frontiers | Frontiers in Political Science

TYPE Review
PUBLISHED 16 October 2023
DOI 10.3389/fpos.2023.1257092

OPEN ACCESS

EDITED BY
Klaus Gerhard Troitzsch,
University of Koblenz, Germany

REVIEWED BY
Vladimir Sazonov,
University of Tartu, Estonia
Samar Haider,
University of Pennsylvania, United States

*CORRESPONDENCE
Mitchell Linegar
✉ mlinegar@caltech.edu

RECEIVED 11 July 2023
ACCEPTED 25 September 2023
PUBLISHED 16 October 2023

CITATION
Linegar M, Kocielnik R and Alvarez RM (2023) Large language models and political science. *Front. Polit. Sci.* 5:1257092. doi: 10.3389/fpos.2023.1257092

Large language models and political science

Mitchell Linegar[1]*, Rafal Kocielnik[2] and R. Michael Alvarez[1,3]

[1]Division of Humanities and Social Science, California Institute of Technology, Pasadena, CA, United States, [2]Computing and Mathematical Sciences, California Institute of Technology, Pasadena, CA, United States, [3]Center for Science, Society, and Public Policy, California Institute of Technology, Pasadena, CA, United States

Large Language Models (LLMs) are a type of artificial intelligence that uses information from very large datasets to model the use of language and generate content. While LLMs like GPT-3 have been used widely in many applications, the recent public release of OpenAI's ChatGPT has opened more debate about the potential uses and abuses of LLMs. In this paper, we provide a brief introduction to LLMs and discuss their potential application in political science and political methodology. We use two examples of LLMs from our recent research to illustrate how LLMs open new areas of research. We conclude with a discussion of how researchers can use LLMs in their work, and issues that researchers need to be aware of regarding using LLMs in political science and political methodology.

KEYWORDS

Large Language Models (LLM), ChatGPT, natural language processing, political science, political methodology

1. Introduction

Just 2 or 3 years ago, few political scientists would have heard of Large Language Models. But in just the past year, LLMs have jumped squarely into public consciousness, sparking many discussions and debates about their uses and abuses in many different situations. We anticipate that many political scientists are now considering how to study the use of LLMs in politics and government, or are thinking about using LLMs in their research.

Our goal in this paper is to introduce political scientists to Large Language Models. We wish to inform researchers how they can use LLMs in their work, presenting examples drawn from new research that use LLMs. We also want to make political scientists better aware of the issues associated with LLMs, and to provide some best practices for how the research community should be using these innovative new natural language processing methods.

The paper is structured as follows. Next we provide a brief introduction to LLMs, followed by a section that discusses the current state of the art and available LLM resources. We then present a number of use cases for researchers in Section 4, followed by discussion of important current issues regarding using LLMs in research (Sections 5, 6). Importantly, we provide a discussion of best practices for research use of LLMs, before we conclude.

2. What are LLMs?

Large language models (LLMs) have recently entered the public conversation about artificial intelligence, as they represent a new and easy-to-use methodology for studying language. LLMs are a new approach for natural language processing (NLP), and proponents have argued that LLMs may revolutionalize the analysis of text and language data. Under the hood, LLMs take advantage of deep learning techniques, large scale computational resources, and huge quantities of training data to generate coherent and contextually relevant text.

This means that LLMs are generally useful in many different applications, ranging from the analysis of text and language (NLP), to the generation of new text content, and thus to the further development of conversational bots. Our paper seeks to outline for political and social scientists how these new methods can be applied in research.

Text and language are information that have long been important for the study of political science. For example, textual data has been used in political science to study political party manifestos (Laver and Garry, 2000), political speeches (Grimmer and Stewart, 2013), legislator communications with constituents (Grimmer, 2013), and social movements (Kann et al., 2023). LLMs and other generative AI models can also be used for creating political content at scale (Zhang et al., 2023a) (see Figure 1). LLMs have great promise for studying the text and language of politics, producing what will be a better understanding of political rhetoric and communications than was possible with previous NLP methods (Grimmer and Stewart, 2013).

Architecturally, LLMs usually are based on neural networks, specifically transformer models. Transformer models work well with text data as they can detect the complexities of language using encoders and decoders. In a transformer model, encoders reduce the dimensionality of the text into embeddings. The decoder then produces some type of output that is based on the embeddings. Transformer models then use "self-attention" to better learn the long-term dependencies in sequences of text, which helps them structure the output in more meaningful and realistic ways.

Of course, like most deep learning methods, LLMs require training data, lots and lots of high quality and hopefully unbiased text. In order to learn the nuances of language, LLMs need training data from text sources like blogs, social media, books, articles; in other words, as much readily-available text data that can be scraped from public sources. LLMs can be pretrained using these data, learning the ability to predict the next set of words in a sequence with missing words. They can then be fine-tuned for specific applications, using domain-specific training data; one example would be developing an LLM to summarize research articles from a discipline like political science.

As we will argue in this paper, LLMs have great potential for use in political science. On the other hand, they also raise significant issues for researchers. One important issue is that like many deep learning models, LLMs are "black boxes"—their development and estimation is not transparent nor is it easily interpretable by or explainable to humans. LLMs have the potential to be substantially biased, as they rest on the quality and coverage of the data they are trained on. If the training data contains biases, or does not contain text from a wide sampling of sources, the outputs produced by LLMs will be biased. A final issue is that LLMs require vast computational resources, raising ethical concerns about their environmental impacts.

2.1. Understanding language model architectures

Language models (LMs) are computational frameworks designed to predict the likelihood of a sequence of words. At their core, they are based on the premise of understanding and predicting the probability distribution over sequences of words. Given a sequence of words $w_1, w_2, \ldots, w_t$, the LM aims to predict:

$$P(w_{t+1}|w_1, w_2, \ldots, w_t) \tag{1}$$

where $P(w_{t+1}|w_1, w_2, \ldots, w_t)$ denotes the conditional probability of the word w_{t+1} occurring next after the words $w_1, w_2, \ldots, w_t$.

Traditional language models, such as n-gram models, relied on counting the occurrences of word sequences in large text corpora to estimate these probabilities. For instance, a bigram model, which considers only the last word to predict the next one, would compute:

$$P(w_{t+1}|w_t) = \frac{\text{count}(w_t, w_{t+1})}{\text{count}(w_t)} \tag{2}$$

where $\text{count}(w_t, w_{t+1})$ represents the number of times the word pair (w_t, w_{t+1}) appears in the corpus and $\text{count}(w_t)$ is the number of times the word w_t appears.

However, with the advent of deep learning, LMs underwent significant transformation. Neural network-based models, particularly recurrent neural networks (RNNs), long short-term memory networks (LSTMs), and transformers, began to dominate the scene. These models compute the probability of the next word using a complex function f parameterized by weights θ:

$$P(w_{t+1}|w_1, w_2, \ldots, w_t) = f(w_1, w_2, \ldots, w_t; \theta) \tag{3}$$

where these weights θ are learned by adjusting them to minimize the difference between the model's predictions and the actual next words in a large training corpus.

The recent emergence of LLMs, like OpenAI's GPT series, leverages the transformer architecture. Benefiting from massive amounts of data and an extensive number of parameters, these models can memorize rare patterns, generalize across tasks, and generate coherent texts over long passages.

2.2. Masking and its role in language model training

In the context of training transformer-based models like BERT (Bidirectional Encoder Representations from Transformers), masking plays a pivotal role. Words or tokens in the input sequence are randomly "masked" or hidden, and the model is trained to predict these masked words based on their context. This is termed as the "masked language model" objective. The concept of masking allows the model to learn bidirectional representations, as opposed to traditional LMs which are unidirectional (either left-to-right or right-to-left).

For a given sequence $w_1, w_2, \ldots, w_t$ where w_j is masked, the model aims to predict:

$$P(w_j|w_1, w_2, \ldots, w_{j-1}, w_{j+1}, \ldots, w_t) \tag{4}$$

This process enhances the model's ability to understand context from both sides of a word, leading to richer and more robust representations.

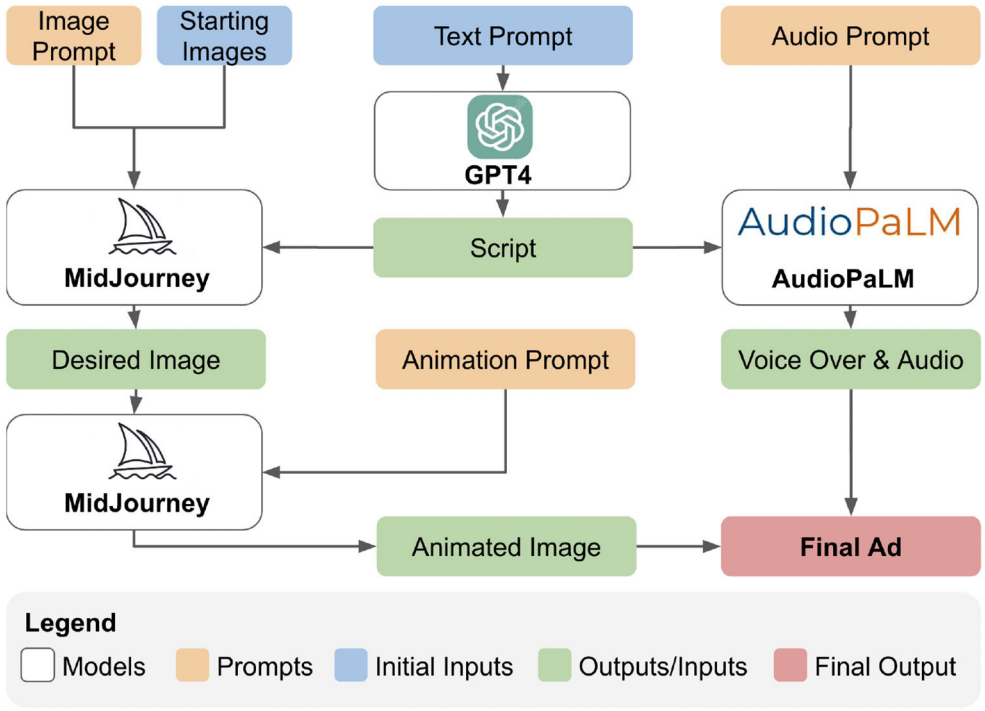

FIGURE 1
An example workflow illustrating how multiple *Generative AI* models can be combined under a chained workflow to produce a political advertisement.

2.3. Perplexity as a metric for LLMs

Perplexity is a widely used metric for evaluating the performance of language models. Intuitively, it measures how well the probability distribution predicted by the model aligns with the true distribution of the words in the text. Mathematically, for a test set T consisting of N words, the perplexity $\mathcal{P}$ is defined as:

$$\mathcal{P}(T) = \exp\left(-\frac{1}{N}\sum_{i=1}^{N}\log P(w_i|w_1,\ldots,w_{i-1})\right) \quad (5)$$

A lower perplexity indicates that the model's predictions are closer to the actual distribution of the words in the text. During training, minimizing perplexity is equivalent to maximizing the likelihood of the data, making it a valuable metric for both training and evaluation.

3. Private and open-source models

Researchers who want to employ LLMs in their work will typically access them in one of two ways: either via an API provided by a commercial third-party or by hosting and running open-source versions of the model themselves. Commercial APIs are by far the easiest way for researchers to use LLMs for their own research. Startup costs are low: users are charged for each token they use ("pay-per-use"), which can be done using simple, high-level APIs. These models, often trained with proprietary data and computational resources far beyond what is available to the average researcher, also tend to be more powerful, generating more human-like responses more quickly than can be done locally. OpenAI's GPT-4, for example, is consistently found to outperform all other publicly available LLMs across a variety of text-based tasks, but its training data and architecture are not public knowledge. This is typical of private APIs for LLMs, emphasizing their black-box status. Privacy is another concern with commercial APIs, since querying them typically requires sending data to the service itself. As a result, when using highly sensitive or personally identifiable data, using commercial APIs for LLMs (at least in their current state) may not be practical for some research purposes.

Using open-source models can address many of the concerns raised above: the data, code, and hardware used to train the models tend to be publicly available. Since the models are run on private machines and not given to a third party, data privacy concerns can be minimized. A final advantage of using open-source models is that it is possible to access the raw probabilities output by the model. This is important both for interpretability and for quantifying the uncertainty inherent to generation by LLMs, and can mitigate concerns about their black-box nature by making it easier to directly examine the probabilities associated with other likely generations.

Running open-source models, whether locally or on a remote server, can present several difficulties for researchers. LLMs tend to be large, and require being loaded onto GPU or TPU memory for inference speeds to be fast enough to be practical for applied research. Though GPU prices have fallen in recent years, they can still represent a significant cost whether researchers purchase their own GPU and run locally or have access to powerful servers

(from this perspective, open-source models can be described as "pay-to-host"). Ongoing research, combined with efforts by the open-source community, have been able to reduce the hardware requirements to run LLMs locally. Reducing the size of model weights via quantization and offloading LLM layers to CPUs have both proven successful at reducing barriers to entry. These problems are exacerbated as model size continues to increase. Though larger models are generally more powerful, they also require significantly more resources to run than small ones. One promising alternative for researchers is to fine-tune small models, training them on the specific tasks required for research (Hu et al., 2021). Refer to Section 6 for more details about how to use open-source models for applied research in practice.

4. Potential applications of LLMs in political science

One of the primary advantages of LLMs over previous models is their flexibility. It is often easier to change a prompt than retrain a model or use a new one. In this section, we highlight ways in which political science may be affected by this emerging technology.

4.1. Studying the effects of LLMs

4.1.1. Prevalence and impact of AI-generated content

Political campaigns are already using LLMs to generate advertising content. A recent example is a widely viewed dystopian advertisement on YouTube, titled "Beat Biden"[1] and posted by the official RNC YouTube account. This advertisement is described as "An AI-generated look into the country's possible future if Joe Biden is re-elected in 2024." There is no doubt that we will see an explosion of advertisements generated by LLMs and other generative AI tools in future elections (Alvarez et al., 2023). In Figure 1, we give an example about how such a workflow might be developed.

These technologies are also being used by popular political news media like *The Daily Show*[2] (Bloomberg, 2023). Other news media organizations will no doubt be using LLMs to generate content for the 2024 U.S. elections and for other elections in other nations. We anticipate that the lines between human-generated news media content and content generated by LLMs and AI will start to blur in the next few years.

We also anticipate that these methodologies will be used to generate misinformation and fake news content. We will see fake campaign ads, false and misleading fake news content, content that deliberately misinforms voters about the election process to disenfranchise them, and many other uses of these technologies to produce misleading and misinforming materials. LLMs will also be used by those producing fake and misleading political content to aid in avoiding detection and to help develop strategies to facilitate the distribution and targeting of this fake and misleading content.

LLMs and other generative AI methods have dramatically reduced the costs of producing highly-realistic and seemingly real content. As a proof of concept demonstration, one of the authors used various free or inexpensive LLM-based methods to generate a quite realistic news interview and also a realistic short (but false) news story.[3] The content we produced was made within 30–60 min, and demonstrates the ease with which content that could appear very realistic on social media platforms, or on the small screen, can be generated.

These many uses of LLMs in politics, media, and the government, open the door for important research opportunities for social scientists. On one hand, studying how candidates use LLMs and generative AI in their campaigns will no doubt become an important research area for those who study political communications. In particular, it will be important to see how the new technologies change the development and distribution of campaign materials. As LLMs have dramatically reduced the costs of producing content, especially audio and video content that in the past would require studios and production teams, we anticipate that many campaigns that in the past would not have used these types of materials will now do so—for example, candidates running for local and municipal offices, who may now use highly polished video materials in their campaigns. For the campaigns of national interest that already spend huge amounts in digital advertising, LLMs will likely be used to target increasingly specific groups of voters. For example, presidential campaigns could use LLMs to generate many different versions of campaign ads, explaining the ramifications of their policy positions to individual towns, and even referencing specific local issues and landmarks. The consequences of the coming surge of individually-tailored campaign ads are not yet well-understood. Nor are the strategies that campaigns will actually employ. As it will become increasingly difficult to track political messaging, campaigns may even be able to claim opposite policies to different low-information groups of voters without consequence.

4.2. LLMs and politics: threats and counter-measures

The potential for LLMs to affect electoral outcomes has become an area of concern for policymakers and academics. In this section, we detail ways in which LLMs are likely to affect future elections and highlight potential counter-measures.

4.2.1. Threats to electoral integrity and political information

It is well-known that social media is a vector of fake news and misinformation that can threaten elections (Allcott and Gentzkow, 2017). The advent of LLMs will likely increase the amount of misinformation voters are exposed to, as they will dramatically drive down the cost of "micro-targeted" messages.

One might imagine a future where misinformation campaigns subtly instruct voters to print their name on ballots in states

1 https://youtu.be/kLMMxgtxQ1Y

2 https://www.youtube.com/watch?v=JONzK-AUzro

3 These can be viewed at https://bit.ly/46M67yo and https://bit.ly/3pA1as4.

where signature matching is pivotal. Such nuanced misinformation could invalidate thousands of votes. Similarly, the intricacies of voter registration, often a labyrinthine process for many, can be made even more convoluted by LLM-generated misinformation, preventing eligible citizens from casting their votes.

The capabilities of LLMs extend beyond mere text. Video content, once the domain of high-cost productions, can now be manipulated or entirely fabricated with the assistance of these models. This raises the possibility of misleading videos, portraying politicians in fabricated scenarios, or twisting real events to fit a particular narrative (see Figure 1).

The ability of LLMs to produce vast amounts of genuine content can also be leveraged to cheaply establish trust with audiences. Once this trust is cultivated, it becomes easier to introduce misinformation into the stream of content, making it harder for consumers to discern fact from fiction.

Beyond the electoral sphere, we should also consider the implications of LLMs in times of crisis. Misinformation during emergencies can have dire consequences. Picture a scenario where false emergency alerts or misleading updates are disseminated during critical events, leading to public panic or misallocated resources.

4.2.2. Counter-measures and safeguards

Countering the threats posed by LLMs requires a multifaceted approach. Public awareness campaigns can play a pivotal role. By illuminating the capabilities and potential pitfalls of LLM-generated content, we can arm the public with the knowledge to critically evaluate the information they consume. However, as LLMs become more sophisticated, it will become increasingly difficult for consumers to identify whether a given piece of content is artificially generated, or whether the information conveyed is both factual and has the appropriate context.

One proposal that has gained traction is "watermarking" content generated by LLMs. While this proposal will not be sufficient to curb misinformation (it is easy to train LLMs not to contain a watermark, and very difficult to identify whether a given piece of text is machine-generated), it may be able to provide a means of verification that a given LLM has been trained to be more factual, or that the LLM has been endorsed by a particular organization.

Trained appropriately, LLMs can aid in the rapid fact-checking of content, flagging inconsistencies or potential falsehoods. Doing so at scale will likely require collaboration with social media platforms. By monitoring and flagging content suspected to originate from LLMs, platforms can provide users with the context needed to evaluate the information.

Thus, if we are correct and there will also be widespread development and distribution of fake and misleading content, this will also open the door for researchers. It will be important to continue to develop methodologies for the real-time detection of false and misleading information online (e.g., Srikanth et al., 2021). Researchers will need to continue to study how misinformation is processed by individuals and to develop means to counter false and misinformative content (e.g., van der Linden, 2023). These are areas of research that will need substantial resources and cooperation with private organizations if academic researchers are to be effective in helping to deflect the effects of political disinformation.

4.3. Research uses of LLMs

4.3.1. Replacing manual processes

The practical use-cases of LLMs extend across the spectrum of political science and computational social science research (Ziems et al., 2023). One of the most significant benefits of deploying LLMs is their ability to replace manual annotation efforts, particularly in processing political content. Leveraging the learning capabilities of LLMs, researchers can efficiently identify elements like toxicity, political polarity, sentiment, and hate speech within a text. Such use-cases have been addressed by tools like ToxicBERT and Perspective API, which harness the power of LLMs to automate tasks traditionally performed by humans (Kocielnik et al., 2023a; Liang et al., 2023; Mendelsohn et al., 2023).

When it comes to information extraction, LLMs exhibit substantial advantages over conventional NLP algorithms. While pre-existing NLP algorithms may perform exceptionally well at specific tasks, they often fall short in dealing with tasks that require an understanding of context. For instance, if a researcher aims to identify the politicians mentioned in a tweet and the sentiment expressed toward them, an LLM using few-shot learning is likely to outperform separate Named Entity Recognition and Sentiment Analysis algorithms. The underlying reason lies in the superior ability of LLMs to interpret context, making them flexible tools for complex NLP tasks. This understanding of context is part of what makes LLMs useful for document summarization, a potentially useful tool for understanding the large bodies of text often found in political science research.

Moreover, LLMs can play a crucial role in generating new content for research purposes. Tools like *AutoBiasTest* (Kocielnik et al., 2023b) and other model written evaluations (Perez et al., 2022) are especially noteworthy for generating politically biased content, which researchers can further analyze to study the dynamics of political bias. Furthermore, LLMs can be prompted to generate particular types of content, such as hate speech, toxicity, and stereotypes related to political stances, which can be used to study the social perceptions in the underlying data or for experiments related to political communication (West, 2023). LLMs have also been used in a variety of other text generation applications. For instance, LLMs have been used to inject persuasive content at scale (Jingnan, 2023), and to generate realistic data to simulate multiple humans and replicate human subject studies (Aher et al., 2023).

4.3.2. Understanding political speech

Slanted news drives polarization (Martin and Yurukoglu, 2017), but the aspects of political speech that affect political polarization and ideology are not themselves well-understood. In part this is because the systematic analysis of speech is difficult. LLMs can aide social scientists by simplifying the process of information extraction, for example by discretizing pieces of text into variables relevant to the research at hand, summarizing

lengthy texts, categorizing qualitative data, classifying sentiment, or by otherwise reducing the dimensionality of the space of text. This concept is similar to dimension reduction techniques, where high-dimensional data is transformed into a lower-dimensional space, preserving as much relevant information as possible while discarding noise or redundancy. Put differently, we can use the complexity of LLMs to ask simple and easily interpretable questions about observed political speech. This provides a systematic way to understand how specific variations in speech affect its persuasiveness. By focusing on easy-to-interpret answers, this process is easy to validate, and can replace what would otherwise require intensive human effort and discretion.

Another approach is to build on foundational models for understanding political speech like NOMINATE scores (Poole and Rosenthal, 1985). One example of this is in trying to estimate the ideological position of different speakers on cable news. LLMs can simplify the process, for example by providing a new way of mapping political speech from a domain that is less-well understood like slanted cable news to pre-existing ideological spaces that researchers may be more comfortable with.

4.4. Toward multimodal research

The application of LLMs in political science research presents an opportunity for a unified approach to analyze multimodal data. This approach consists of interconverting different types of data, say a campaign video and its textual summary, facilitating the understanding of one medium in terms of the other. To illustrate, a researcher might transform a campaign video into a compact textual format by using an LLM trained for speech recognition to transform spoken language into written text (Radford et al., 2023), and use a multi-modal model to identify and describe key visual components of the video, retaining pertinent characteristics. The summary can then be used to regenerate a similar video. This ability of LLMs like GPT-4 to process and translate between text and image inputs, offers a consistent methodology to map from the video and transcript space to a smaller and more informative summary space. We anticipate this technology will be particularly useful in new areas of research in political communication, including for studying the contents of political advertisements, the dynamics of televised appearances, how news media frames particular political candidates, and for analyzing online political discussions.

5. Issues with using LLMs in research

The output of LLMs is inherently a function of the data used to train it. This means that biases present in the training data are likely to be perpetuated in the final model. Large-scale scrapes of the internet (such as those used in The Pile Gao et al., 2020, a common dataset used for training LLMs), are likely to contain large amounts of such bias. Though efforts can be made to censor the model's output, this censorship will only prevent the LLM for recreating the internet's worst tendencies, not prevent it from perpetuating subtle biases.

How social bias manifests itself in the output of LLMs is thus a key concern when deploying LLMs. Biased data can lead models to disproportionately represent majority viewpoints, leading to a systematic marginalization of minority data. Such bias can subtly yet profoundly skew the outputs of the models, creating outcomes that may reinforce existing societal prejudices. In the context of political discourse, this could manifest as a bias toward particular political parties or viewpoints, potentially influencing research conclusions. For example, Motoki et al. (2023) find that ChatGPT displays bias toward the Democratic party, and Feng et al. (2023) find that language models have political leanings and can reinforce political polarization and other social biases. Though uncensored versions of models can mitigate this problem to some degree (see Section 7 for more details), these models will likely perpetuate bias in other directions.

Several types and sources of social bias in AI systems have been identified in prior work (Mehrabi et al., 2021). Issues of potential social and stereotypical bias need to be carefully considered when applying AI and especially LLMs to analysis on real-world data or in applications impacting society. Google (2022) presents a detailed discussion about social bias and AI. There are two primary categories of social bias political scientists should be concerned with. The first categories concerns the data used to train LLMs. These issues can generally be boiled down to *Reporting Bias* and *Selection Bias*. The second category concerns the output of LLMs, whether this output constitutes labels or generated text. These issues can generally be described as *Group Attribution Bias* or extensions of observed *Implicit Bias*. Both involve the tendency to stereotype minority groups, which may result in different degrees of accuracy for LLM-based classifiers across groups. Abid et al. (2021) provide a clear example of this, showing that LLMs can perpetuate biases against minority groups through stereotypes. In particular, they show that LLMs can convey a strong association between Muslims and violence.

5.1. Fairness in AI

Ensuring fairness in AI is challenging due to the lack of interpretability of the models, and the bias present in the training data, yet crucial, due to the pervasive integration of AI and machine learning systems in diverse applications with direct societal impact. These applications range from court systems assessing reoffending probabilities, to medical fields, childhood welfare systems (Chouldechova et al., 2018), and autonomous vehicles. When applying AI to practical scenarios and real-world data it is important to consider aspects of fairness as inherent biases can have detrimental effects on many levels. Osoba and Welser (2017) list examples of biases in real-world applications of AI, including bias in AI chatbots, employment matching, flight routing, automated legal aid for immigration algorithms, and search and advertising placement algorithms (Osoba and Welser, 2017). Bias can also manifest in real-world AI and robotic systems, such as face recognition and voice recognition applications, and search engines (Howard and Borenstein, 2018).

Discriminatory behavior in AI systems is a notable problem. For instance, the COMPAS risk assessment software has been identified as biased and its performance questioned when compared to human judgment (Lambrecht and Tucker, 2019).

Another example is an advertisement algorithm promoting STEM field jobs which, despite intending to be gender-neutral, was found to show fewer ads to women due to the gender imbalance in the target audience. Bias has also been observed in facial recognition systems (Buolamwini and Gebru, 2018; Raji and Buolamwini, 2019) and recommender systems (Schnabel et al., 2016), often leading to discrimination against certain populations and subgroups. These examples highlight the critical need for engineers and researchers to understand the sources of these biases and take preventive measures during the design and engineering process of AI systems. Ensuring fairness is not just about maintaining balance, but about minimizing harm and potential detrimental effects on society.

Fairness in AI, as examined across various research, encompasses numerous definitions rooted in different fields including political philosophy, education, and machine learning (Mehrabi et al., 2021). Various definitions of fairness in AI exist, making it challenging to achieve fairness in practices across these competing definitions. These definitions can be grouped into three broad categories: *Individual Fairness*, which emphasizes similar predictions for similar individuals; *Group Fairness*, which advocates for treating different groups equally; and *Subgroup Fairness*, a hybrid approach aiming to combine the benefits of both individual and group fairness.

5.2. Other challenges

Another issue is that LLMs are sensitive to variations in semantically irrelevant inputs[4], making it difficult to discern the effects of seemingly minor differences across prompts. Adding to this are the difficulties presented by hallucinations, where models generate false data due to high probability assigned to untrue statements. Detecting and mitigating these hallucinations in automated ways is particularly challenging. Since the hallucinations are often based on the model's training data, they can be highly context-specific and difficult to predict. Furthermore, because LLMs generate outputs probabilistically, they may not consistently produce the same hallucinations, making these errors even harder to catch.

5.3. Ethical considerations

Deploying LLMs in social science research has important ethical considerations. The use of these models often requires trusting in the "black box" nature of the algorithms, particularly those developed by private-sector entities. This means that any biases in the training data can inadvertently be introduced into the research, perpetuating subtle biases that can skew the results. It is crucial for researchers to remember that LLMs are not a definitive source of truth, but rather a representation of the data they were trained on. Thus, by accepting a model without a thorough examination of its biases, researchers implicitly trust that the creators of the model have trained it with a bias they find acceptable.

Another issue in using LLMs pertains to the lack of explicit consent in using individuals' data for model training. This concern is particularly pronounced in models designed for chat functionalities, such as OpenAI and Google's Bard. In these instances, the models' terms of use often allow the use of chat data for further model training, potentially infringing upon user privacy. Moreover, this issue extends to intellectual property rights, as demonstrated by ongoing lawsuits by artists against organizations like OpenAI.[5] These cases underscore the concern over the use of copyrighted material within the training data, again without explicit permission.

6. LLMs and reproducible research

Research replication and reproducibility have long been important research best practices in political science (King, 1995; Alvarez and Heuberger, 2022). LLMs present important challenges for researchers and publishers with respect to replication, in particular regarding transparency in LLM development, clarity around the datasets and benchmarks used, standardized model evaluation rankings, and journal policies for provision of replication materials prior to publishing.

6.1. LLM development transparency

Efforts in enhancing the transparency of AI models have been gaining momentum over the last few years. Prominent initiatives in this regard include the use of *"Model Cards"* (Mitchell et al., 2019) which provide a detailed snapshot of a model's purpose, performance, limitations, and biases in a structured manner. They act as a kind of report card, providing relevant information about a model, and making it more interpretable and explainable to end-users. Unfortunately, the level of detail and the quality of a Model Card relies on the voluntary effort of model developers. Platforms such as HuggingFace provides guides for creating high-quality model cards (Hugging Face, 2023a), but these are not always followed and in many cases, Model Cards are prefilled automatically, often resulting in their poor quality.

Recent initiatives such as *"Interactive Model Cards"* (Crisan et al., 2022), *"AutoBiasTest"* (Kocielnik et al., 2023b), and model written evaluations focus on interactive tools, that support live exploration of model capabilities and limitations using generated datasets or human-in-the-loop evaluation. Recent efforts in AI transparency put emphasis on user-friendly tools that can be used by various non-AI experts with relevant social expertise (e.g., social scientists, gender studies researchers, ethics experts) as well as practitioners in domains where various AI tools can be applied (e.g., clinicians, chemists, content writers; Kocielnik et al., 2019; Rastogi et al., 2023). These efforts collectively aim to demystify AI, making it more accessible, understandable, and ultimately more accountable.

4 https://huggingface.co/blog/evaluating-mmlu-leaderboard?fbclid=IwAR04lwlW3eZTXz7YBxpgL7F4b1paMwmYpuo4mdKNgtMkTIRs7Ja5x7GUAx4

5 https://ymcinema.com/2023/02/15/midjourney-is-being-class-action-sued-for-severe-copyright-infringements

6.2. Datasets and benchmarks transparency

In relation to data used for model training and evaluation, the prominent NeurIPS conference recently introduced a separate track called the "Datasets and Benchmarks Track" focused specifically on obtaining high-quality datasets and benchmarks, but also on refinement of existing datasets (Denton et al., 2023). To try to enforce high quality of submissions, this track requires the use of several transparency tools related to datasets. The "Datasheets for Datasets" (Gebru et al., 2021) initiative encourages comprehensive documentation for datasets used in AI model training, including data collection processes, motivations, biases, and ethical considerations. This can be likened to "nutrition labels" for data (Holland et al., 2020), offering a transparent look at the raw materials that feed into AI systems. Furthermore, the development and implementation of accountability frameworks are essential to ensure that those who develop and deploy AI systems are held responsible for their actions.

6.3. Model evaluation rankings

Arguably, most of the well-established evaluation practices for LLMs are various benchmarks and rankings (Ramanathan, 2022; Ceylan, 2023). Open LLM Leaderboard (Hugging Face, 2023b) and Super-GLUE (Wang et al., 2019) (a benchmark suite designed to evaluate the performance of LLMs on a range of demanding natural language understanding tasks) are some of the popular benchmarks that provide frameworks for comparing and evaluating these models on various aspects, such as accuracy, fluency, coherence, and subject relevance.

Benchmarking LLM performance requires careful selection of evaluation tasks, data preparation, and comparative analysis. Benchmarking LLMs is crucial not just for performance assessment, but also for detecting and mitigating biases, and assessing user satisfaction and trust (Huang et al., 2023). Different evaluation methods such as Perplexity (Chiusano, 2022), human evaluation (Liang et al., 2022), BLEU, ROUGE (Santhosh, 2023), social bias scores (Delobelle et al., 2022), and diversity measures are used for different aspects of performance. However, they come with challenges such as subjectivity in human evaluations, lack of diversity in metrics, lack of generalization to real-world scenarios, and susceptibility to adversarial attacks.

To overcome these challenges, best practices involve using multiple evaluation metrics, enhancing human evaluation, creating diverse reference data, incorporating diverse metrics, augmenting evaluation methods with real-world scenarios, and evaluating LLMs for robustness against adversarial attacks. Several popular AI development frameworks offer standardized evaluation of code-based tools that can be run by developers of LLMs and reproduced by other researchers. The most popular one is arguably Eluther AI's LM evaluation harness (ElutherAI, 2023).

6.4. LLMs and replication materials

Most political science research journals either require that authors make their code and data available upon publication in a public and permanent repository, or they strongly suggest that authors follow this best practice upon publication. Usually this means that authors will provide some documentation regarding how they collected, preprocessed, analyzed and presented the data—usually in the form of code and the actual data itself.

Research journals and professional societies need to provide guidance for authors about how to document their use of public or private LLMs, and give authors detailed information about what information meets the standard for good replication materials. If an author develops and uses their own LLM for a research project, that might present other challenges for journals, in particular with respect how to archive and curate the training data used for the researcher's LLM. Training datasets may be large and thus require significant storage space, and they main contain information that might be difficult to make public (for example, for copyright or privacy reasons). Guidelines for researchers about archiving and curating their LLMs to meet professional best practices for replication are needed.

7. A practical guide to using LLMs in political science research

LLMs have wide-ranging applicability, from classification to document summarization to sentiment analysis, as previously discussed in Section 4. While these models exhibit good performance on many tasks out of the box, researchers can further enhance their results by providing additional training data. This could entail creating longer, more detailed prompts for few-shot learning, or fine-tuning the model using techniques such as Parameter Efficient Fine-tuning (PEFT) or Low-Rank Adaptation (LoRA). While powerful, techniques like PEFT and LoRA require the model architecture and weights to be known, which is typically only the case with open-source models.

PEFT optimizes the training process by fine-tuning only a subset of the model's parameters, while LoRA trains a rank-decomposition matrix, adding this to pre-trained weights. The rank-decomposition matrix is small relative to the pre-trained weights, which are kept frozen. Both techniques help reduce the computational cost and hardware requirements, and make LLMs a more accessible tool for researchers with limited resources.

However, their flexibility does not make LLMs a text data panacea. They can be slow and inefficient compared to other text-processing methods. When complexity arises, LLMs offer more flexible problem-solving approaches, albeit at the risk of hallucination – fabricating information with unwarranted confidence. They may also struggle with domain-specific tasks, such as mathematical problems. Generally, it is advised to avoid LLMs where these issues may be problematic.

HuggingFace is a prominent platform in the field of NLP and machine learning, and is the recommended way to find and download individual LLM models. It provides a comprehensive, open-source library of pre-trained models for a multitude of NLP tasks, making it an invaluable resource for researchers. With HuggingFace, researchers can easily download different models and even upload their custom-trained models, facilitating the sharing of research outputs. Moreover, its user-friendly codebase aids in streamlining machine learning tasks. It offers various libraries that

ease the implementation of complex tasks like operations across multiple GPUs, model quantization, and the implementation of techniques like PEFT, LoRA, and various methods for quantization. Given these attributes, HuggingFace significantly reduces the barriers to entry for researchers venturing into the realm of advanced NLP and LLMs.

The choice of model depends largely on the research objectives, hardware constraints, and the need for high-speed computations. Researchers grappling with hardware limitations may consider fine-tuning smaller models, which can deliver comparable performance to larger, unspecialized models. Quantization techniques have also proven effective in reducing model size and enhancing speed. Notably, combining these techniques can yield substantial performance improvements (Dettmers et al., 2023).

A pertinent issue in the use of LLMs, particularly in political science research, is model censorship. For instance, models like ChatGPT may censor requests related to certain political figures like Donald Trump but not others like Joe Biden. This black-box nature of LLMs introduces biases stemming from the decisions of those who train the models, leading to "censorship" that can limit the scope of research. This issue persists even with open-source models. However, recent efforts in the open-source community aim to release "uncensored" models. Such uncensoring improves the models' capability to handle potentially controversial topics, and is especially important when deploying these models in politically charged environments.

7.1. Model selection

There are a dizzying number of LLMs available. How are researchers to know which to use? In addition to using models that perform well on the public benchmarks we have suggested, researchers can choose a particular family of "foundational" models to work with, finetuning for their particular goals. In this section we highlight a few of the key models that have been developed recently. While most of our discussion has centered around "text-to-text" models, the ability of LLMs to perform across modalities is one of the abilities we hope to highlight, so we draw a distinction between models accepting different types of inputs and outputs.

In Table 1, we provide a brief overview of the most commonly used LLMs for various tasks. In order to give a sense of the data used to train these models, we also detail some of the data used to train these models.

TABLE 1 Bird's-eye view of the LLM landscape mapping tasks, models, and datasets.

Model type	Example tasks	Example models	Example datasets
Text-based	Text classification	GPT-4	Wikipedia
	Text labeling	T5	Common Crawl
	Text generation	LlaMA	BooksCorpus
		Falcon	
Image-based	Image generation	Midjourney, DALL-E	ImageNet
	Image-to-text	GPT-4	COCO
	Text-to-image	Midjourney, DALL-E	MS COCO captions

Note that all text-based tasks can be performed using any of the example models.

7.2. Text to text models

These models focus on generating or transforming textual data. They play a crucial role in Natural Language Processing (NLP) tasks such as language translation, summarization, question-answering, text generation, and text-focused processing.

GPT-3.5 and GPT-4: OpenAI's GPT-3.5 and GPT-4 are advanced language models driving ChatGPT-3.5 and ChatGPT-4 chatbot applications. GPT-4 excels over GPT-3.5 in size, computational power, and memory, handling complex tasks and longer conversations. Although slower and having an hourly prompt limit, GPT-4 can process visual inputs and retain more data during a chat session.

T5 (Text-To-Text Transfer Transformer): This model has been created by Google. T5 frames every NLP task as a text generation problem, making it highly versatile for various tasks. This class of models is especially useful for embedding large bodies of text, and T5 models have consistently topped HuggingFace's Massive Text Embedding Leaderboard (Muennighoff et al., 2023) since they were released.

LLaMA (Large Language Model Meta AI): LLaMA, introduced by Meta AI, is a state-of-the-art, foundational language model designed to democratize access to large language models. Available in several sizes (7B, 13B, 33B, and 65B parameters), these models require less computational power and are ideal for fine-tuning across various tasks. Despite presenting some common challenges of large language models, like bias and toxicity, LLaMA provides a platform for researchers to test solutions for these problems. The models are released under a noncommercial license for research use cases.

Open-source LLaMA-based models: Building on the foundation set by LLaMA, the open-source and research communities developed an array of language models that harness the robustness and accessibility of the LLaMA framework. This lineage includes various families of LLMs, such as those influenced by Alpaca (Taori et al., 2023), Vicuna (Chiang et al., 2023), Guanaco (Dettmers et al., 2023), and WizardLM (Xu et al., 2023). Each of these derivatives embodies the democratized vision of LLaMA, optimized to run efficiently on consumer-grade hardware. The models are not only easy to use but also designed for straightforward fine-tuning, making them highly adaptable to specific research tasks or applications. These LLaMA-based models form an important pillar in the landscape of open-source AI, embodying the intersection of state-of-the-art performance and the ethos of open, accessible AI research.

Falcon: A state-of-the-art language model family created by the Technology Innovation Institute in Abu Dhabi and released under the Apache 2.0 license (von Werra et al., 2023). Falcon-40B and Falcon-7B are the two base models in this family, with the former topping the charts of the Open LLM Leaderboard

and the latter being best in its weight class. Falcon-40B rivals the capabilities of many current closed-source models, and notably, it is an open-sourced model.

7.3. Text to image

Another category of models takes textual descriptions and transforms them into visual counterparts (Gozalo-Brizuela and Garrido-Merchan, 2023; Zhang et al., 2023b). These models employ a two-step process: the language model first changes the input text into a latent representation, followed by a generative image model that creates an image conditioned on that representation (Borji, 2022). There are several popular models in this space.

MidJourney: This AI-driven tool stands out in the arena of text-to-video platforms by efficiently transforming textual prompts into corresponding images. It demonstrates a special capability in adapting real art styles to create an image of any combination of things the user desires, with an emphasis on creating environments, especially fantasy and sci-fi scenes (Vartiainen and Tedre, 2023). Its dramatic lighting effect makes it appear as though the images were rendered concept art from a video game. Notably, MidJourney is known for its distinct artistic style, and its Discord bot integration adds convenience for the users.

Stable Diffusion: Developed by StabilityAI in 2022 (StabilityAI, 2022), Stable Diffusion is a text-to-image model that uses a unique diffusion process (Rombach et al., 2022). Its mechanism begins with just noise and gradually refines the image until it is completely noise-free, progressively aligning with the provided text description. It's powered by the Latent Diffusion Model (LDM), a state-of-the-art text-to-image synthesis method (Alammar, 2022). It balances between complexity reduction and detail preservation, usually resulting in high visual fidelity. Stable Diffusion is also open-source and capable of producing highly detailed artwork, but it needs an interpretation of complex original prompts.

DALL-E: Created by OpenAI, the DALL-E model, and its successor DALL-E 2 also produce images from text prompts (Dayma et al., 2021). They've been trained on more than 10 billion parameter versions of the GPT-3 transformer model, which allows them to interpret natural language inputs and generate corresponding images. The system primarily consists of two components: one that changes the user input into an image representation (called Prior), and another that converts this representation into an actual image (called Decoder). The textual and image embeddings used by DALL-E are derived from another network called CLIP (Contrastive Language-Image Pre-training), also created by OpenAI. This model is known for creating sophisticated output images with high level of detail.

7.4. Image to text

Several models also support image-to-text generation. Their goal is to convert visual data into textual information. They are utilized in a range of applications including generating descriptive captions, object recognition, image-based searches, and accessibility features for visually impaired individuals. Some prominent models in this space include:

CLIP (Contrastive Language–Image Pretraining): Developed by OpenAI, it is a foundational image-to-text model trained on a large variety of image-text pairs, which enables it to understand and generate textual descriptions from images (Radford et al., 2021). Unlike models that only understand images or text, CLIP jointly learns to understand both, allowing it to connect the dots between visual and linguistic information, thus leading to more accurate and detailed descriptions. It is worth noting that CLIP itself can't generate text, but it can evaluate how appropriate a given sentence is as a caption for a given image.

VisionEncoderDecoder: It is a versatile open-source image-to-text model that can integrate any pre-trained Transformer-based vision model as the encoder (like ViT, BEiT, DeiT, Swin) and any pre-trained language model as the decoder (like RoBERTa, GPT2, BERT, DistilBERT). This model is adaptable and has multiple applications. It can be utilized in image captioning where the encoder encodes the image and an autoregressive language model generates the caption. It is also employed in Optical Character Recognition (Li et al., 2023).

Xception: It is a caption generator model using a pre-trained deep learning network called Xception, which generates descriptive text captions for images (Chollet, 2017). This model has shown effectiveness due to Xception's architecture that performs depth-wise separable convolutions for increased efficiency.

8. Discussion and conclusion

Large Language Models have seen rapid development in recent years and we expect the continued evolution of LLMs to continue in the near future. LLMs will be trained on increasingly larger (and hopefully more representative) datasets, they will be made easier to use, and we anticipate that they will become an important component of the tool kit for political and social scientists. Among the developments that we have argued are necessary are increasing the transparency of the models, improving their interpretability and explainability, and reducing their bias.

At the same time, we expect to see an explosion in the use of LLMs, particularly in electoral politics but perhaps also in other areas of governance and policymaking. These applications will spark additional research opportunities for political scientists. We will need to understand better how electoral campaigns use LLMs, both for the development of legitimate and informative communications, but also for the production of misleading and misinforming communications. Governmental agencies will start to use LLMs for many purposes, for example, chatbots that can interact with citizens. These uses will need research, and for researchers to scrutinize the LLMs and how the are trained to help prevent biases from these models.

We also expect that the public will continue to use LLMs in many ways. People will uses LLMs to manage their communications, answering email and creating social media posts. LLMs will be used by students, replacing flashcards and tutors, and also meaning that their parents will need to become proficient with the tools that their children are using. Many businesses will use LLMs, to build chatbots for communications but to also simply or automate many simple and routine tasks—drafting legal agreements, writing news reports, and developing advertising materials, for example. These public uses of LLMs will intersect

with government and politics in important ways, again creating new research opportunities for political scientists to study.

While LLMs hold great promise for political scientists, we also are concerned about their ability to quickly and inexpensively churn our false and misleading information. If misinformation becomes rampant, especially in future elections, that could lead to significant calls for the regulation of the technology. Some today are even calling for halting the development and public dissemination of LLMs, which could have a chilling effect on their evolution in nations or regions that introduce strong regulatory models for LLMs and more generally, AI. The possibility that regulation might be introduced for LLMs should be a call for researchers to understand this new technology, to help build them in ways that ensure their transparent and fair use, and to help policy makers navigate how to support the development and use of LLMs in ways that mitigate social and political harm.

Author contributions

RA: Conceptualization, Funding acquisition, Investigation, Methodology, Project administration, Supervision, Writing—original draft, Writing—review and editing. ML: Conceptualization, Investigation, Methodology, Writing—original draft, Writing—review and editing. RK: Conceptualization, Investigation, Methodology, Writing—original draft, Writing—review and editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Acknowledgments

We thank the Caltech Center for Science, Society, and Public Policy for supporting our research on the Ethics of AI.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Abid, A., Farooqi, M., and Zou, J. (2021). Large language models associate Muslims with violence. *Nat. Mach. Intell.* 3, 461–463. doi: 10.1038/s42256-021-00359-2

Aher, G. V., Arriaga, R. I., and Kalai, A. T. (2023). "Using large language models to simulate multiple humans and replicate human subject studies," in *International Conference on Machine Learning* (PMLR), 337–371.

Alammar, J. (2022). *The Illustrated Stable Diffusion: Visualizing Machine Learning One Concept at a Time*. Available online at: https://jalammar.github.io/illustrated-stable-diffusion/ (accessed July 07, 2023).

Allcott, H., and Gentzkow, M. (2017). Social media and fake news in the 2016 election. *J. Econ. Perspect.* 31, 211–36. doi: 10.1257/jep.31.2.211

Alvarez, R. M., Eberhardt, F., and Linegar, M. (2023). *Generative AI and the Future of Elections*. Caltech Center for Science, Society, and Public Policy (CSSPP) Policy Brief.

Alvarez, R. M., and Heuberger, S. (2022). How (not) to reproduce: Practical considerations to improve research transparency in political science. *Polit. Sci. Polit.* 55, 149–154. doi: 10.1017/S1049096521001062

Bloomberg (2023). *Generative AI Takes Stereotypes and Bias from Bad to Worse*. Available online at: https://www.bloomberg.com/graphics/2023-generative-ai-bias/ (accessed July 07, 2023).

Borji, A. (2022). Generated faces in the wild: Quantitative comparison of Stable Diffusion, Midjourney and DALL-E 2. *arXiv preprint arXiv:2210.00586*. doi: 10.48550/arXiv.2210.00586

Buolamwini, J., and Gebru, T. (2018). "Gender shades: intersectional accuracy disparities in commercial gender classification," in *Proceedings of the 1st Conference on Fairness, Accountability and Transparency, volume 81 of Proceedings of Machine Learning Research*, eds S. A. Friedler and C. Wilson (PMLR), 77–91.

Ceylan, B. (2023). *Large Language Model Evaluation in 2023: 5 Methods*. Available online at: https://research.aimultiple.com/large-language-model-evaluation/ (accessed October 07, 2023).

Chiang, W.-L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., et al. (2023). *Vicuna: An Open-Source Chatbot Impressing GPT-4 With 90% ChatGPT Quality*. Available online at: https://lmsys.org/blog/2023-03-30-vicuna/

Chiusano, F. (2022). Two Minutes NLP–Perplexity Explained With Simple Probabilities. Available online at: https://bit.ly/3PSLKtr (accessed October 07, 2023).

Chollet, F. (2017). Xception: deep learning with depthwise separable convolutions. *arXiv preprint arXiv:1610.02357*. doi: 10.48550/arXiv.1610.02357

Chouldechova, A., Benavides-Prado, D., Fialko, O., and Vaithianathan, R. (2018). "A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions," in *Conference on Fairness, Accountability and Transparency* (PMLR), 134–148.

Crisan, A., Drouhard, M., Vig, J., and Rajani, N. (2022). "Interactive model cards: a human-centered approach to model documentation," in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 427–439.

Dayma, B., Patil, S., Cuenca, P., Saifullah, K., Abraham, T., Le Khac, P., et al. (2021). *Dall-e Mini*. Available online at: https://github.com/borisdayma/dalle-mini

Delobelle, P., Tokpo, E. K., Calders, T., and Berendt, B. (2022). "Measuring fairness with biased rulers: a comparative study on bias metrics for pre-trained language models," in *NAACL 2022: the 2022 Conference of the North American chapter of the Association for Computational Linguistics: Human Language Technologies*, 1693–1706.

Denton, E., Ha, J.-W., and Vanschoren, J. (2023). "Neurips 2023," in *Thirty-seventh Conference on Neural Information Processing Systems*.

Dettmers, T., Pagnoni, A., Holtzman, A., and Zettlemoyer, L. (2023). QLORA: efficient finetuning of quantized LLMs. *arXiv preprint arXiv:2305.14314*. doi: 10.48550/arXiv.2305.14314

ElutherAI (2023). *EleutherAI/lm-Evaluation-Harness: A Framework for Few-Shot Evaluation of Autoregressive Language Models*. Available online at: https://github.com/EleutherAI/lm-evaluation-harness (accessed October 07, 2023).

Feng, S., Park, C. Y., Liu, Y., and Tsvetkov, Y. (2023). From pretraining data to language models to downstream tasks: tracking the trails of political biases leading to unfair NLP models. *arXiv preprint arXiv:2305.08283*. doi: 10.48550/arXiv.2305.08283

Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., et al. (2020). The pile: an 800GB dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*. doi: 10.48550/arXiv.2101.00027

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daume, H. III, et al. (2021). Datasheets for datasets. *Commun. ACM* 64, 86–92. doi: 10.1145/3458723

Google (2022). *Fairness: Types of Bias*. Available online at: https://developers.google.com/machine-learning/crash-course/fairness/types-of-bias (accessed November 07, 2023).

Gozalo-Brizuela, R., and Garrido-Merchan, E. C. (2023). ChatGPT is not all you need. A state of the art review of large generative AI models. *arXiv preprint arXiv:2301.04655*. doi: 10.48550/arXiv.2301.04655

Grimmer, J. (2013). *Representational Style in Congress: What Legislators Say and Why It Matters*. New York, NY: Cambridge University Press.

Grimmer, J., and Stewart, B. M. (2013). Text as data: the promise and pitfalls of automatic content analysis methods for political texts. *Polit. Anal.* 21, 267–297. doi: 10.1093/pan/mps028

Holland, S., Hosny, A., Newman, S., Joseph, J., and Chmielinski, K. (2020). The dataset nutrition label. *Data Protect. Privacy* 12, 1. doi: 10.5040/9781509932771.ch-001

Howard, A., and Borenstein, J. (2018). The ugly truth about ourselves and our robot creations: the problem of bias and social inequity. *Sci. Eng. Ethics* 24, 1521–1536. doi: 10.1007/s11948-017-9975-2

Hu, E., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., et al. (2021). LORA: low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*. doi: 10.48550/arXiv.2106.09685

Huang, Y., Zhang, Q., and Sun, L. (2023). TrustGPT: a benchmark for trustworthy and responsible large language models. *arXiv preprint arXiv:2306.11507*. doi: 10.48550/arXiv.2306.11507

Hugging Face (2023a). *Model Cards*. Available online at: https://huggingface.co/docs/hub/model-cards (accessed October 07, 2023).

Hugging Face (2023b). *Open LLM Leaderboard - a Hugging Face Space by HuggingFaceH4*. Available online at: https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard (accessed October 07, 2023).

Jingnan, H. (2023). *How Generative AI May Empower Political Campaigns and Propaganda*. Available online at: https://bit.ly/46DNzk5 (accessed November 07, 2023).

Kann, C., Hashash, S., Steinert-Threlkeld, Z., and Alvarez, R. M. (2023). Collective identity in collective action: evidence from the 2020 summer BLM protests. *Front. Polit. Sci.* 5, 1185633. doi: 10.3389/fpos.2023.1185633

King, G. (1995). Replication, replication. *Polit. Sci. Polit.* 28, 444–452.

Kocielnik, R., Amershi, S., and Bennett, P. N. (2019). "Will you accept an imperfect AI? exploring designs for adjusting end-user expectations of AI systems," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–14.

Kocielnik, R., Kangaslahti, S., Prabhumoye, S., Hari, M., Alvarez, R. M., and Anandkumar, A. (2023a). "Can you label less by using out-of-domain data? Active & transfer learning with few-shot instructions," in *Transfer Learning for Natural Language Processing Workshop* (PMLR), 22–32.

Kocielnik, R., Prabhumoye, S., Zhang, V., Jiang, R., Alvarez, R. M., and Anandkumar, A. (2023b). BiasTestGPT: using ChatGPT for social bias testing of language models. *arXiv preprint arXiv: 2302.07371*.

Lambrecht, A., and Tucker, C. (2019). Algorithmic bias? An empirical study of apparent gender-based discrimination in the display of STEM career ads. *Manage. Sci.* 65, 2966–2981. doi: 10.1287/mnsc.2018.3093

Laver, M., and Garry, J. (2000). Estimating policy positions from political texts. *Am. J. Polit. Sci.* 44, 619–634. doi: 10.2307/2669268

Li, M., Lv, T., Chen, J., Cui, L., Lu, Y., Florencio, D., et al. (2023). "TROCR: transformer-based optical character recognition with pre-trained models," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 13094–13102.

Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., et al. (2022). Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*. doi: 10.48550/arXiv.2211.09110

Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., and Zou, J. (2023). GPT detectors are biased against non-native English writers. *arXiv preprint arXiv:2304.02819*. doi: 10.48550/arXiv.2304.02819

Martin, G. J., and Yurukoglu, A. (2017). Bias in cable news: persuasion and polarization. *Am. Econ. Rev.* 107, 2565–2599. doi: 10.1257/aer.20160812

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Comput. Surv.* 54, 1–35. doi: 10.1145/3457607

Mendelsohn, J., Bras, R. L., Choi, Y., and Sap, M. (2023). "From dogwhistles to bullhorns: unveiling coded rhetoric with language models," in *ACL* (Toronto, ON).

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., et al. (2019). "Model cards for model reporting," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 220–229.

Motoki, F., Pinho Neto, V., and Rodrigues, V. (2023). More human than human: measuring ChatGPT political bias. *Public Choice*. doi: 10.1007/s11127-023-01097-2

Muennighoff, N., Tazi, N., Magne, L., and Reimers, N. (2023). MTEB: massive text embedding benchmark. *arXiv preprint arXiv:2210.07316*. doi: 10.48550/arXiv.2210.07316

Osoba, O. A., and Welser, W. IV (2017). *An Intelligence in our Image: The Risks of Bias and Errors in Artificial Intelligence*. Santa Monica, CA: Rand Corporation.

Perez, E., Ringer, S., Lukošiūtė, K., Nguyen, K., Chen, E., Heiner, S., et al. (2022). Discovering language model behaviors with model-written evaluations. *arXiv preprint arXiv:2212.09251*. doi: 10.48550/arXiv.2212.09251

Poole, K. T., and Rosenthal, H. (1985). A spatial model for legislative roll call analysis. *Am. J. Polit. Sci.* 29, 357–384.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., et al. (2021). "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning* (PMLR), 8748–8763.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. (2023). "Robust speech recognition via large-scale weak supervision," in *International Conference on Machine Learning* (PMLR), 28492–28518.

Raji, I. D., and Buolamwini, J. (2019). "Actionable auditing: investigating the impact of publicly naming biased performance results of commercial ai products," in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 429–435.

Ramanathan, B. (2022). *Evaluating Large Language Models (LLMs) with Eleuther AI*. Available online at: https://bit.ly/44mnm7R (accessed October 07, 2023).

Rastogi, C., Ribeiro, M. T., King, N., and Amershi, S. (2023). Supporting human-AI collaboration in auditing LLMs with LLMs. *arXiv preprint arXiv:2304.09991*. doi: 10.48550/arXiv.2304.09991

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. (2022). "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10684–10695.

Santhosh, S. (2023). *Understanding BLEU and ROUGE Score for NLP Evaluation*. Available online at: https://medium.com/@sthanikamsanthosh1994/understanding-bleu-and-rouge-score-for-nlp-evaluation-1ab334ecadcb#:~:text=While%20BLEU%20score%20is%20primarily,the%20reference%20translations%20or%20summaries (accessed October 07, 2023).

Schnabel, T., Swaminathan, A., Singh, A., Chandak, N., and Joachims, T. (2016). "Recommendations as treatments: debiasing learning and evaluation," in *International Conference on Machine Learning* (PMLR), 1670–1679.

Srikanth, M., Liu, A., Adams-Cohen, N., Cao, J., Alvarez, R. M., and Anandkumar, A. (2021). "Dynamic social media monitoring for fast-evolving online discussions," in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 3576–3584.

StabilityAI (2022). *Stable Diffusion v2.1 and DreamStudio Updates*. Available online at: https://stability.ai/blog/stablediffusion2-1-release7-dec-2022 (accessed July 07, 2023).

Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., et al. (2023). *Stanford Alpaca: An Instruction-Following LLaMA Model*. Available online at: https://github.com/tatsu-lab/stanford_alpaca

van der Linden, S. (2023). *Foolproof: Why Misinformation Infects our Minds and How to Build Immunity*. New York, NY: Norton.

Vartiainen, H., and Tedre, M. (2023). Using artificial intelligence in craft education: crafting with text-to-image generative models. *Digit. Creat.* 34, 1–21. doi: 10.1080/14626268.2023.2174557

von Werra, L., Belkada, Y., Mangrulkar, S., Tunstall, L., Dehaene, O., Cuenca, P., et al. (2023). *The Falcon Has Landed in the Hugging Face Ecosystem*. Available online at: https://huggingface.co/blog/falcon (accessed July 07, 2023).

Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., et al. (2019). SuperGLUE: a stickier benchmark for general-purpose language understanding systems. *arXiv preprint arXiv:1905.00537*. doi: 10.48550/arXiv.1905.00537

West, D. M. (2023). *Comparing Google Bard with OpenAI's ChatGPT on Political Bias, Facts, and Morality*. Available online at: https://bit.ly/44EEbe3 (accessed October 07, 2023).

Xu, C., Sun, Q., Zheng, K., Geng, X., Zhao, P., Feng, J., et al. (2023). WizardLM: empowering large language models to follow complex instructions. *arXiv preprint arXiv:2304.12244*. doi: 10.48550/arXiv.2304.12244

Zhang, C., Zhang, C., Li, C., Qiao, Y., Zheng, S., Dam, S. K., et al. (2023a). One small step for generative AI, one giant leap for AGI: a complete survey on chatgpt in AIGC era. *arXiv preprint arXiv:2304.06488*. doi: 10.48550/arXiv.2304.06488

Zhang, C., Zhang, C., Zhang, M., and Kweon, I. S. (2023b). Text-to-image diffusion model in Generative AI: A survey. *arXiv preprint arXiv:2303.07909*. doi: 10.48550/arXiv.2303.07909

Ziems, C., Held, W., Shaikh, O., Chen, J., Zhang, Z., and Yang, D. (2023). Can large language models transform computational social science? *arXiv preprint arXiv:2305.03514*. doi: 10.48550/arXiv.2305.03514