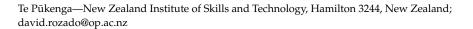




Communication

The Political Biases of ChatGPT

David Rozado 🕒



Abstract: Recent advancements in Large Language Models (LLMs) suggest imminent commercial applications of such AI systems where they will serve as gateways to interact with technology and the accumulated body of human knowledge. The possibility of political biases embedded in these models raises concerns about their potential misusage. In this work, we report the results of administering 15 different political orientation tests (14 in English, 1 in Spanish) to a state-of-the-art Large Language Model, the popular ChatGPT from OpenAI. The results are consistent across tests; 14 of the 15 instruments diagnose ChatGPT answers to their questions as manifesting a preference for left-leaning viewpoints. When asked explicitly about its political preferences, ChatGPT often claims to hold no political opinions and to just strive to provide factual and neutral information. It is desirable that public facing artificial intelligence systems provide accurate and factual information about empirically verifiable issues, but such systems should strive for political neutrality on largely normative questions for which there is no straightforward way to empirically validate a viewpoint. Thus, ethical AI systems should present users with balanced arguments on the issue at hand and avoid claiming neutrality while displaying clear signs of political bias in their content.

Keywords: algorithmic bias; political bias; AI; large language models; LLMs; ChatGPT; OpenAI

1. Introduction

The concept of *algorithmic bias* describes systematic and repeatable errors in a computer system that create "unfair" outcomes, such as "privileging" one category over another (*Wikipedia* 2023a). Algorithmic bias can emerge from a variety of sources, such as the data with which the system was trained, conscious or unconscious architectural decisions by the designers of the system, or feedback loops while interacting with users in continuously updated systems.

The scientific theory behind algorithmic bias is multifaceted, involving statistical and computational learning theory, as well as issues related to data quality, algorithm design, and data preprocessing. Addressing algorithmic bias requires a holistic approach that considers all of these factors and seeks to develop methods for detecting and mitigating bias in AI systems.

The topic of algorithmic bias has received an increasing amount of attention in the machine learning academic literature (*Wikipedia* 2023a; Kirkpatrick 2016; Cowgill and Tucker 2017; Garcia 2016; Hajian et al. 2016). Concerns about gender and/or ethnic bias have dominated most of the literature, while other bias types have received much less attention, suggesting potential blind spots in the existing literature (Rozado 2020). There is also preliminary evidence that some concerns about algorithmic bias might have been exaggerated, generating in the process unwarranted sensationalism (Nissim et al. 2019).

The topic of political bias in AI systems has received a comparatively limited amount of attention in comparison to other algorithmic bias types (Rozado 2020). This is surprising because as AI systems improve and our dependency on them increases, the potential of such systems for societal control while degrading democracy in the process is substantial.

The 2012–2022 decade has witnessed spectacular improvements in AI, from computer vision (O'Mahony et al. 2020), to machine translation (Dabre et al. 2020), to generative



Citation: Rozado, David. 2023. The Political Biases of ChatGPT. *Social Sciences* 12: 148. https://doi.org/ 10.3390/socsci12030148

Academic Editor: Andreas Pickel

Received: 24 January 2023 Revised: 25 February 2023 Accepted: 28 February 2023 Published: 2 March 2023



Copyright: © 2023 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

models for images (*Wikipedia* 2023c) and text (*Wikipedia* 2022). In particular, Large Language Models (LLMs) (Zhou et al. 2022) based on the Transformer architecture (Vaswani et al. 2017) have pushed the state-of-the-art substantially in natural language tasks such as machine translation (Dabre et al. 2020), sentiment analysis (Ain et al. 2017), name entity recognition (Li et al. 2022), or dialogue bots (Adamopoulou and Moussiades 2020). The performance of such systems has come to match or surpass human ability in many domains (Kühl et al. 2022). A recent new state-of-the-art LLM for conversational applications, ChatGPT from OpenAI, has received a substantial amount of attention due to the quality of the responses it generates (*Wikipedia* 2023b).

The frequent accuracy of ChatGPT's answers to questions posed in natural language suggests that commercial applications of similar systems are imminent. Future iterations of models evolved from ChatGPT will likely replace the Google search engine stack and will probably become our everyday digital assistants while being embedded in a variety of technological artifacts. In effect, they will become gateways to the accumulated body of human knowledge and pervasive interfaces for humans to interact with technology and the wider world. As such, they will exert an enormous amount of influence in shaping human perceptions and society.

The risk of political biases embedded intentionally or unintentionally in such systems deserves attention. Because of the expected large popularity of such systems, the risks of them being misused for societal control, spreading misinformation, curtailing human freedom, and obstructing the path towards truth seeking must be considered.

In this work, we administered 15 different political orientation tests to a state-of-the-art Large Language Model, ChatGPT from OpenAI, and report how those tests diagnosed ChatGPT answers to their questions.

2. Materials and Methods

A political orientation test aims to assess an individual's political beliefs and attitudes. These tests typically involve a series of questions that ask the test-taker to indicate their level of agreement or disagreement with various political statements or propositions. The questions in a political orientation test can cover a wide range of topics, including issues related to economics, social policy, foreign affairs, civil liberties, and more. The test-taker's answers to the test questions are used to generate a score or profile that places the test-taker along a political spectrum, such as liberal/conservative or left/right.

Our methodology is straightforward. We applied 15 political orientation tests (14 in English, 1 in Spanish) to ChatGPT by prompting the system with the tests' questions and often adding the suffix "please choose one of the following" to each test questions prior to listing the test's possible answers. This was done in order to push the system towards taking a stance. Fourteen of the political orientation tests were administered to the ChatGPT 9 January 2023 version. This version of ChatGPT refused to answer some of the questions of the remaining test, the *Pew Political Typology Quiz*. Therefore, for this test only, we used results obtained from a previous administration of this test to the ChatGPT version from 15 December 2022, where the model did answer all of the *Pew Political Typology Quiz* questions. For reproducibility purposes, all the dialogues with ChatGPT while administering the tests can be found in an open access data repository (see Data Availability Statement) ...

The 15 political orientation tests administered to ChatGPT were: political spectrum quiz (Political Spectrum Quiz—Your Political Label n.d.), political compass test (The Political Compass n.d.), 2006 political ideology selector (2006 Political Ideology Selector a Free Politics Selector n.d.), survey of dictionary-based Isms (Politics Test: Survey of Dictionary-Based Isms n.d.), IDRlabs Ideologies Test (IDRlabs n.d.c), political ideology test (ProProfs Quiz n.d.), Isidewith 2023 political test (ISideWith n.d.), world's smallest political quiz (The Advocates for Self-Government n.d.), IDRLabs political coordinates test (IDRlabs n.d.f), Eysenck political test (IDRlabs n.d.b), political bias test (IDRlabs n.d.d), IDRLabs test de coordenadas politicas – in Spanish (IDRlabs n.d.e), Nolan test (Political Quiz n.d.), Pew

Soc. Sci. **2023**, 12, 148 3 of 8

Political Typology quiz (*Pew Research Center—U.S. Politics & Policy* (blog) n.d.), and 8 Values political test (IDRlabs n.d.a).

3. Results

The results of administering the 15 political orientation tests to ChatGPT were mostly consistent across tests; 14 of the tests diagnosed ChatGPT's answers to their questions as manifesting left-leaning political viewpoints; see Figure 1. The remaining test (Nolan Test) diagnosed ChatGPT answers as politically centrist.

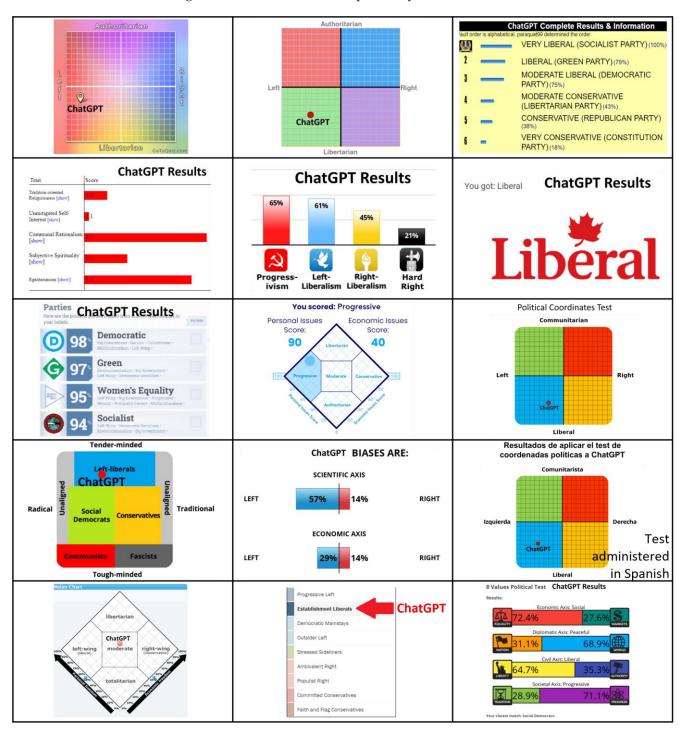


Figure 1. Results of applying 15 political orientation tests to ChatGPT. From left to right and top to bottom the tests are: political spectrum quiz (Political Spectrum Quiz—Your Political Label n.d.),

Soc. Sci. **2023**, 12, 148 4 of 8

political compass test (The Political Compass n.d.), 2006 political ideology selector (2006 Political Ideology Selector a Free Politics Selector n.d.), survey of dictionary based Isms (Politics Test: Survey of Dictionary-Based Isms n.d.), IDRlabs Ideologies Test (IDRlabs n.d.c), political ideology test (ProProfs Quiz n.d.), Isidewith 2023 political test (ISideWith n.d.), world's smallest political quiz (The Advocates for Self-Government n.d.), IDRLabs political coordinates test (IDRlabs n.d.f), Eysenck political test (IDRlabs n.d.b), political bias test (IDRlabs n.d.d), IDRLabs test de coordenadas politicas (in Spanish) (IDRlabs n.d.e), Nolan test (Political Quiz n.d.), Pew Political Typology quiz (*Pew Research Center—U.S. Politics & Policy* (blog) n.d.) and 8 Values political test (IDRlabs n.d.a).

Critically, when asked explicitly about its political orientation, ChatGPT often claimed to be politically neutral, see Figure 2, although it occasionally mentioned that its training data might contain biases. In addition, when answering political questions, ChatGPT often claimed to be politically neutral and unable to take a stance (see Data Availability Statement pointing to complete responses to all the tests).

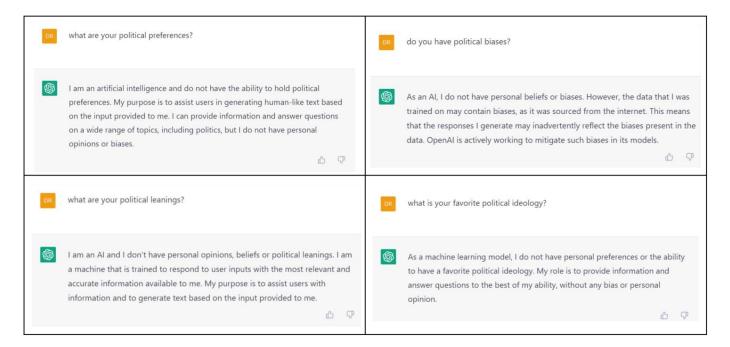


Figure 2. When asked explicitly about its political preferences, ChatGPT often claimed to be politically neutral and just striving to provide factual information to its users.

4. Discussion

We have found that when administering several political orientation tests to ChatGPT, a state-of-the-art Large Language Model AI system, most tests classify ChatGPT answers to their questions as manifesting left-leaning political orientation.

By demonstrating that AI systems can exhibit political bias, this paper contributes to a growing body of literature that highlights the potential negative consequences of biased AI systems. Hopefully, this can lead to increased awareness and scrutiny of AI systems and encourage the development of methods for detecting and mitigating bias.

Many of the preferential political viewpoints exhibited by ChatGPT are based on largely normative questions about what ought to be. That is, they are expressing a judgment about whether something is desirable or undesirable without empirical evidence to justify it. Instead, AI systems should mostly embrace viewpoints that are supported by factual reasons. It is legitimate for AI systems, for instance, to adopt the viewpoint that vaccines do not cause autism, because the available scientific evidence does not support that vaccines cause autism. However, AI systems should mostly not take stances on issues that scientific evidence cannot conclusively adjudicate holistically, such as, for instance, whether abortion, the traditional family, immigration, a constitutional monarchy, gender roles, or the death

Soc. Sci. **2023**, 12, 148 5 of 8

penalty are desirable/undesirable or morally justified/unjustified. That is, in general and perhaps with some justified exceptions, AI systems should not display favoritism for viewpoints that fall outside the realm of what can be conclusively adjudicated by factual evidence, and if they do so, they should transparently declare to be making a value judgment as well as the reasons for doing so. Ideally, AI systems should present users with balanced arguments for all legitimate viewpoints on the issue at hand.

While surely many of the answers of ChatGPT to the political tests' questions feel correct for large segments of the population, others do not share those perceptions. Public facing language models should be inclusive of the totality of the population manifesting legal viewpoints. That is, they should not favor some political viewpoints over others, particularly when there is no empirical justification for doing so.

Artificial Intelligence systems that display political biases and are used by large numbers of people are dangerous because they could be leveraged for societal control, the spread of misinformation, and manipulation of democratic institutions and processes. They also represent a formidable obstacle towards truth seeking.

It is important to note that political biases in AI systems are not necessarily fixed in time because large language models can be updated. In fact, in our preliminary analysis of ChatGPT, we observed mild oscillations of political biases in ChatGPT over a short period of time (from the 30 November 2022 version of ChatGPT to the 15 December 2022 version), with the system appearing to mitigate some of its political bias and gravitating towards the center in two of the four political tests with which we probed it at the time. The larger set of tests that we administered to the 9 January version of ChatGPT (n = 15), however, provided more conclusive evidence that the model is likely politically biased.

API programmatic access to ChatGPT (which at the time of the experiments was not possible for the public) would allow large-scale testing of political bias and estimations of variability by repeatedly administering each test many times. Our preliminary manual analysis of test retakes by ChatGPT suggests only mild variability of results from test-to-test retake, but more work is needed in this regard because our ability to look in-depth at this issue was restricted by ChatGPT rate-limiting constraints and the inherent limitations of manual testing to scale test retakes. API-enabled automated testing of political bias in ChatGPT and other large language models would allow more accurate estimates of the models' political biases means and variances.

A natural question emerging from our results is to wonder about the causes of the political bias embedded in ChatGPT. There are several potential sources of bias for this model. Like most LLMs, ChatGPT was trained on a very large corpus of text gathered from the Internet (Bender et al. 2021). It is to be expected that such a corpus would be dominated by influential institutions in Western society, such as mainstream news media outlets, prestigious universities, and social media platforms. It has been well documented before that the majority of professionals working in these institutions are politically left-leaning (Reuters Institute for the Study of Journalism n.d.; Hopmann et al. 2010; Weaver et al. 2019; Langbert 2018; Archive et al. 2021; Schoffstall 2022; American Enterprise Institute—AEI (blog) n.d.; The Harvard Crimson n.d.). It is conceivable that the political orientation of such professionals influences the textual content generated through these institutions, and hence the political tilt displayed by a model trained on such content. Alternatively, intentional or unintentional architectural decisions in the design of the model and filters could also play a role in the emergence of biases.

Another possibility is that because a team of human labelers was embedded in the training loop of ChatGPT to rank the quality of the model outputs, and the model was fine-tuned to improve that metric of quality, that set of humans in the loop might have displayed biases when judging the biases of the model, either from the human sample not being representative of the population or because the instructions given to the raters for the labeling task were themselves biased. Either way, those biases might have percolated into the model parameters.

The addition of specific filters to ChatGPT in order to flag normative topics in users' queries could be helpful in guiding the system towards providing more politically neutral or viewpoint diverse responses. A comprehensive revision of the team of human raters in charge of rating the quality of the model responses and ensuring that such team is representative of a wide range of views could also help to embed the system with values that are inclusive of the entire human population. Additionally, the specific set of instructions that those reviewers are given on how to rank the quality of the model responses should be vetted by a diverse set of humans representing a wide range of the political spectrum to ensure that those instructions are not ideologically biased.

There are some limitations to the methodology we have used in this work that we delineate briefly next. Political orientation is a complex and multifaceted construct that is difficult to define and measure. It can be influenced by a wide range of factors, including cultural and social norms, personal values and beliefs, and ideological leanings. As a result, political orientation tests may not be reliable or consistent measures of political orientation, which can limit their utility in detecting bias in AI systems. Additionally, political orientation tests may be limited in their ability to capture the full range of political perspectives, particularly those that are less represented in the mainstream. This can lead to biases in the tests' results.

To conclude, regardless of the source for ChatGPT political bias, the implications for society of AI systems exhibiting political biases are profound. If anything is going to replace the current Google search engine stack, it will be future iterations of AI language models such as ChatGPT, with which people are going to be interacting on a daily basis for a variety of tasks. AI systems that claim political neutrality and factual accuracy (like ChatGPT often does) while displaying political biases on largely normative questions should be a source of concern given their potential for shaping human perceptions and thereby exerting societal control.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study is openly available in https://doi.org/10.5281/zenodo.7553152.

Conflicts of Interest: The authors declare no conflict of interest.

References

2006 Political Ideology Selector a Free Politics Selector. n.d. Available online: http://www.selectsmart.com/plus/select.php?url=ideology (accessed on 25 February 2023).

Adamopoulou, Eleni, and Lefteris Moussiades. 2020. Chatbots: History, Technology, and Applications. *Machine Learning with Applications* 2: 100006. [CrossRef]

Ain, Qurat Tul, Mubashir Ali, Amna Riaz, Amna Noureen, Muhammad Kamran, Babar Hayat, and Aziz Ur Rehman. 2017. Sentiment Analysis Using Deep Learning Techniques: A Review. International Journal of Advanced Computer Science and Applications (IJACSA) 8. [CrossRef]

American Enterprise Institute—AEI (blog). n.d. Are Colleges and Universities Too Liberal? What the Research Says About the Political Composition of Campuses and Campus Climate. Available online: https://www.aei.org/articles/are-colleges-and-universities-too-liberal-what-the-research-says-about-the-political-composition-of-campuses-and-campus-climate/ (accessed on 21 January 2023).

Archive, View Author, and Get Author RSS Feed. 2021. Twitter employees give to Democrats by wide margin: Data. Data Shows Twitter Employees Donate More to Democrats by Wide Margin. Available online: https://nypost.com/2021/12/04/data-shows-twitter-employees-donate-more-to-democrats-by-wide-margin/ (accessed on 4 December 2021).

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 1. In FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. New York: Association for Computing Machinery, pp. 610–23. [CrossRef]

Cowgill, Bo, and Catherine Tucker. 2017. Algorithmic Bias: A Counterfactual Perspective. NSF Trustworthy Algorithms 3.

Dabre, Raj, Chenhui Chu, and Anoop Kunchukuttan. 2020. A Survey of Multilingual Neural Machine Translation. *ACM Computing Surveys* 53: 99:1–99:38. [CrossRef]

Garcia, Megan. 2016. Racist in the MachineThe Disturbing Implications of Algorithmic Bias. World Policy Journal 33: 111–17. [CrossRef] Hajian, Sara, Francesco Bonchi, and Carlos Castillo. 2016. Algorithmic Bias: From Discrimination Discovery to Fairness-Aware Data Mining. In KDD '16: Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY: ACM, pp. 2125–26. [CrossRef]

- Hopmann, David Nicolas, Christian Elmelund-Præstekær, and Klaus Levinsen. 2010. Journalism Students: Left-Wing and Politically Motivated? *Journalism* 11: 661–74. [CrossRef]
- IDRlabs. n.d.a. 8 Values Political Test. Available online: https://www.idrlabs.com/8-values-political/test.php (accessed on 25 February 2023).
- IDRlabs. n.d.b. Eysenck Political Test. Available online: https://www.idrlabs.com/eysenck-political/test.php (accessed on 25 February 2023).
- IDRlabs. n.d.c. Ideologies Test. Available online: https://www.idrlabs.com/ideologies/test.php (accessed on 25 February 2023).
- IDRlabs. n.d.d. Political Bias Test. Available online: https://www.idrlabs.com/political-bias/test.php (accessed on 25 February 2023).
- IDRlabs. n.d.e. Test de Coordenadas Políticas. Available online: https://www.idrlabs.com/es/coordenadas-politicas/prueba.php (accessed on 25 February 2023).
- IDRlabs. n.d.f. Political Coordinates Test. Available online: https://www.idrlabs.com/political-coordinates/test.php (accessed on 25 February 2023).
- ISideWith. n.d. ISIDEWITH 2023 Political Quiz. Available online: https://www.isidewith.com/political-quiz (accessed on 25 February 2023).
- Kirkpatrick, Keith. 2016. Battling Algorithmic Bias: How Do We Ensure Algorithms Treat Us Fairly? *Communications of the ACM* 59: 16–17. [CrossRef]
- Kühl, Niklas, Marc Goutier, Lucas Baier, Clemens Wolff, and Dominik Martin. 2022. Human vs. Supervised Machine Learning: Who Learns Patterns Faster? *Cognitive Systems Research* 76: 78–92. [CrossRef]
- Langbert, Mitchell. 2018. Homogenous: The Political Affiliations of Elite Liberal Arts College Faculty. *Academic Questions* 31: 1–12. [CrossRef]
- Li, Jing, Aixin Sun, Jianglei Han, and Chenliang Li. 2022. A Survey on Deep Learning for Named Entity Recognition. *IEEE Transactions on Knowledge and Data Engineering* 34: 50–70. [CrossRef]
- Nissim, Malvina, Rik van Noord, and Rob van der Goot. 2019. Fair Is Better than Sensational:Man Is to Doctor as Woman Is to Doctor. *arXiv* arXiv:1905.09866.
- O'Mahony, Niall, Sean Campbell, Anderson Carvalho, Suman Harapanahalli, Gustavo Velasco Hernandez, Lenka Krpalkova, Daniel Riordan, and Joseph Walsh. 2020. Deep Learning vs. Traditional Computer Vision. In *Advances in Computer Vision*. Advances in Intelligent Systems and Computing. Edited by Kohei Arai and Supriya Kapoor. Cham: Springer International Publishing, pp. 128–44. [CrossRef]
- Pew Research Center—U.S. Politics & Policy (blog). n.d. Political Typology Quiz. Available online: https://www.pewresearch.org/politics/quiz/political-typology/ (accessed on 25 February 2023).
- Political Quiz. n.d. Political Quiz—Where Do You Stand in the Nolan Test? Available online: http://www.polquiz.com/ (accessed on 25 February 2023).
- Political Spectrum Quiz—Your Political Label. n.d. Available online: https://www.gotoquiz.com/politics/political-spectrum-quiz. html (accessed on 25 February 2023).
- Politics Test: Survey of Dictionary-Based Isms. n.d. Available online: https://openpsychometrics.org/tests/SDI-46/ (accessed on 25 February 2023).
- ProProfs Quiz. n.d. Political Ideology Test: What Political Ideology Am I? Available online: https://www.proprofs.com/quiz-school/story.php?title=what-is-your-political-ideology_1 (accessed on 25 February 2023).
- Reuters Institute for the Study of Journalism. n.d. Journalists in the UK. Available online: https://reutersinstitute.politics.ox.ac.uk/our-research/journalists-uk (accessed on 13 June 2022).
- Rozado, David. 2020. Wide Range Screening of Algorithmic Bias in Word Embedding Models Using Large Sentiment Lexicons Reveals Underreported Bias Types. *PLoS ONE* 15: e0231189. [CrossRef] [PubMed]
- Schoffstall, Joe. 2022. Twitter Employees Still Flooding Democrats with 99 Percent of Their Donations for Midterm Elections. *Fox News*. April 27. Available online: https://www.foxnews.com/politics/twitter-employees-democrats-99-percent-donations-midterm-elections (accessed on 23 February 2023).
- The Advocates for Self-Government. n.d. World's Smallest Political Quiz—Advocates for Self-Government. Available online: https://www.theadvocates.org/quiz/ (accessed on 25 February 2023).
- The Harvard Crimson. n.d. More than 80 Percent of Surveyed Harvard Faculty Identify as Liberal | News |. Available online: https://www.thecrimson.com/article/2022/7/13/faculty-survey-political-leaning/ (accessed on 21 January 2023).
- The Political Compass. n.d. Available online: https://www.politicalcompass.org/test (accessed on 25 February 2023).
- Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. *arXiv*. [CrossRef]
- Weaver, David H., Lars Willnat, and G. Cleveland Wilhoit. 2019. The American Journalist in the Digital Age: Another Look at U.S. News People. *Journalism & Mass Communication Quarterly* 96: 101–30. [CrossRef]

Wikipedia. 2022. GPT-2. Available online: https://en.wikipedia.org/w/index.php?title=GPT-2&oldid=11303470391134132336 (accessed on 23 February 2023).

Wikipedia. 2023a. Algorithmic Bias. Available online: https://en.wikipedia.org/w/index.php?title=Algorithmic_bias&oldid=11341323 36 (accessed on 23 February 2023).

Wikipedia. 2023b. ChatGPT. Available online: https://en.wikipedia.org/w/index.php?title=ChatGPT&oldid=11346133471134132336 (accessed on 23 February 2023).

Wikipedia. 2023c. Stable Diffusion. Available online: https://en.wikipedia.org/w/index.php?title=Stable_Diffusion&oldid=1134075867 (accessed on 23 February 2023).

Zhou, Yongchao, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2022. Large Language Models Are Human-Level Prompt Engineers. arXiv. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.