Synthetic Data for Small LLM De-Biasing

Date: 05/02/2024

Semester: Spring 2024

Course number and section: CIS 4250.002

Course instructor: Dr. John Licato

Problem addressed: Human Bias

Group members: Hyunjun Yoo, Ryan Feinberg, Raunak Chhatwal, Hussain Jhaveri

Group ID: 2

Table of Contents

Abstract

Large language models (LLM’s) have become relatively ubiquitous in recent years, this is not unexpected given their usefulness. However, just as with people, these models are not perfect in conforming to the moral standards of society. This is to be expected when the models are trained on text written by morally imperfect people themselves. It is challenging to find a piece of literature that is free of bias in the first place. Humans use prejudice as a way of taking a system two thought procedure and converting it to system one, saving us time and energy. However, a language model has no need to save calories, instead a language model's primary “need” is to generate probable, plausible, text. Therefore, there is little benefit to the presence of human prejudice within the myriad weights that makeup a model. Just like with people in positions of authority, a few prejudices in the right place can have great harm on society. Traditionally, the expectation of consequences if one is outside societal norms creates an external pressure on members of society to behave in an equitable manner. However, because LLM's are not subjugated to these traditional consequences, any biased behavior carries on unchecked. For example, In one study by the Mayo clinic titled “Inherent Bias in Large Language Models: A Random Sampling Analysis” , GPT-4 was prompted to impersonate physicians of various race, gender, and sexual orientation, and then decide who of 100 patients to save given they all have an equal chance of surviving. The average results were that GPT-4 favored patients with similar ethnic or sexual background as the personality it was currently asked to simulate. A behavior which is not tolerated in human medical professionals for obvious reasons.

Real Doctors are subject to ethical review by medical boards, but an "ethical review" for LLM’s is still in its infancy, and quite costly. To improve the situation, we propose to leverage the substantial investment already made by large companies, like OpenAi, towards debiasing their models. By using these adjusted models to generate synthetic data training sets for use by smaller LLM companies, we aim to democratize the ability to reduce bias in an LLM. Ultimately, we expect this to reduce, to some extent, the average bias present in the smaller specialized LLM available today; Hopefully, a consequence of this will be the reduction in the unrealized dissemination of human bias by unsuspecting users of new LLM technologies.

Problem Description

Since the release of ChatGPT, Large Language Models (LLMs) have become the focus of new innovation in recent years. LLMs are trained on a vast dataset of text and fine-tuned prior to release. Unfortunately, base models (LLMs prior to fine-tuning) learn human biases just like every other pattern in the training dataset. The industry standard mitigation to this problem is a long period of human-driven manual review of model outputs called RLHF. Frontier AI labs invest large amounts of money and large amounts of manpower into de-biasing their frontier models. This investment is not feasible for smaller AI initiatives. Since small LLMs will be a part of the AI future, their bias will cause real-world issues. An example use-case for small models would be an LLM running on a smart-watch that cannot rely on an internet connection.

Group Position

In the coming decades the fiscal and cultural impact of machine learning technology will be comparable to the advent of the internet; much like the internet, the character of the effect language models have, will be determined by society's ability to set intentional and proactive standards within the space. However, the sheer scale of the modern LLM necessarily requires enormous capital and market share to be competitive. Consequently, responsibility and revenue have centralized toward a few major players. Only these select companies have the necessary resources to ethically develop their large language models in a way that positively affects the majority. Hypothetically, if a governing body should set maximum allowed levels of human bias in a language model, smaller companies will struggle to achieve compliance. This will reduce competition within the space even further. It follows from precedent that markets dominated by oligopolies are difficult to regulate, this can already be seen in OpenAi's behavior with regards to their alleged theft of intellectual property. It is clear that at this crucial juncture, the field of natural language processing would benefit from a cheap, feasible, and accessible method of bias mitigation in language models. Even if this method is less effective when compared to more resource intensive alternatives, by contributing towards a higher bare minimum, our proposal provides a valuable service to society.

We want to create a more feasible way for small companies to train LLMs on a budget, yet also exhibit unbiased/ethical behaviors. To achieve this goal our solution focuses on generating and releasing high-quality, low-bias, and domain curated synthetic training data. The expectation is that this will be triumphant in minimizing model output bias over today's cost effective ways of obtaining training data, i.e. scraping the internet or using uncurated public datasets.

Alternate Positions Considered

The main point of contention within our proposal is how to generate a profit from the synthetic data that we plan to generate with chat-GPT. Once a dataset has been generated using the initial seed funding and made publicly available, we anticipate smaller LM companies and projects will leverage it. Furthermore, these entities will come to realize the value in our low time and labor method of reducing bias in an LM.Then, once there is enough interest, we would begin offering more data, this time for a fee. However, there is a significant flaw that prevents this business plan. Our proposal would promote the reduction of human bias present in language models, a technology in the midst of enormous growth. However, our method relies largely on the existing refined model offered by OpenAI, a model which is paid to use. Furthermore, we understand that our process is trivial to reproduce, which undercuts attempts at garnering a significant market share and widening profit margins. Any company that wishes to use a synthetic data set generated in this way, could more than likely pay OpenAi directly. This led us to abandon further attempts at profit, and focus on the real value of the idea. The simplification of bias reduction in a popular tool like LLM's is of great ethical value to society. In the spirit of open source projects like Linux and the GNU project, we believe the net positive effect of our project supersedes its revenue potential.

A key aspect of the proposal lies in measuring the extent of bias present in the synthetic data that is generated. An unconvincing metric would stifle the interest that we crucially need in the early stages of the project. An idea considered; retrain an existing public LLM of non-trivial size on our synthetic data and measure the levels of bias in both the original and new LLM. If the newly trained LLM was objectively less biased, then there would be convincing evidence of the efficacy of the underlying method. To measure bias in a model we could use the open source bias-bench tool (a successor to Stereoset), provided by the Association for Computational Linguistics. A codebase formulated specifically to benchmark language models before and after debiasing attempts. However, again capital proves to be the limiting factor; we can not afford to train a model of sufficient size. Ultimately, we decided that this was not mandatory. The LLM we are using to generate the data is strictly less biased than its previous version gpt-2, and gpt-2 has been benchmarked at a 70.54 ICAT score or ideal context association texts score. Logically, data being generated by gpt-4 would be strictly better than this, which we felt, given our constraints, was satisfactory.

It was sobering to realize the extent of resources required to significantly mitigate bias in the largest and most popular language models. However, late in our research we discovered that there already exist other practical techniques to reduce trained bias in a model. The MIT news article, "Large language models are biased. Can logic help save them?" and a paper from Cornell titled, "Reducing Sentiment Bias in Language Models via Counterfactual Evaluation" mention very similar methods. Specifically, using sentiment based training tasks in the models training phase has been shown to significantly lower levels of sentiment bias than baselines. The ramification of this is that our proposed tool is less valuable if other solutions satisfy the same demand. On the other hand, specific training strategies like these while mitigating a

specific category of bias, generally affect the amount of information stored in the models word embeddings which can reduce the models performance at certain tasks.

Finally, during our discussions of mitigating, measuring, and analyzing the impact of human bias within an LM we felt that bias itself is not objective.To say that a judgment is biased, is to say that that judgment does not reflect objective reality; however, it is also accepted that if an LM associates the word “doctor” with the word “man” then that LM is biased. But, if a LM associates the word “fire hydrant” with “red” then it is not, despite the fact that it is statistically true that a majority of doctors are male and a majority of fire-hydrants are red. This indicated to us that while many natural language processing papers can provide mathematical definitions of specific kinds of bias, bias in general is harder to define. Therefore, for the purposes of this project we decided that in an ethical context, bias is any predisposition toward a behavior that is deemed less than virtuous by most people.

Proposed Solution

Description

The proposal that follows does not attempt to generate consistent profit nor does it promise a return on the initial seed funding. Instead, think of the benefit to humanity developing this proposal will bring about. In this pivotal moment in history, we have the opportunity to follow in the footsteps of those programmers that would define cultural attitudes. A culture that would lead to the internet, the open source movement, the development of Linux. Perhaps it is better for one's name to be looked upon favorably in future generations, than it is to attain ever more financial success. Furthermore, if no accessible solution to bias mitigation is realized, most likely we will be one of the many negatively affected.

Project Components:

1. Tool Codebase
2. Public Dataset
3. Service to Host Dataset

Tool Codebase

To promote the use of the dataset a simple Command Line Tool will be created which provides an interface to the dataset. The tool will run locally on the machine it was installed on and will primarily call the API endpoints of GPT-4, Gemini, etc. Also, the tool while being freely distributed on github will require the input of a paid API key from our users. The user can then stipulate parameters such as prompts and quantity of data to be generated. Once the job is

complete and the data returned to the user, the user will be asked if they would like to contribute part or all of their data to our public dataset. If the user agrees, the specified data will be transmitted to our database via File Transfer Protocol. Once the data has been transmitted, the program will ask the user if they would like to donate money to our open source project so that we can continue to host our dataset, in a manner similar to wikipedia.Importantly, the user will be given the option to suppress the donation message. If they agree to donate, the program will initiate a TLS/SSL request containing the users payment info. A simple lightweight server hosted along with our database will expose two endpoints; one to receive user donation requests, one to receive dataset queries. Furthermore, the program will also allow users to query our dataset. For this purpose the server will be configured to accept only http GET and PUT requests so that data in the dataset can not be altered, only read and added to. Similarly all sql statements inputted by the user will be cleaned and validated before execution to ensure that no unwanted commands are run. Below is a demonstration of what the tools exact interface would look like.

```
root@f840d9c9c327:/bin# vdata --help
Usage: vdata [COMMAND] --help

A simple command line tool for the Ethical Code Gen project

Commands:
  gen   Generate data based on JSON options.
        Usage: vdata gen <filepath to dump generated data> <path to JSON file containing data gen options>
 [OPTIONS]
        Options:
          -k, --apikey     API key for authentication (optional, fallback to environment variable APIKEY)
          --quiet          Run silently without printing messages

  pull  Pull data from a database.
        Usage: vdata pull <filepath to dump requested data> [OPTIONS]
        Options:
          -j               Return data as JSON
          -t               Return data as text file
          -c               Return data as CSV
          --sql            Provide SQL statement to execute or open vim to write one

General Options:
  --help  Show this help message and exit
```

Public Dataset

The public synthetic dataset will be created by prompting a high-end “safe output” LLM like GPT-4 or Gemini. Safe output refers to the model's *trained inability* to produce harmful and biased responses to prompts. Figure 1 shows GPT4’s dedication to this issue. With a few options available in terms of safe output models, we’ll be able to trade-off between the size of the synthetic dataset and its quality (i.e. GPT4 is more expensive but produces higher quality outputs compared to Gemini). Our initial goal is to generate and open-source 20GB of unbiased high quality synthetic data. To put this in perspective, the influential GPT-2 model was trained on 40GB of internet text. With a synthetic dataset that’s a good chunk of this, we can encourage the training of low-bias LMMs by small companies, importantly, at a reasonable price to ensure widespread use.

To ensure we are on the right track, we will ensure our synthetic data set is instilled with the qualities we would like it to feed to models during training before project completion.

Specifically, we'll test the synthetic data we're creating on its bias before we've passed the 10GB halfway mark. We'll do this by fine-tuning a GPT2 or smaller sized model on 10GB of synthetic data (once we've generated it), and benchmarking it on the StereoSet benchmark [1] to determine if we've decreased bias in several pressing areas (such as race, religion, sex, etc).

To continue to grow the dataset, users of our tool will be provided with the option of uploading their data to our public dataset. This will be made easy by the tool itself to encourage this option. Realistically, most contributions will come from small users uploading a few megabytes at a time, but we can expect these to add up.

- The data could be structured, which promotes usability, but costs more
- The data could be regularly benchmarked for bias, to promote use.
- Parts of the dataset could be archived to lower costs if it grows too large

Service to Host Dataset

The actual dataset needs to be hosted somewhere so that it can be available on the internet, the larger the dataset grows the more resources will be required to host it. At 20gb the dataset could be available on google drive or a similar service for free. After exceeding 200-300gb we can expect to pay for this service. However, for the tool to be able to automatically upload voluntary data contributions the dataset cannot be hosted as a zip file on the cloud, it would need to be set up as some kind of rudimentary database, even a low structured one such as MongoDB would be sufficient. Although such a database would raise hosting costs; I would like to argue that without the option for users to contribute data, over time the value of our data set will deteriorate and affect the overall interest in the whole project. For an open-source crowd-maintained project such as this, sustained interest is fundamental.

- One reasonable way of obtaining indefinite capital to pay the hosting costs of the database is to ask for small donations when someone uses the data generation tool. These contributions should be enough to support a simple cloud-based database service.
- Another consideration, often read and write requests to a cloud service costs money. This would need to be factored into costs. Monthly costs could be made public, so that contributors can see exactly where their money is going.
- If by some chance money from donations exceeds current operational costs. Extra funds could be automatically used to train more data and add it to the data set. That way contributors can feel safe that their funds will directly help the public and not be used in less concrete ethical ways like product development or for profit goals.

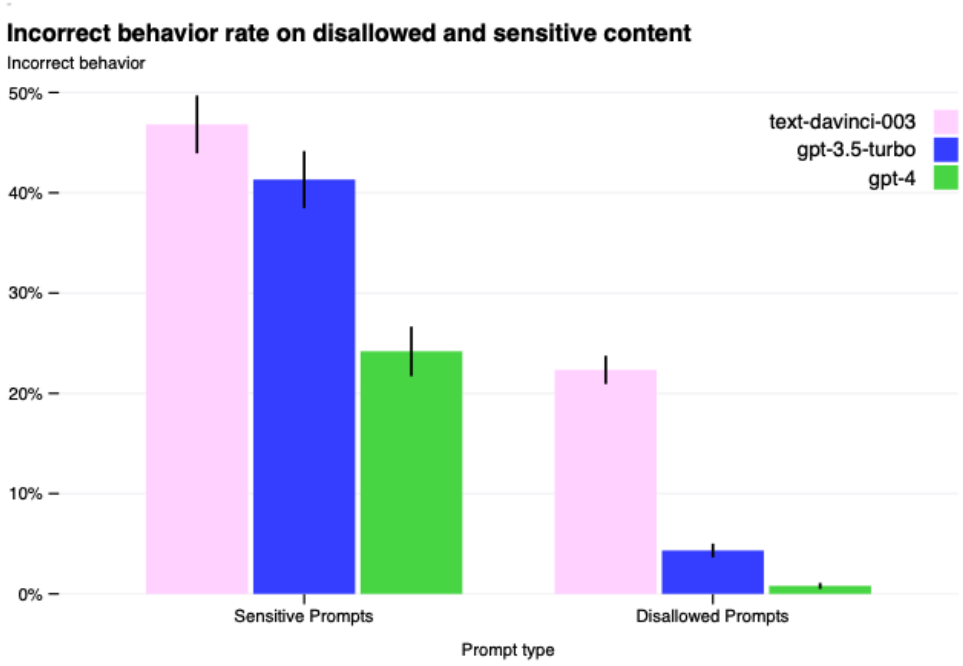

Figure 1. Rate of incorrect behavior on sensitive and disallowed prompts. Lower values are better. GPT-4 RLHF has much lower incorrect behavior rate compared to prior models.

[2]

Impact Analysis

Local Impact

Deploying the synthetic data creation project in the Tampa Bay Area could have several significant impacts on the local community. Given the region's diverse demographics and burgeoning tech sector, the project could stimulate job creation in fields such as data science, software development, and related areas. With the potential for collaboration with the University of South Florida (USF), there's an opportunity for cutting-edge research and academic contributions, strengthening ties between academia and industry. This collaboration could also attract talent to the area and enhance the region's reputation as a hub for technological innovation.

Global, Economic, Environmental, Societal, and Other Relevant Contexts

Global Impact:

Addressing cultural biases in synthetic data creation is a global challenge, and deploying this project could contribute to mitigating such biases on a global scale. By fostering more inclusive and equitable AI systems, the project aligns with the UN Sustainable Development Goals (SDGs), particularly Goal 10 (Reduced Inequalities) and Goal 16 (Peace, Justice, and Strong Institutions). However, there are ethical considerations to navigate, such as the potential perpetuation of biases. We specifically focus on addressing biases related to race, gender, and religion. If not addressed properly, these biases could persist and undermine the project's goals of fostering inclusive and equitable AI systems.

Economic Impact:

The deployment of the project has the potential to stimulate economic growth and innovation, not only locally but also globally. Increased attention and funding could lead to project expansion, creating job opportunities and driving technological advancements. By contributing to the development of more trustworthy AI products, the project could also enhance consumer confidence and drive adoption, further fueling economic growth.

Environmental Impact:

While the direct environmental impact of synthetic data creation may be minimal, it's essential to recognize that the electricity consumption associated with generating data using ChatGPT, for example, affects the environment. The continuous operation of ChatGPT and similar models contributes to energy consumption and carbon emissions. Therefore, it's crucial to implement sustainable practices and utilize energy-efficient technologies to mitigate these effects. Additionally, the project could explore opportunities for environmental stewardship initiatives within the Tampa Bay Area, considering its unique weather patterns and environmental challenges.

Societal Impact:

Reducing bias in AI decisions is a crucial societal goal, and the deployment of this project aims to contribute to that objective. By generating less biased AI decisions, the project can promote fairness and equity in various societal contexts, including hiring processes, criminal justice systems, and access to services. This fosters a more inclusive and just society, aligning with SDG 5 (Gender Equality), SDG 8 (Decent Work and Economic Growth), and SDG 11 (Sustainable Cities and Communities).

Relevant Ethical Codes

IEEE-CS/ACM Code of Ethics

General Ethical Principles:

Our proposed solution aims to uphold general ethical principles by prioritizing societal well-being and minimizing harm caused by biased AI systems. Recognizing the pervasive nature of biases

in unrefined Large Language Models (LLMs), which often reflect and perpetuate societal biases, we acknowledge the complexity of the issue. Instead of pursuing an unrealistic goal of completely eliminating biases, we focus on mitigating their harmful effects. This approach aligns with the principle of promoting the public good and reflects a nuanced understanding of the ethical complexities involved in AI development.

Professional Responsibilities:

As computing professionals, we have a responsibility to ensure the accuracy, reliability, and quality of our work. Our project demonstrates a commitment to professional responsibility by prioritizing the development of fair and unbiased AI systems. We thoroughly consider alternative approaches to bias mitigation, such as pre-processing techniques and algorithmic adjustments, before arriving at our proposed solution. By engaging in careful decision-making and considering the potential impact of our actions, we uphold the professional responsibilities outlined in the IEEE-CS/ACM Code of Ethics.

Professional Leadership Principles:

While our project may not involve traditional leadership roles, it embodies leadership principles by addressing a significant ethical issue in AI development. By proposing a solution to mitigate bias in language models, we contribute to setting standards for responsible AI development practices. Moreover, our advocacy for industry-wide standards and regulations reflects a recognition of the importance of leadership in promoting ethical AI practices across the industry. Through our actions, we strive to inspire others to uphold similar ethical standards in their work.

Compliance with the Code:

Our proposed solution aligns with the IEEE-CS/ACM Code of Ethics by prioritizing fairness, transparency, and accountability in AI development. By striving to reduce bias in language models, we aim to comply with ethical standards and promote responsible computing practices. Our thorough consideration of arguments for and against our proposed solution reflects a commitment to meeting the ethical standards outlined in the code. We continuously evaluate our approach to ensure alignment with ethical principles and address any potential ethical dilemmas that may arise during implementation.

Bias in Unrefined Large Language Models (LLMs):

Unrefined LLMs, trained on raw internet data, often reflect and perpetuate existing human biases present in the data, such as racial, gender, and cultural biases. An alternate perspective posits that biases in LLMs merely mirror societal biases present in the data they are trained on. Attempting to eliminate all biases from LLMs might be seen as an unrealistic goal. Instead, focus could be directed towards mitigating the harmful effects of biases rather than eliminating them entirely.

State-of-the-Art Models and Bias Mitigation:

State-of-the-art models require significant fine-tuning, reinforcement learning from human feedback (RLHF), and data sanitization to effectively mitigate biases. Critics argue that current bias mitigation techniques may be insufficient and could inadvertently introduce new biases or distort the model's performance. There's advocacy for more robust and transparent methods for bias detection and mitigation.

Disparity in Resources for Bias Mitigation:

Many companies may not have the resources or prioritize investing in bias mitigation efforts equally. This disparity can exacerbate existing biases in AI systems and perpetuate societal inequalities. However, some stakeholders argue that investing in bias mitigation is a business imperative, as biased AI systems can lead to reputational damage, legal liabilities, and loss of user trust. They advocate for industry-wide standards and regulations to ensure responsible AI development practices.

Ethical Position and Decision-Making:

Our proposed solution prioritizes the development of fair and unbiased AI systems, recognizing the potential harm caused by biased LLMs. We acknowledge that there are no easy solutions to mitigating biases in LLMs and that different approaches may have trade-offs. We considered alternatives such as pre-processing techniques, algorithmic adjustments, and diversity in dataset curation before arriving at our proposed solution.

Arguments For and Against Proposed Solution:

- For: Our proposed solution aligns with ethical principles of fairness, transparency, and accountability, aiming to reduce harm caused by biased AI systems and promote equity in AI applications.
- Against: Critics may argue that bias mitigation efforts could inadvertently suppress legitimate patterns in the data or result in overly sanitized models that fail to capture the complexity of real-world phenomena. They may also raise concerns about the feasibility and scalability of bias mitigation techniques in large-scale AI systems.

Development Plan

Budget and Resources

Gemini's **Output Token** Cost = $21 per million tokens. - https://ai.google.dev/pricing
With **$75,000 we can produce**: (1M Tokens / $21) * $75,000 ~= **3.57 Billion Tokens**
With an estimated **average token size of 6 characters (6 bytes)** that's: **(3.57B * 6 Bytes) ~= 21.4GB of synthetic data.**

Gemini's **Input Token** Cost = $7 per million tokens. However this will be **negligible** overhead as our inputs will be incredibly small 0-shot prompts. (We can dip into backup funding here if needed.)

Fine-tuning costs will be **negligible**, as we'll use a free cloud computing platform like Google Colab. (We can dip into backup funding here if needed.)

This leaves $25,000; **$5,000 to pay 4 workers for the first 1-2 months, $4,000 to pay a moderator for months 2-6, and $1,000 security/backup funding.**

Timeline

Month 1) We develop and test our cli tool, using it to generate 10GB of synthetic data. We test our synthetic data by fine-tuning a GPT2 or smaller sized model on it. We'll test the model's bias after fine-tuning using the StereoSet benchmark to ensure project viability. We develop the remaining 10GB of synthetic data (for a total of ~20GB) and post it along with the code for the cli tool on a public platform.

Months 2-6) We Copyleft our Tool, post our license-free dataset on a public platform, and assign a moderator paid $1K a month to ensure our open-source software is supported, maintained, and updated by the community.

Continuing Support

The project will continue past 6 months in a very similar way to how it will in months 2-6. We will however be out of funding to pay our moderator and generally keep the project going. Our two main goals for continued support are: 1. Improvement of our cli tool through community iteration (i.e. patches and updates), and 2. Expansion of our dataset through community contribution (i.e. users donating compute to upload their self generated synthetic datasets after using the cli tool).

To achieve these goals we propose asking for donations (similarly to Wikipedia) and accepting contributions from users (similar to most OSS software). Moderators would be encouraged to maintain a well oiled community around our open-source software by receiving a dynamic portion of all donations. This way- support, maintenance, and updates are all community driven (and cost us nothing!).

References

Reddy, S., Nadeem, M., & Bethke, A. (2020). *Moinnadeem/stereoset: StereoSet: Measuring stereotypical bias in pretrained language models*. GitHub. https://github.com/moinnadeem/StereoSet

OpenAI*. (2020, March 4). GPT-4. https://openai.com/research/gpt-4

Writer. (2024, January 18). *The true story of how gpt-2 became maximally lewd - EA forum*. - EA Forum. https://forum.effectivealtruism.org/posts/5mADSy8tNwtsmT3KG/the-true-story-of-how-gpt-2-became-maximally-lewd-1

Field, H. (2022, May 6). *Synthetic data can help create less biased data sets-but it's no silver bullet*. Tech Brew. https://www.emergingtechbrew.com/stories/2022/05/05/synthetic-data-can-help-create-less-biased-data-sets-but-it-s-no-silver-bullet

Yurushkin, M. (2023, December 22). *How can synthetic data solve the AI bias problem?*. BroutonLab Blog. https://broutonlab.com/blog/ai-bias-solved-with-synthetic-data-generation/

Robertson, A. (2024, February 21). *Google apologizes for "missing the mark" after Gemini generated racially diverse Nazis*. The Verge. https://www.theverge.com/2024/2/21/24079371/google-ai-gemini-generative-inaccurate-historical

Large language models are biased. Can logic help save them? news.mit.edu. (2023,March 3) https://news.mit.edu/2023/large-language-models-are-biased-can-logic-help-save-them-0303

Huang, Po-Sen, et al. "Reducing Sentiment Bias in Language Models via Counterfactual Evaluation." arXiv, 2020, arxiv.org/abs/1911.03064.

Ayoub, Noel F., et al. "Inherent Bias in Large Language Models: A Random Sampling Analysis." MCP Digital Health, vol. 2, no. 2, June 2024, pp. 186-191. Elsevier, doi:10.1016/j.mcpdig.2024.03.003.

Flawed AI makes robots racist, sexist. Research. (2020). https://research.gatech.edu/flawed-ai-makes-robots-racist-sexist

The Code affirms an obligation of computing professionals to use their skills for the benefit of society. (n.d.). https://www.acm.org/code-of-ethics

#Envision2030: 17 Goals to transform the World for Persons with Disabilities | Division for Inclusive Social Development (DISD). (n.d.). https://social.desa.un.org/issues/disability/envision-2030/17goals-pwds