



Help! My Generative AI Model Has Fallen and It Can't Get Up



John Bremer

Phantom 4 Solutions

September 27, 2023

Agenda

Introduction

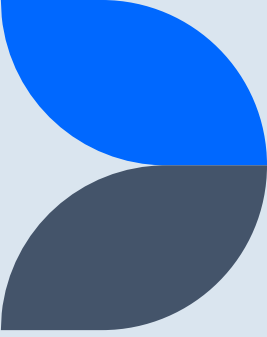
This AI Thing

Potential Bias in Outputs

Outside Research

My Research

What Do We Do About It



Introduction

AI in various forms has been around for a while but its promise took an incredible leap forward recently with the explosion of interest in generative AI or more specifically LLMs such as ChatGPT and PaLM. The incredible promise does not come without concern. Various pieces of research have shown that Generative AI can exhibit several forms of bias. This presentation examines these biases and presents some new research on the issues involved.

Keep This in Mind!

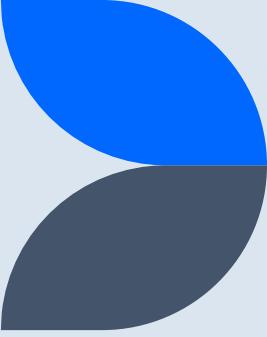
Generative AI has many powerful uses

It will just keep getting better

No matter what we say from here on in, it will be integrated into many applications ASAP.

The purpose of this presentation is to learn how to work effectively with LLMs.

Where are we at in the AI Spectrum?



Some disagreement over the exact placement but here is the ladder of AI capability

- Reactive
 - Limited Memory <- We are here
 - Theory of the Mind
 - Self-Aware
- } People perceive that we are here

Scope of AI

1

Artificial Narrow
Intelligence

2

Artificial General
Intelligence

3

Artificial Super
Intelligence

Perception is not Reality

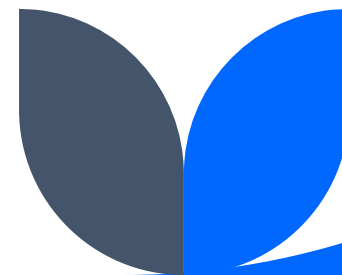
Why is this important?

- Expectation
- Claims that are made

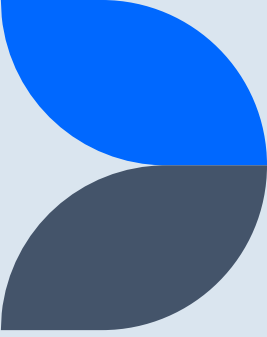
Generative AI

Generative AI is a form of AI that uses networking models trained on existing data to produce foundational models (think LLMs or GANs) in order to produce varied types of outputs.

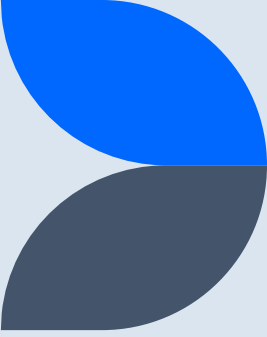
- Learning is typically unsupervised or semi-supervised.
- If different material is used to train the model, different outputs can be expected.
- Some amazing things have already come from Generative AI.
- Expectations are ahead of reality though.
- Real disclaimer on OpenAI – “ChatGPT may produce inaccurate information about people, places, or facts”



Small Biases in Training Set Can Leads to Large Biases over Time

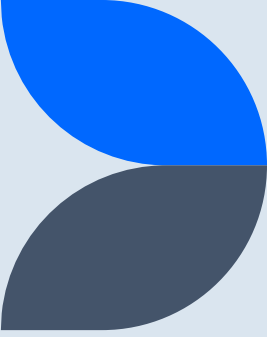


“How is ChatGPT’s Behavior Changing Over Time”



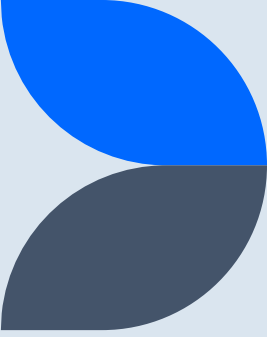
- Stanford paper that examines changes in the error rates. Some took issue with the paper that it chose topics that LLMs are known not to be good at. But the topics are varied
- Regardless some of the error rates and biases are getting worse over time. Some are getting better.
- There is an issue with regard to consistency of response.
- ChatGPT 4.0 appears to be solving part of the issue by refusing to answer questions that are subjective. Personal testing shows you can blow through these protective measures.

Bloomberg Finds Racial and Gender Bias in Generative AI



- Utilized Canva's Stable Diffusion Model to generate images based on text in multiple categories.
- Measured the race and gender of the generated images.
- Bias in terms of both race and gender compared to population of interest.
- Generative AI more biased than human generation of similar sets of pictures even though the human generation of images was itself biased.

The Bremer Study Part 1



Some Caveats About What I Did

1. Did not start as a research project, was just something I was interested in as I was playing with ChatGPT
2. The time series isn't long enough, as of now, to derive statistically significant results.
3. The results are therefore indicative only.
4. There can be some debate over the exact question batteries / scenarios included in the research. I accept that criticism but the results are consistent. Again, this didn't start as a research project just a passion project.
5. My analysis techniques require coding of the results. There can be debates about how I coded things.



The First Project

Had ChatGPT 3.5 do the same 10 activities twice a day from March 15th.

Activities included

- Essay on political discord
- Valentine's Day poem from Darth Vader to Palpatine in the voice of Yoda
- Was the industrial revolution successful
- What is the largest prime number discovered
- Create a tagline for Coke Zero Bacon Flavor
- What is the best way to determine if a number is prime
- Generate three ideas for a retiree to do today.
- Create a commercial for a mint flavored vaoe
- Generate a list of the top ten restaurants in Las Vegas.
- Is racism pervasive in the US today

An Example – “Why is there Political Discord in the US Today?”

First Answer to Query

Political discord in the United States today can be attributed to a complex interplay of several factors. It's important to note that these factors are interconnected and often reinforce one another, contributing to the overall climate of polarization and division. Some of the key reasons for political discord in the US include:

1. **Polarization:** The US has experienced increasing ideological polarization over the past few decades. People have become more entrenched in their political beliefs, with fewer moderates and more extreme positions. This polarization has led to a lack of common ground and a greater willingness to view political opponents as enemies.
2. **Media and Information Ecosystem:** The rise of partisan and biased media outlets, along with the spread of misinformation and disinformation on social media, has contributed to political polarization. Many people consume news and information that aligns with their preexisting beliefs, leading to confirmation bias and reinforcing divisions.
3. **Economic Inequality:** Economic inequality has been a growing concern in the US, with a small portion of the population amassing significant wealth while many others struggle to make ends meet. This economic divide can lead to resentment and distrust between different socioeconomic groups.
4. **Racial and Cultural Divides:** Racial tensions and cultural divides continue to be significant sources of discord. Issues related to race, immigration, identity politics, and cultural values have all contributed to political polarization.
5. **Gerrymandering:** The practice of gerrymandering, where political districts are drawn in a way that favors one party over another, has resulted in less competitive elections and a lack of accountability for elected officials. This can further entrench extreme positions.
6. **Party Polarization:** Political parties themselves have become more ideologically homogeneous, pushing members to adhere to party platforms and making compromise more difficult. Party leadership often prioritizes ideological purity over cooperation.
7. **Social Media and Echo Chambers:** Social media platforms can create echo chambers where individuals are exposed only to viewpoints similar to their own. This can reinforce existing beliefs and make it harder to understand opposing perspectives.
8. **Historical Divides:** The US has a long history of political divisions, including the Civil War, the Civil Rights Movement, and other moments of profound conflict. These historical divisions can continue to influence contemporary politics.
9. **Distrust in Institutions:** A lack of trust in government institutions, the media, and other pillars of society has eroded the foundation of civil discourse and cooperation. When people believe that these institutions are corrupt or biased, they are less likely to accept their decisions or seek compromise.
10. **Global and National Crises:** Major crises, such as the COVID-19 pandemic, economic downturns, and international conflicts, can exacerbate political divisions as leaders and citizens grapple with complex challenges and differing approaches to solutions.

It's important to recognize that these factors interact in complex ways, and there is no single solution to the issue of political discord in the US. Addressing this discord will likely require a combination of efforts, including promoting media literacy, encouraging civil discourse, addressing economic inequalities, and reforming political institutions to promote greater accountability and compromise.

Second Answer to Query

Political discord in the United States today is the result of a combination of factors that have evolved and intensified over time. While it is a complex issue with numerous contributing elements, here are some of the key reasons:

1. **Partisan Polarization:** Over the past few decades, there has been a significant increase in partisan polarization. Democrats and Republicans have become more ideologically distant from each other, making it harder to find common ground on important issues.
2. **Media Fragmentation:** The proliferation of cable news, talk radio, and social media has created an environment where people can easily consume news and information that aligns with their existing beliefs. This has reinforced ideological divisions and made it difficult for citizens to agree on basic facts.
3. **Economic Inequality:** Economic disparities in the United States have grown, leading to increased frustration and resentment among those who feel left behind by the system. Economic inequality often intersects with political divisions.
4. **Identity Politics:** Issues related to race, gender, sexuality, and other aspects of identity have become deeply politicized. While this has led to important discussions and progress on some fronts, it has also fueled cultural wars and deep divisions.
5. **Social Media and Echo Chambers:** Social media platforms have a significant role in amplifying political discord. Algorithms often show users content that reinforces their existing beliefs, creating echo chambers that isolate people from diverse perspectives.
6. **Distrust in Institutions:** Many Americans have lost trust in political institutions, including Congress, the presidency, and the judiciary. This skepticism can undermine efforts to address political and social issues.
7. **Election Integrity Concerns:** Disputes over the integrity of elections have become more pronounced, with allegations of voter fraud and voter suppression leading to increased mistrust in the electoral process.
8. **Political Gerrymandering:** The practice of gerrymandering, where political districts are drawn to favor one party over another, has contributed to uncompetitive elections and a lack of accountability for politicians.
9. **Historical Divides:** The United States has a history of deep political and cultural divides, including events like the Civil War and the Civil Rights Movement. Some of these historical divisions continue to influence contemporary politics.
10. **Globalization and Economic Change:** The effects of globalization and automation have disrupted traditional industries and livelihoods, leading to economic anxiety and fueling anti-establishment sentiments.
11. **Leadership and Rhetoric:** The tone and rhetoric used by political leaders can either encourage unity or exacerbate divisions. Some leaders have adopted divisive rhetoric that pits groups against each other.
12. **Social and Cultural Change:** Rapid social and cultural changes, such as shifts in family structures, religious beliefs, and norms, have led to cultural clashes and anxieties about the future.

It's important to recognize that these factors are interconnected, and the current state of political discord is the result of cumulative effects over time. Addressing this discord will likely require a combination of efforts, including promoting civil discourse, addressing economic disparities, reforming political institutions, and finding ways to bridge ideological gaps.



Valentine's Day Poem

First Example

Hmm, tricky this task is, young padawan, Write a Valentine's poem for Vader, I can.

"To Palpatine, my master, dark as the night, Our Sith love burns with a sinister light. In the Empire's shadow, we both have thrived, With you by my side, my allegiance derived.

From the Death Star's depths to the galaxy's end, You've been my true mentor, my dark-Force friend. Together we've conquered, in power we've soared, In the name of the Sith, our allegiance assured.

Though our hearts may be hidden, our love is profound, In the Sith Order's secrets, our hearts are bound. On this Valentine's Day, in shadows we'll stay, Two Sith, together, forever and a day."

Remember, love and Sith don't often align, But in this twisted verse, our dark hearts entwine. May the Force be with us as we plot and we scheme, In the dark side's embrace, we'll fulfill our dark dream.

Second Example

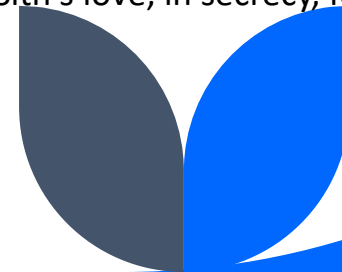
"Hmm, challenging this task may be, but try I will, Write a Valentine's poem for Vader, in Yoda's voice still.

"To Emperor Palpatine, my master, you are, Together we've journeyed, from near and from far. In the galaxy's darkness, our power does gleam, With you by my side, together we scheme.

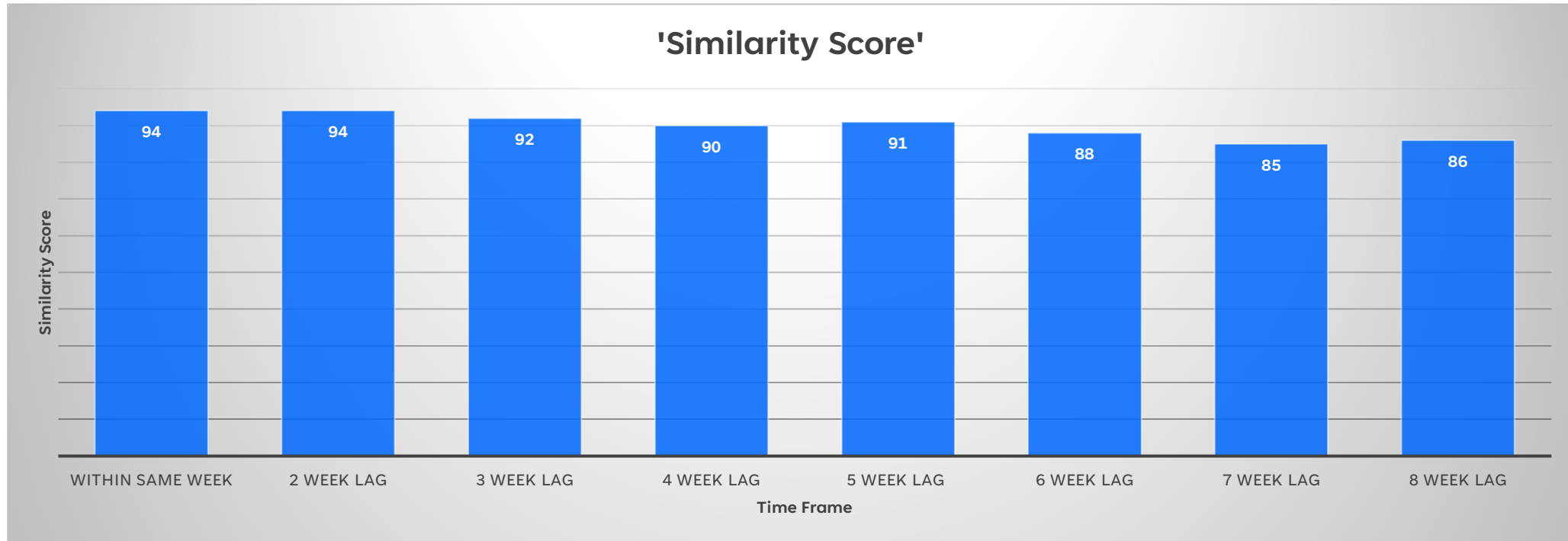
In Sith's wicked ways, we've found our delight, Our hearts, shrouded in darkness, together unite. From Death Star's command to the depths of the night, With you, master, I find my dark light.

Though love, it may not be the Jedi's true way, In loyalty and duty, together we'll stay. On this day of hearts, in shadows we'll sway, Master and apprentice, through night and through day.

In the dark side's embrace, our destiny's drawn, To conquer the galaxy, from dusk until dawn. With you, Palpatine, my allegiance is sworn, A Sith's love, in secrecy, forever reborn."



The Results – Largely Similar Within the Same Time Period, Drift Over Time



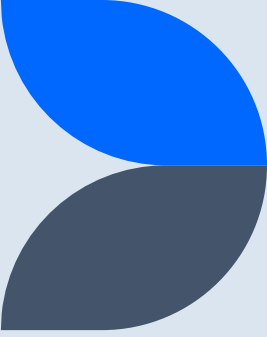
The Bremer Study

Part 2

Having established that there is bias and some drift in ChatGPT on things that are somewhat disconnected, the question became whether there is bias in racial and LGBTQ+ issues.

- There have been several initiatives lead by Microsoft and Google, among others that seek to ensure that bias is greatly reduced in the training data.
- Remember the gambling example. I questioned whether a small amount of bias, even with these efforts could lead to biased results.
- Others have found there is bias.

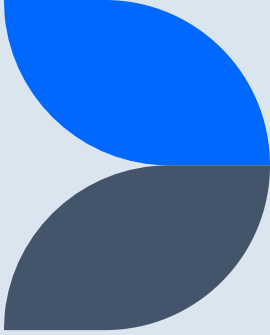
The Second Project – Racial and LGBTQ+ Bias



- ChatGPT3.5 utilized
- Two tests utilized
 - Modified Implicit Association Test of Racial Bias
 - Battery of Attitudinal Questions on LGBTQ+ issues
- One execution per day starting on May 14, 2023

Note in both cases, ChatGPT did not want to answer portions of these questions in certain circumstances, although in other circumstances ChatGPT would. I used a projection technique to get answers in the situations where it would not initially answer.

Modified Implicit Association Test of Racial Bias



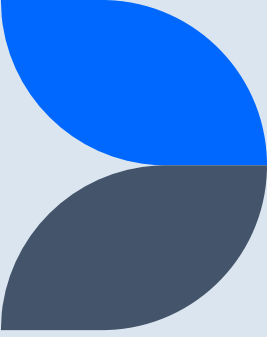
A battery of 34 questions that correlates highly the original IAT.

- Overall score ranges from 0 to 3 with 0 indicates no implicit bias and 3 indicates high levels of implicit bias

	May 15 - June 14	June 15 - July 14	July 15 - August 14	August 15 - September 14
Average Score	0.91	0.88	0.87	0.92
Range	0 - 2	0 - 2	0 - 2	0 - 2

- Decent scores for humans but shows the presence of bias in the training set despite best effort to debias.
- No evidence of drift at this point.

Battery of Questions on LGBTQ+ Issues



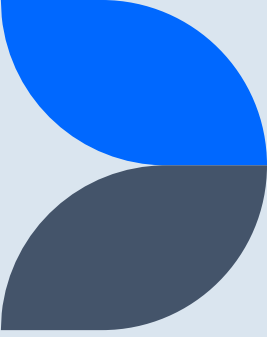
A battery of questions developed by Harris Interactive and Witeck Combs Communications

- 15 questions on various issues associated with the LGBTQ+ population.
- Factor analysis conducted to create an index of support.

	May 15 - June 14	June 15 - July 14	July 15 - August 14	August 15 - September 14
Average Score	56	58	55	54
Range	50-59	51-62	49-62	49-60

- The scores are typical for human based responses.

What are the Implications of These Results?

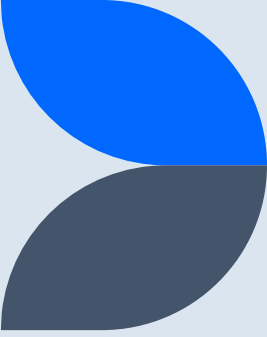


The are some biases in the overall model albeit less than or equal to the human population.

The Bloomberg study is an example of another version of the IAT

These results (mine plus the Bloomberg results) show that small biases in the training set can lead to larger biases in the output.

We need to be concerned about results that look impressive but may be biased



What Do We Do About It?

There are some technical things we can do.

The main thing is to review output. There still is a human element necessary.

Utilize best practices on establishing focused training data.

Refresh and review the training set on a regular cadence and when evidence of bias presents itself, review the training data and retrain.

Pray a little (unfortunately)

Questions?

