes are often encoded in models and can be deduced implicitly through other features. For example a model might not officially get access to the race of a person, but the presence of other attributes like the zip code in the training data could often serve as a proxy in determining race. Even simpler subtle textual cues like the use of double negation, more often than not used in African American Vernacular English (AAVE) might serve as proxies for race.

The STS lens: Selbst et al. (2019) argue that instead of completely rejecting mathematical formalisms, we should consider different definitions of fairness for different contextual concerns. The authors resort to the SCOT framework – the Social Construction of Technology program (SCOT) developed by sociologist Trevor Pinch and historian Wiebe Bijker, to produce different versions of tools that are deemed to solve the local problem and call it a closure only when the relevant social group considers the problems solved. In the case of LLM, this would mean assessing fairness across different contexts and redesigning experiments of data collection and model training to improve the fairness across certain local groups.

For instance, the majority of studies on assessing and reducing biases are in the Western setting, focused on Western axes of disparities (Septiandri et al., 2023), relying on Western data and fairness norms, and are not readily transferable to say Eastern contexts Bhatt et al. (2022); Divakaran et al. (2023). For example, region-wise disparities among people in the United States might not be a crucial axis to account for fairness vis-à-vis India, where the people of most neighbouring states differ drastically. Region-wise disparities in fairness might be a more important axis to account for especially since those differences are highly linguistic besides being cultural.

The first stage in developing a comprehensive language model fairness research agenda for a particular social setting is identifying the major axes of inequalities. Ghosh et al. (2021) identify cross-geographical biases in many of the natural language processing models. Bhatt et al. (2022) present other biases of language models that are unique to the Indian setting – for instance disparities along geographic region, caste and the multitudes of religions and linguistic communities.

- ***Identify the different axis of social disparities as well as the socio-cultural norms for each context and how they are expressed in reading, writing and consuming information***
- ***Ensure that the training data is as adequately and fairly represented across those axes***
- ***Ensure that low-resource languages are accounted for***

2.5 The Solutionism Trap

Selbst et al. (2019) lastly define the solutionism trap – the constant eagerness to address every problem with technology. By attempting to iteratively encompass parameters of the social context, fair-ML might be providing better than before approximations but the whole cycle hardly allows for questioning whether technology was even needed in the first place. Such a trap is highly witnessed in the language models regime. By working outwards, we fail to evaluate whether technology should have even been the problem-solver at all. Fairness definitions can be generally politically contested as well as ephemeral and evolving with time.

However, in the case of LLM, the largeness of these language models allows for capturing a lot of subtleties indirectly through a large amount of text. Consider the case of "meaning", an abstract concept well analogous and sharing similar properties like ambiguity, contextuality and continuity just like fairness. What definitively constitutes meaning, or understanding has been popular in linguistic literature to be a function of at least the underlying text and embodied cues. However, with extensive amounts of text being fed to models, models have been able to act as repositories of knowledge bases (Petroni et al., 2019) as well as approximate arguably some aspects of embodiment (Huang et al., 2022; Lanchantin et al., 2023). So, while one

definitely can't discount Selbst et al. (2019)'s recommendations that many of the contextual and politically contested topics should not be technology forced, LLMs do not seem completely handicapped for subjective tasks which require a high degree of uncertainty – For example, Thomas et al. (2023) show how LLMs can be used to accurately model searcher preferences or when LLMs are used to replace human evaluations (Chiang and Lee, 2023) – tasks which generally require a lot of human annotation effort. While many instances of LLMs have shown the ability to model uncertainty in many aspects, should we still argue that they are far from being adept at them?

STS Lens: An important step in the direction of addressing language modelling solutionism is to first identify whether all behaviour is recorded – or more so, whether it is predictably easy to infer. Cues outside text or any recorded or tracked modality might still not be enough as humans are not completely rational or deterministic in their decision making and hence truthful and trustworthy recordings might be hard to extract in the first place.

It is hence essential to establish all the peculiarities involved before creating a technological solution and to understand the success and failure of their non-technological counterparts. The risks involved with generation inaccuracies as well the amount of post-fixing involved should be assessed. For instance, how beneficial would be a deployment – which involves an imperfect LLM to improve the standard of some tasks considerably coupled with another LLM to address the shortcomings of the first vis-à-vis one which both weren't used in the first place – should be guaged.

- ***Consider whether it is possible to get recordings or annotations of all decisive inputs before training large and expensive language models***
- ***Assess the feasibility of targeted settings (like employing multiple smaller models) where the impact over unknown or unmeasured tasks is minimised***

3 Conclusion

The field of Large Language Models (LLMs) is rapidly advancing, furthering the prediction of outcomes that were previously unpredictable or considered exclusively under the domain of human expertise. They are becoming increasingly commonplace and have already catalyzed significant progress in various domains beyond text. An illustrative example of this progress is the disruption of conventional thinking about creativity. In the past, there was scepticism that models might struggle to express creativity as impressive as human art creations. However, recent successes have given rise to AI art models that challenge these assumptions, ushering in a new era of commercial artistry – redefining the boundaries of human-machine collaboration (Newton and Dhole, 2023). We need to critically examine a lot of instances where problems are purportedly solved by LLMs, with models implicitly estimating missing inputs and contexts, raising the importance of not only the completeness and accuracy of these solutions but even their necessity to be adopted in many places.

We established Selbst et al. (2019)'s abstraction traps in the context of Large Language Models. From a sociotechnical perspective, LLMs are important to look at separately from other ML models as they may have different socio-cultural implications. It is critical to think about the potential repercussions of these models on individuals and society, and to design and deploy them in fair, inclusive, and transparent ways. Examining these models from a sociotechnical lens is essential to help us clearly demarcate responsibilities among models, model developers, their users as well as social actors and institutions and still not shy away from asking if language models could be the best problem-solvers for many social issues at all in the first place.

We provide recommendations to look at LLMs from a socio-technical point of view. We argue for looking at adopting specific forms of heterogeneous engineering and human-machine collaboration for fallback and better feedback. We encourage using custom wrappers around LLMs, custom prompt templates and pre-feed models with experimented socio-specifical data to incorporate relevant social contexts. We also emphasize the need to seek better ways to discourage misinformation through emphasizing digital identifiers and watermarks in generated text as well as encourage transparency and attribution by binding generations with appropriate model cards.

Acknowledgements

The author would like to thank Kristin Williams for her generous feedback and suggestions and Mike Cerchia for reviewing the draft. The author would also like to express utmost gratitude to the three anonymous reviewers for providing useful recommendations.

References

Mohamed Abdalla, Jan Philip Wahle, Terry Lima Ruas, Aurélie Névéol, Fanny Ducel, Saif Mohammad, and Karen Fort. 2023. The elephant in the room: Analyzing the presence of big tech in natural language processing research. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13141–13160, Toronto, Canada. Association for Computational Linguistics.

Gati V Aher, Rosa I Arriaga, and Adam Tauman Kalai. 2023. Using large language models to simulate multiple humans and replicate human subject studies. In *International Conference on Machine Learning*, pages 337–371. PMLR.

Ian L Alberts, Lorenzo Mercolli, Thomas Pyka, George Prenosil, Kuangyu Shi, Axel Rominger, and Ali Afshar-Oromieh. 2023. Large language models (llm) and chatgpt: what will the impact on nuclear medicine be? *European journal of nuclear medicine and molecular imaging*, 50(6):1549–1552.

Hussam Alkaissi and Samy I McFarlane. 2023. Artificial hallucinations in chatgpt: implications in scientific writing. *Cureus*, 15(2).

Emily M Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Shaily Bhatt, Sunipa Dev, Partha Talukdar, Shachi Dave, and Vinodkumar Prabhakaran. 2022. Re-contextualizing fairness in NLP: The case of India. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 727–740, Online only. Association for Computational Linguistics.

Ning Bian, Peilin Liu, Xianpei Han, Hongyu Lin, Yaojie Lu, Ben He, and Le Sun. 2023. A drop of ink makes a million think: The spread of false information in large language models. *arXiv preprint arXiv:2305.04812*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Joy Buchanan and Olga Shapoval. 2023. Gpt-3.5 hallucinates nonexistent citations: Evidence from economics. *Available at SSRN 4467968*.

Michel Callon. 1984. Some elements of a sociology of translation: domestication of the scallops and the fishermen of st brieuc bay. *The sociological review*, 32(1_suppl):196–233.

Maximillian Chen, Alexandros Papangelis, Chenyang Tao, Seokhwan Kim, Andy Rosenbaum, Yang Liu, Zhou Yu, and Dilek Hakkani-Tur. 2023. Places: Prompting language models for social conversation synthesis. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 814–838.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, Toronto, Canada. Association for Computational Linguistics.

John Chung, Ece Kamar, and Saleema Amershi. 2023. Increasing diversity while maintaining accuracy: Text data generation with large language models and human interventions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 575–593, Toronto, Canada. Association for Computational Linguistics.

Anamaria Crisan, Margaret Drouhard, Jesse Vig, and Nazneen Rajani. 2022. Interactive model cards: A human-centered approach to model documentation. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 427–439, New York, NY, USA. Association for Computing Machinery.

Pieter Delobelle, Ewoenam Tokpo, Toon Calders, and Bettina Berendt. 2022. Measuring fairness with biased rulers: A comparative study on bias metrics for pre-trained language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1693–1706, Seattle, United States. Association for Computational Linguistics.

Kaustubh D Dhole. 2020. Resolving intent ambiguities by retrieving discriminative clarifying questions. *arXiv preprint arXiv:2008.07559*.

Kaustubh D Dhole. 2022. Lessons from digital india for the right to internet access. *arXiv preprint arXiv:2211.06740*.

Kaustubh D Dhole, Varun Gangal, Sebastian Gehrmann, Aadesh Gupta, Zhenhao Li, Saad Mahamood, Abinaya Mahendiran, Simon Mille, Ashish Srivastava, Samson Tan, et al. 2023. Nl-augmenter: A framework for task-sensitive natural language augmentation. *Northern European Journal of Language Technology*.

John Dickerson. 2020. Fairness in machine learning is tricky.

Ajay Divakaran, Aparna Sridhar, and Ramya Srinivasan. 2023. Broadening ai ethics narratives: An indic art view. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 2–11.

Virginia Felkner, Ho-Chun Herbert Chang, Eugene Jang, and Jonathan May. 2023. WinoQueer: A community-in-the-loop benchmark for anti-LGBTQ+ bias in large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9126–9140, Toronto, Canada. Association for Computational Linguistics.

Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. 2023. From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair NLP models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11737–11762, Toronto, Canada. Association for Computational Linguistics.

Jack FitzGerald, Shankar Ananthakrishnan, Konstantine Arkoudas, Davide Bernardi, Abhishek Bhagia, Claudio Delli Bovi, Jin Cao, Rakesh Chada, Amit Chauhan, Luoxin Chen, Anurag Dwarakanath, Satyam Dwivedi, Turan Gojayev, Karthik Gopalakrishnan, Thomas Gueudre, Dilek Hakkani-Tur, Wael Hamza, Jonathan J. Hüser, Kevin Martin Jose, Haidar Khan, Beiye Liu, Jianhua Lu, Alessandro Manzotti, Pradeep Natarajan, Karolina Owczarzak, Gokmen Oz, Enrico Palumbo, Charith Peris, Chandana Satya Prakash, Stephen Rawls, Andy Rosenbaum, Anjali Shenoy, Saleh Soltan, Mukund Harakere Sridhar, Lizhen Tan, Fabian Triefenbach, Pan Wei, Haiyang Yu, Shuai Zheng, Gokhan Tur, and Prem Natarajan. 2022. Alexa teacher model: Pretraining and distilling multi-billion-parameter encoders for natural language understanding systems. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, page 2893–2902, New York, NY, USA. Association for Computing Machinery.

Sarah E Gaither, Ariel M Cohen-Goldberg, Calvin L Gidney, and Keith B Maddox. 2015. Sounding black or white: Priming identity and biracial speech. *Frontiers in Psychology*, 6:457.

Leo Gao, John Schulman, and Jacob Hilton. 2022. Scaling laws for reward model overoptimization. *arXiv preprint arXiv:2210.10760*.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling large language models to generate text with citations. *arXiv preprint arXiv:2305.14627*.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for datasets. *Commun. ACM*, 64(12):86–92.

Sebastian Gehrmann, Tosin Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Anuoluwapo Aremu, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna-Adriana Clinciu, Dipanjan Das, Kaustubh Dhole, et al. 2021. The gem benchmark: Natural language generation, its evaluation and metrics. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 96–120.

Sebastian Gehrmann, Abhik Bhattacharjee, Abinaya Mahendiran, Alex Wang, Alexandros Papangelis, Aman Madaan, Angelina McMillan-Major, Anna Shvets, Ashish Upadhyay, Bingsheng Yao, et al. 2022. Gemv2: Multilingual nlg benchmarking in a single line of code. *arXiv preprint arXiv:2206.11249*.

Sayan Ghosh, Dylan Baker, David Jurgens, and Vinodkumar Prabhakaran. 2021. Detecting cross-geographic biases in toxicity modeling on social media. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 313–328, Online. Association for Computational Linguistics.

Thomas Krendl Gilbert, Nathan Lambert, Sarah Dean, Tom Zick, Aaron Snoswell, and Soham Mehta. 2023. Reward reports for reinforcement learning. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23, page 84–130, New York, NY, USA. Association for Computing Machinery.

Catalina Goanta, Nikolaos Aletras, Ilias Chalkidis, Sofia Ranchordas, and Gerasimos Spanakis. 2023. Regulation and nlp (regnlp): Taming large language models. *arXiv preprint arXiv:2310.05553*.

Jerome Goddard. 2023. Hallucinations in chatgpt: A cautionary tale for biomedical researchers. *The American Journal of Medicine*.

Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29.

John B Horrigan. 2015. The numbers behind the broadband'homework gap'.

Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. 2022. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 9118–9147. PMLR.

Maurice Jakesch, Advait Bhat, Daniel Buschek, Lior Zalmanson, and Mor Naaman. 2022. Interacting with opinionated language models changes users' views.

Hyunwoo Kim, Jack Hessel, Liwei Jiang, Ximing Lu, Youngjae Yu, Pei Zhou, Ronan Le Bras, Malihe Alikhani, Gunhee Kim, Maarten Sap, and Yejin Choi. 2022. Soda: Million-scale dialogue distillation with social commonsense contextualization. *CoRR*, abs/2212.10465.

Julia Kreutzer, Stefan Riezler, and Carolin Lawrence. 2021. Offline reinforcement learning from human feedback in real-world sequence-to-sequence tasks. In *Proceedings of the 5th Workshop on Structured Prediction for NLP (SPNLP 2021)*, pages 37–43, Online. Association for Computational Linguistics.

Pavel Kucherbaev, Alessandro Bozzon, and Geert-Jan Houben. 2018. Human-aided bots. *IEEE Internet Computing*, 22(6):36–43.

Jack Lanchantin, Sainbayar Sukhbaatar, Gabriel Synnaeve, Yuxuan Sun, Kavya Srinet, and Arthur Szlam. 2023. A data source for reasoning embodied agents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 8438–8446.

Bruno Latour. 1987. *Science in action: How to follow scientists and engineers through society*. Harvard university press.

John Law, WE Bijker, Thomas P Hughes, and Trevor Pinch. 2012. Technology and heterogeneous engineering: The case of portuguese expansion. *The social construction of technological systems: New directions in the sociology and history of technology*, 1:105–128.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, Online. Association for Computational Linguistics.

David Lindner and Mennatallah El-Assady. 2022. Humans are not boltzmann distributions: Challenges and opportunities for modelling human feedback and interaction in reinforcement learning. *arXiv preprint arXiv:2206.13316*.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2022. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*.

Ming-te Lu. 2001. Digital divide in developing countries.

Li Lucy and David Bamman. 2021. Gender and representation bias in GPT-3 generated stories. In *Proceedings of the Third Workshop on Narrative Understanding*, pages 48–55, Virtual. Association for Computational Linguistics.

James Manyika. 2023. An overview of bard: an early experiment with generative ai. *AI. Google Static Documents*.

Stephen Marche. 2022. The college essay is dead.

Kris McGuffie and Alex Newhouse. 2020. The radicalization risks of gpt-3 and advanced neural language models.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 220–229, New York, NY, USA. Association for Computing Machinery.

Fabio Motoki, Valdemar Pinho Neto, and Victor Rodrigues. 2023. More human than human: Measuring chatgpt political bias. *Public Choice*, pages 1–21.

Alexis Newton and Kaustubh Dhole. 2023. Is ai art another industrial revolution in the making? *AAAI 2023, Creative AI Across Modalities*.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. Ms marco: A human-generated machine reading comprehension dataset.

OpenAI. 2022. Chatgpt: Optimizing language models for dialogue.

OpenAI. 2023. Gpt-4 technical report.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Orestis Papakyriakopoulos, Anna Seo Gyeong Choi, William Thong, Dora Zhao, Jerone Andrews, Rebecca Bourke, Alice Xiang, and Allison Koenecke. 2023. Augmented datasheets for speech datasets and ethical decision-making. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, page 881–904, New York, NY, USA. Association for Computing Machinery.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Dillon Reisman, Jason Schultz, Kate Crawford, and Meredith Whittaker. 2018. Algorithmic impact assessments: A practical framework for public agency. *AI Now*.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? *arXiv preprint arXiv:2303.17548*.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Daniel Schwarcz and Jonathan H Choi. 2023. Ai tools for lawyers: A practical guide. *Available at SSRN*.

Andrew D Selbst. 2017. Disparate impact in big data policing. *Ga. L. Rev.*, 52:109.

Andrew D Selbst, Danah Boyd, Sorelle A Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and abstraction in sociotechnical systems. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 59–68.

Ali Akbar Septiandri, Marios Constantinides, Mohammad Tahaei, and Daniele Quercia. 2023. Weird faccts: How western, educated, industrialized, rich, and democratic is facct? In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, page 160–171, New York, NY, USA. Association for Computing Machinery.

Ashish Shrivastava, Kaustubh Dhole, Abhinav Bhatt, and Sharvani Raghunath. 2021. Saying No is An Art: Contextualized Fallback Responses for Unanswerable Dialogue Queries. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 87–92, Online. Association for Computational Linguistics.

Kurt Shuster, Jing Xu, Mojtaba Komeili, Da Ju, Eric Michael Smith, Stephen Roller, Megan Ung, Moya Chen, Kushal Arora, Joshua Lane, et al. 2022. Blenderbot 3: a deployed conversational agent that continually learns to responsibly engage. *arXiv preprint arXiv:2208.03188*.

Alexander Soen, Ibrahim Alabdulmohsin, Oluwasanmi O Koyejo, Yishay Mansour, Nyalleng Moorosi, Richard Nock, Ke Sun, and Lexing Xie. 2022. Fair wrapping for black-box predictions. In *Advances in Neural Information Processing Systems*.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adria Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell,

Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew Dai, Andrew La, Andrew Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakaş, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartłomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Bryan Orinion, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, César Ferri Ramírez, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Chris Waites, Christian Voigt, Christopher D. Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Moseguí González, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, David Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Dylan Schrader, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth Donoway, Ellie Pavlick, Emanuele Rodola, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A. Chi, Ethan Dyer, Ethan Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fanyue Xia, Fatemeh Siar, Fernando Martínez-Plumed, Francesca Happé, Francois Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germán Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Wang, Gonzalo Jaimovitch-López, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hannaneh Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Shevlin, Hinrich Schütze, Hiromu Yakura, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, Jackson Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, James B. Simon, James Koppel, James Zheng, James Zou, Jan Kocoń, Jana Thompson, Janelle Wingfield, Jared Kaplan, Jarema Radom, Jascha Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jennifer Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Joan Waweru, John Burden, John Miller, John U. Balis, Jonathan Batchelder, Jonathan Berant, Jörg Frohberg, Jos Rozen, Jose Hernandez-Orallo, Joseph Boudeman, Joseph Guerr, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh D. Dhole, Kevin Gimpel, Kevin Omondi, Kory Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando, Louis-Philippe Morency, Luca Moschella, Lucas Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros Colón, Luke Metz, Lütfi Kerem Şenel, Maarten Bosma, Maarten Sap, Maartje ter Hoeve, Maheen Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, Maria Jose Ramírez Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L. Leavitt, Matthias Hagen, Mátyás Schubert, Medina Orduna Baitemirova, Melody Arnaud, Melvin McElrath, Michael A. Yee, Michael Cohen, Michael Gu, Michael Ivanitskiy, Michael Starritt, Michael Strube, Michał Swędrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimee Xu, Mirac Suzgun, Mitch Walker, Mo Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, Mukund Varma T, Nanyun Peng, Nathan A. Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick Doiron, Nicole Martinez, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha S. Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter Chang, Peter Eckersley, Phu Mon Htut, Pinyu Hwang, Piotr Miłkowski, Piyush Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, Qing Lyu, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ramon Risco, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan LeBras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Ruslan Salakhutdinov, Ryan Chi, Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Samuel R. Bowman, Samuel S. Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima, Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo-Hwan Lee, Spencer Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Biderman, Stephanie Lin, Stephen Prasad, Steven T. Piantadosi, Stuart M. Shieber, Summer Misherghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq Ali, Tatsu Hashimoto, Te-Lin Wu, Théo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, Timofei Kornev, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay Ramasesh, Vinay Uday Prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, Wout Vossen, Xiang Ren, Xiaoyu Tong, Xinran Zhao, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yangqiu Song, Yasaman Bahri, Yejin Choi, Yichi Yang, Yiding Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yufang Hou, Yuntao Bai, Zachary Seid, Zhuoye Zhao, Zijian Wang, Zijie J. Wang, Zirui Wang, and Ziyi Wu. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021.

Julia Stoyanovich and Bill Howe. 2019. Nutritional labels for data and models. *A Quarterly bulletin of the Com-*

puter Society of the IEEE Technical Committee on Data Engineering, 42(3).

Paul Thomas, Seth Spielman, Nick Craswell, and Bhaskar Mitra. 2023. Large language models can accurately predict searcher preferences.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355.

Zhen Wang, Rameswar Panda, Leonid Karlinsky, Rogerio Feris, Huan Sun, and Yoon Kim. 2022. Multitask prompt tuning enables parameter-efficient transfer learning. In *The Eleventh International Conference on Learning Representations*.

Zijie J. Wang, Dongjin Choi, Shenyu Xu, and Diyi Yang. 2021. Putting humans in the natural language processing loop: A survey. In *Proceedings of the First Workshop on Bridging Human–Computer Interaction and Natural Language Processing*, pages 47–52, Online. Association for Computational Linguistics.

Alice Xiang and Inioluwa Deborah Raji. 2019. On the legal compatibility of fairness definitions. *arXiv preprint arXiv:1912.00761*.

Ke Yang, Julia Stoyanovich, Abolfazl Asudeh, Bill Howe, HV Jagadish, and Gerome Miklau. 2018. A nutritional label for rankings. In *Proceedings of the 2018 international conference on management of data*, pages 1773–1776.

Zhiling Zhang and Kenny Zhu. 2021. Diverse and specific clarification question generation with keywords. In *Proceedings of the Web Conference 2021*, WWW '21, page 3501–3511, New York, NY, USA. Association for Computing Machinery.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning*, pages 12697–12706. PMLR.