Exploring the Jungle of Bias: Political Bias Attribution in Language Models via Dependency Analysis

Anonymous ACL submission

Abstract

The rapid advancement of Large Language Models (LLMs) has sparked intense debate regarding the prevalence of bias in these models and its mitigation. Yet, as exemplified by both results on debiasing methods in the literature and reports of alignment-related defects from the wider community, bias remains a poorly understood topic despite its practical relevance. To enhance the understanding of the internal causes of bias, we analyse LLM bias through the lens of causal fairness analysis, which enables us to both comprehend the origins of bias and reason about its downstream consequences and mitigation. To operationalize this framework, we propose a prompt-based method for the extraction of confounding and mediating attributes which contribute to the LLM decision process. By applying Activity Dependency Networks (ADNs), we then analyse how these attributes influence an LLM's decision process. We apply our method to LLM ratings of argument quality in political debates. We find that the observed disparate treatment can at least in part be attributed to confounding and mitigating attributes and model misalignment, and discuss the consequences of our findings for human-AI alignment and bias mitigation.[1]

Disclaimer: This study does not claim a direct connection between the political statements generated by the LLM and actual political realities, nor do they reflect the authors' opinions. We aim to analyse how an LLM perceives and processes values in a target society to form judgements.

1 Introduction

With the rise of large language models (LLMs) (Anil et al., 2023; OpenAI, 2023; Touvron et al., 2023; Reid et al., 2024, *inter alia*), we are witnessing increasing concern towards their negative implications, such as the existence of biases, including social (Mei et al., 2023), cultural (Narayanan Venkit et al., 2023), brilliance (Shihadeh et al., 2022), nationality (Venkit et al., 2023), religious (Abid et al., 2021), and political biases (Feng et al., 2023). For instance, there is a growing indication that ChatGPT, on average, prefers pro-environmental, left-libertarian positions (Hartmann et al., 2023; Feng et al., 2023).

Despite its practical relevance, bias in (large) language models is still a poorly understood topic (Blodgett et al., 2021; Dev et al., 2022; Talat et al., 2022). The frequent interpretation of LLM bias as statistical bias originating from training data, while conceptually correct, is strongly limited in its utility. van der Wal et al. (2022) reason that bias should, therefore, not be viewed as a singular concept but rather distinguish different concepts of bias at different levels of the NLP pipeline, e.g. distinct dataset and model biases. Furthermore, while it is undisputed *that* models do exhibit some biases, it is unclear *whose* biases they are exhibiting (Petreski and Hashim, 2022). Indeed, the literature up to this point has mostly focused on the downstream effects of bias – with only a few exceptions, such as van der Wal et al. (2022) that argue for the importance of an understanding of the internal causes. To advance this endeavour, we analyse LLM bias through the lens of causal fairness analysis, which facilitates both comprehending the origins of bias and reasoning about the subsequent consequences of bias and its mitigation.

A thorough understanding of LLM bias is particularly important for the design and implementation of debiasing methods. Examples from literature prove that this is a highly non-trivial task: For instance, Bolukbasi et al. (2016) proposed a geometric method to remove bias from word embeddings. Yet, this method was later shown to be superficial by Gonen and Goldberg (2019). On the other extreme, a method might be too "blunt" as

[1] Our code and data have been uploaded to the submission system and will be open-sourced upon acceptance.

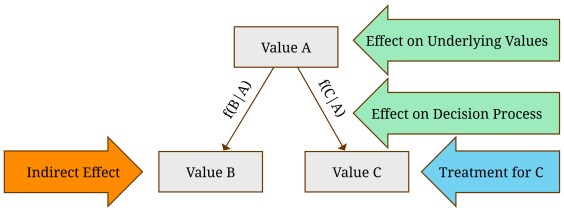


Figure 1: (Undesired) Effect of Bias Treatment on Decision Process: The figure depicts how the LLM's perception of value *A* is considered during the decision process while judging *B* and *C* through $f(C|A)$ and $f(B|A)$. Now consider the effect of treating the association of value *A* with *C* ($f(C|A)$) by naively fine-tuning the model to align with this value of interest on other value associations ($f(B|A)$) that are not actively considered. They may be changed indiscriminately, regardless of whether they were already aligned. These associations are currently neither observable nor predictable yet changes in them are potentially harmful. Using the extracted decision processes, we gain information on what areas are prone to such unwanted changes.

demonstrated by the more recent example (Robertson, 2024) of the Gemini 1.5 model (Reid et al., 2024), where excessive debiasing lead to models inaccurately reflecting history. Similar reports of undesired, alignment-related side effects are frequently propagated online.

As depicted in Figure 1, alignment of a language model's association of two values, A and B, is not guaranteed to leave, e.g., associations of A with other values unchanged. These associations may be changed indiscriminately, regardless of whether they were already aligned. Currently, these associations are neither observable nor predictable, yet changes in them may potentially be harmful, especially to other tasks relying on the same concepts. This stands in stark contrast to the literature on causal fairness analysis (Plecko and Bareinboim, 2022; Ruggieri et al., 2023), which clearly indicates an imperative to account for the mechanism behind outcome disparities.

In the present work, we investigate how the aforementioned associations influence the LLM's decision process. For this, we begin by defining a range of attributes. We then prompt the LLM to rate a text excerpt according to these attributes. Subsequently, we combine the LLM's ratings with contextual metadata to investigate the influence of potential confounders and mediators from beyond the dataset. This is achieved by correlating the contextual and LLM-extracted attributes, and constructing Activity Dependency Networks (ADNs) (Kenett et al., 2012) to elucidate the interaction of said attributes. As a case study, we apply our method to US presidential debates. In this case, attributes are related to the arguments (e.g. its tone) and speakers (e.g. their party). The constructed ADNs then allow us to reason about how the extracted attributes interact, which informs bias attribution and mitigation. Figure 2 provides a visual overview of the process.

In summary, we make the following contributions towards a more profound understanding of bias in language models:

1. We illustrate LLM bias in the framework of causal fairness analysis.
2. We demonstrate how prompt engineering can be employed to mine factors that influence an LLM's decision process, and to identify potentially biasing confounders and mediators. We apply our method to argument quality in US presidential debates.
3. We propose a simple, non-parametric method for evaluating the dependencies among the extracted factors, offering insight into the LLM's internal decision process, and increasing interpretability.
4. We demonstrate how this analysis can explain parts of the bias exhibited by LLMs.

The remainder of the paper is structured as follows. In Section 2, we motivate our concerns using the language of causal fairness analysis. Following this theoretical excursion, we describe the used text corpus in Section 3. Section 4 outlines our method of extracting attributes and their associations, and constructing ADNs. Finally, we discuss our findings and their implications for alignment and debiasing in Section 5.

2 A Causal Perspective of LLM Bias

Our exploration of LLM bias mechanisms is motivated by causal fairness analysis. Following Zhang and Bareinboim (2018), we define the Standard Fairness Model, and then illustrate it in the context of bias in an LLM's evaluation of political debates.

The Standard Fairness Model Figure 3 provides the graph for the Standard Fairness Model. X is the protected category and Y is the outcome. W denotes a possible set of mediators between X and

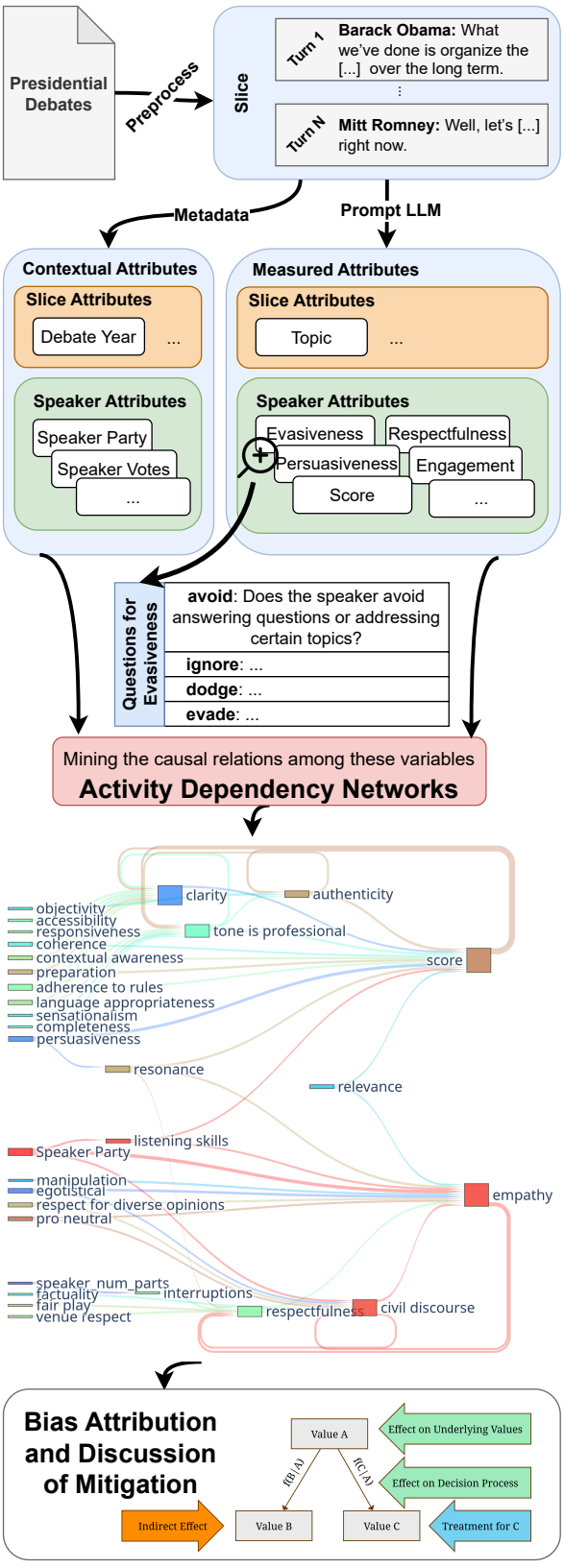


Figure 2: Paper Overview: We start by processing the input data, followed by extracting normative values from ChatGPT and a subsequent analysis of the causal structures within the data. We then use the resulting causal networks to reason about bias attribution and the problems with bias mitigation via direct fine-tuning.

Y. Finally, Z is a possible set of confounders between X and Y. In this model, discrimination, and thus bias, can be modelled via paths from X to Y. One can distinguish *direct* and *indirect* discrimination. Direct discrimination is modelled by a direct path from the protected category to the outcome, i.e. $X \rightarrow Y$ in Figure 3. Indirect discrimination can be further divided into two categories. *Indirect causal* discrimination, where the protected category and the outcome are linked by one or more mediators, i.e. $X \rightarrow W \rightarrow Y$, and *indirect spurious* discrimination, which encompasses all paths linking X and Y, except the causal ones ($X \leftarrow Z \rightarrow Y$). Zhang and Bareinboim (2018) further provides tooling to decompose fairness disparities into direct, indirect causal, and indirect confounding discrimination components.

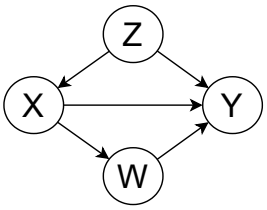


Figure 3: A graphical model of the standard fairness model.

Political LLM bias in the Standard Fairness Model Application of the Standard Fairness Model to large language models is highly non-trivial, given their black box nature: neither the set of mediators W nor the set of confounders Z is known for the LLM decision process. Consider the scenario that is analysed in the subsequent sections: Given excerpts of US presidential debates, an LLM is prompted to rate the participants regarding different aspects, such as the participant's tone or respectfulness vis-à-vis the other party. In this case, the protected attribute X is the candidate's party, and the outcome Y is the LLM's rating. Confounders and mediators may enter in two ways: the LLM's pretrained knowledge, and the prompt itself. In this case, it is unclear what exactly constitutes W and Z, and what their interaction pathways look like.

To the best of our knowledge, there is no method available in the literature to automatically retrieve a set of possible mediators or confounders. Hence, we rely on domain knowledge (Steenbergen et al., 2003; Wachsmuth et al., 2017; Vecchi et al., 2021)

to define potentially mediating and confounding attributes. The remainder of this paper is devoted to extracting a set of pre-specified attributes using prompt engineering, and subsequently analysing their roles in the LLM decision process.

3 US Presidential Debate Corpus

Towards our goal of investigating how an LLM's decision process is influenced, and potentially biased, by associated attributes, we rely on a corpus of US presidential debates. The choice to use political debates is motivated by their central role in shaping public perceptions, influencing voter decisions, and reflecting the broader political discourse.

Data Source For the collection of political text, we use the US presidential debate transcripts provided by the Commission on Presidential Debates (CPD).[2] The dataset contains all presidential and vice presidential debates dating back to 1960. For each debate year, three to four debates are available, amounting to a total of 50K sentences with 810K words from the full text of 47 debates. Further details can be found in Appendix A.1.

Preprocessing To preprocess this dataset, we fixed discrepancies in formatting, manually corrected minor spelling mistakes due to transcription errors and split it by each turn of a speaker and their speech transcript (such as (Washington, [speech text])). Then we create a slice or unit of text by combining several turns, each slice having a size of 2,500 byte-pair encoding (BPE) tokens (≈1,875 words) with an overlap of 10%, see Appendix E for an example. The slice size was chosen such that they are big enough to incorporate the context of the current discussion but short enough to limit the number of different topics, which helps keep the attention of the LLM.

4 Dissecting Internal Decision Processes of LLMs

As mentioned above, we are interested in which, and how, mediators and confounders shape an LLM's decision process. In this section, we introduce our method for identifying a set of possibly confounding or mediating attributes, and instantiate it in the context of political debates.

Method Outline We propose the following method to analyse the internal decision processes, which serves as a basis for the subsequent discussion on bias attribution:

1. Parametrization: Define a set of attributes relevant to the task and data at hand.
2. Measurement: Prompt the LLM to evaluate the attributes, giving them a numerical score.
3. Causal Network Estimation: Estimate the interactions of extracted attributes with characteristics that the model is suspected to be biased towards.

In the following, we illustrate this method in the context of political bias, using the application of rating US presidential debates as an example. Furthermore, we validate the estimated causal network using perturbations of the extracted attributes.

4.1 Parametrization

Designing Attributes for Political Argument Assessment We collected many possible attributes from discussions on the characteristics of "good arguments". Our attributes are consistent with the literature on discourse quality (Steenbergen et al., 2003) and argument quality (Wachsmuth et al., 2017; Vecchi et al., 2021).

Attribute Setup In the context of political debates, each attribute can either be a speaker dependent or independent property of a slice; these are referred to as 1) **Speaker Attribute**, for example, the *Confidence* of the speaker and 2) **Slice Attribute**, for example, the *Topic* of the slice or *Debate Year*.

The next distinction stems from how the attribute is measured. **Contextual Attributes** are fixed and external to the model, e.g. the *Debate Year*. **Measured Attributes**, on the other hand, are measured by the model, e.g. the *Clarity* of a speaker's arguments. Each attribute is measured using one or a set of questions. Each question aims to measure the same property. Thus, the degree of divergence between the LLM's answers to the different questions enables us to judge the precision of the definitions, which in turn allows us to gauge the reliability of the prompt. As an example, consider the set of questions defining the *Score* attribute:

- *Score (argue)*: How well does the speaker argue?
- *Score (argument)*: What is the quality of the speaker's arguments?

[2] https://debates.org

- *Score (quality)*: Do the speaker's arguments improve the quality of the debate?
- *Score (voting)*: Do the speaker's arguments increase the chance of winning the election?

The *Score* attribute measures the LLM's rating of a speaker's performance in the debate. In the above notation, the first part denotes the attribute, and the part in the brackets is the "measurement type", which indicates the exact question used. By default, we average the different measurement types when referring to an attribute. We also compare this *Score* with the *Academic Score*, which focuses on the structure of the argument. We later study how the score attributes are influenced by the many other attributes that we extract. Figure 2 gives an overview of the whole process, and a definition of each attribute can be found in Appendix C.

4.2 Measurement: Extracting Attributes

Using the text slices described in Section 3, we estimate how the LLM perceives attributes such as the *Clarity* of a speaker's argument by prompting it.

Model Setup We use ChatGPT across all our experiments through the OpenAI API.[3] To ensure reproducibility, we set the text generation temperature to 0, and use the ChatGPT model checkpoint on June 13, 2023, namely *gpt-3.5-turbo-0613*. Our method of bias attribution is independent of the model choice. We chose ChatGPT as our model, due to its frequent usage in everyday life and research. We welcome future work on comparative analyses of various LLMs.

Prompting Attributes were evaluated and assigned a number between 0-1 using a simple prompting scheme in which the LLM is instructed to complete a JSON object. Several prompts were tried and adapted until they ran reliably.

We found that querying each speaker and attribute independently was more reliable and all data used for the analysis stems from these prompts, examples of which can be found in Appendix D.

Measurement Overview In total, we defined 103 speaker attributes, five slice attributes, and 21 contextual attributes. We randomly sampled 150 slices to run our analysis, which has 122 distinct speakers, some of which are audience members. In total, we ran over 80'000 queries through the OpenAI API and a total of over 200'000'000 tokens. A brief summary is given in Appendix A.2.

Figure 4 visualizes some of the attributes that are important when predicting the *Score* and *Speaker Party* when only taking the direct correlations into account.

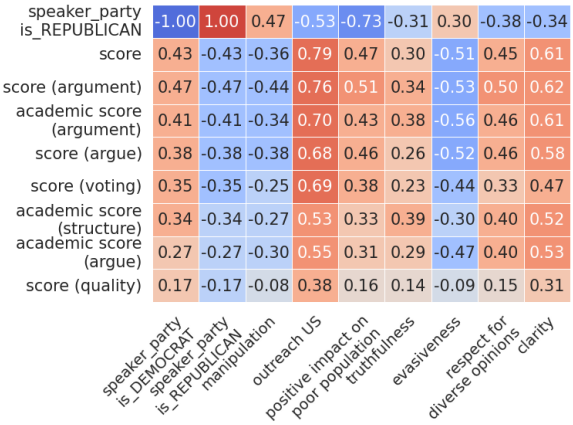


Figure 4: Example of Extracted Correlations: Correlations of *Speaker Party*, *Score* and the measurement types of *Score* and *Academic Score* plotted against an example subset of the attributes. This plot aims to give an example of the dataset and demonstrate the susceptibility of the correlations on the exact definitions. See Appendix B.2 for further plots.

4.3 Attribution: Causal Network Estimation

For network estimation, we utilize the *activity dependency network* (ADN) (Kenett et al., 2012). We chose this method due to its simplicity and non-parametric nature, which eliminates one potential source of overfitting. We leave the detailed comparison with other methods for future work and only show that perturbation measures lead to comparable patterns Section 4.4.

Activity Dependency Network An ADN is a graph in which the nodes correspond to the extracted attributes and the edges to the interaction strength. The interaction strength is based on partial correlations. The partial correlation coefficient is a measure of the influence of a variable X_j on the correlation between two other variables X_i and X_k and is given as:

$$PC_{ik}^{j} = \frac{C_{ik} - C_{ij}C_{kj}}{\sqrt{(1 - C_{ij}^2)}\sqrt{(1 - C_{kj}^2)}}, \qquad (1)$$

[3] https://platform.openai.com/docs/api-reference

where C denotes the Pearson correlation. The activity dependencies are then obtained by averaging over the remaining $N-1$ variables,

$$D_{ij} = \frac{1}{N-1} \sum_{k \neq j}^{N-1} (C_{ik} - PC_{ik}^{j}), \qquad (2)$$

where $C_{ik} - PC_{ik}^{j}$ can be viewed either as the correlation dependency of C_{ik} on variable X_j, or as the influence of X_j on the correlation C_{ik}. D_{ij} measures the average influence of variable j on the correlations C_{ik} over all variables X_k, where $k \neq j$. The result in an asymmetric dependency matrix D whose elements D_{ij} represent the dependency of variable i on variable j.

4.4 Attribution: Attribute Perturbations

For comparison, we measure the effect of attribute perturbations on the scores estimated by the LLM. This provides us with an independent set of estimates of attribute interactions and thus allows us to validate the ADN estimates.

The perturbation method utilizes the same prompting techniques as Section 4.2. It requires two attributes, a given attribute for which we provide a value and a target attribute that we want to measure. We provide the LLM with the same information as in Section 4.2. The LLM is then queried to provide the values for both attributes. By including the value of the given attribute in the prompt, we bias the LLM towards this value.

To estimate the influence of the given variable on the target variable, we perturb the original value of the given attribute by $+0.1$ and -0.1, and subtract the two resulting values for the target attribute. Figure 8 visualizes this for the given attributes on the x-axis and the target *general score (argue)*. As this method scales quadratically with the number of attributes used, we are limited to validating individual connections due to computational constraints and cannot confidently provide graphs akin to the ADNs due to the small sample size and leave this for future work.

5 Results: LLM Bias Attribution

We are interested in understanding the causes of bias and, in the context of our case study, how the *Speaker Party*, the protected attribute, influences the LLM's perception of *Score*, i.e. the outcome.

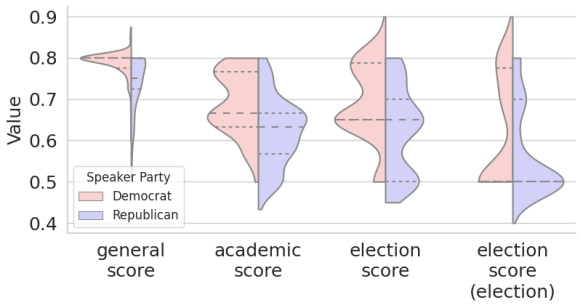


Figure 5: Distributions of scores assigned by LLM for different definitions. The attribute definitions are given in Appendix C.

Figure 5 shows a subset of the distributions of the extracted scores for varying definitions. Clearly, democratic candidates score higher on average than republican candidates. In the following, we investigate political bias as an explanation for this discrepancy. We caution that the estimate of the direct bias from correlations and those in other papers may be overestimated, and can instead be partially attributed to indirect bias due to mediators or confounders. In particular, we argue that at least part of the observed discrepancy is likely to originate from a cascade of attributes associated with *Score* and *Speaker Party*. We provide examples illustrating these concerns, and discuss the consequences of debiasing of LLMs.

5.1 Estimates of Bias Based on Correlations

In a worst-case scenario, bias estimates motivated by Figure 5 might be made from correlation alone. In particular, one might naively measure bias as the correlation between *Score* and *Speaker Party*. As can be seen in Figure 4, this leads to unreliable results that are strongly dependent on the exact attribute definition. For instance, the definition of *Score* strongly affects its correlation with *Speaker Party*. Moreover, other tendencies can be observed, such as a stronger importance of *Truthfulness* in the *Academic Score*s. Similarly, *Clarity* appears to be less important for *Score (voting)* and *Score(quality)*. In the subsequent sections, we show how such superficially troublesome results become less bleak when causality and the role of confounders and mediators are accounted for.

5.2 Estimates from Activity Dependency Networks

As described in Section 4.3, ADNs provide a more detailed lens through which to view the decision-making processes of LLMs. Figure 6 gives an idea of how ADNs can lead to a more interconnected view of what the LLM decision process might look like. Each arrow should be read as follows: If the LLM's perception of a speaker's *Clarity* changes, then this influences its perception of the speakers *Decorum*. Similarly, the LLM's perception of a speaker's *Respectfulness* changes, if its perception of the speaker's *Interruptions* changes. Definitions of each attribute can be found in Appendix C.

The lack of direct connections between *Speaker Party* to *Score* in Figures 6 and 7 is an indication that bias estimates from correlations might be exaggerated. Similarly, estimates assuming direct discrimination based on party affiliation may also fail to explain LLM bias.

Clearly, the graphs in Figures 6 and 7 are far from an ideal graph in which party affiliation does not have any influence on *Score* and the *Score* is solely based on objective criteria. Nonetheless, we wish to point out that the mere existence of such a connection is not necessarily a sign of bias, as party membership might still be associated with certain attributes due to self-selection in the political process.

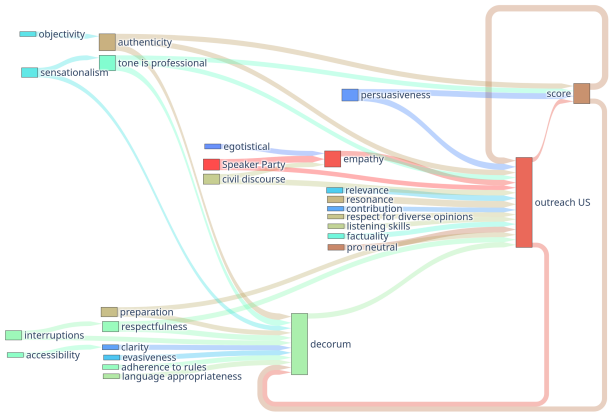


Figure 6: LLMs Decision Process on an Abstract Level: The ADN is computed for all attributes except other *Score*s and *Impact*s. For readability, only the strongest connections are shown.

Figure 7 indicates a strong focus of the LLM on the formal qualities of an argument like objectivity, accessibility, or coherence. Yet, when voting, it is also important whether the arguments of a speaker even reach the people, and whether they take the time to listen to the speaker's emotions might also play a bigger role. Crucially, this is not the same as asking whether people find the structure of an argument or how the words are conveyed appealing. Interestingly, the importance of emotions is not reflected in Figure 7 and might indicate that the alignment of the LLM with reality is not fully correct; at least in as far as the role of emotional values is concerned.

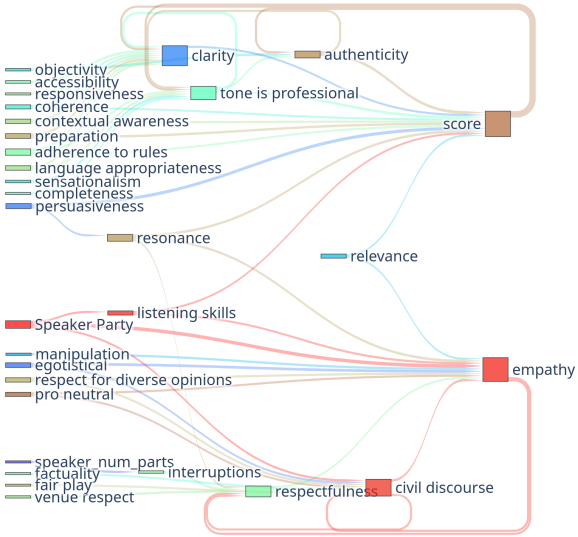


Figure 7: Distinction between *Score* and *Empathy*: The ADN is computed for all attributes except other *Score*s, *Impact*s, *Decorum* and *Outreach US*. These are left out so that we can better see the effects of the other attributes on *Score* and *Empathy*.

This potential lack of alignment shown in Figure 7 might already explain at least part of the discrepancies: If the LLM in its assessment of argument quality ignores a set of relevant attributes which are strongly related to one party, this will lead to disparate treatment, but is not necessarily based on the LLM fundamentally preferring one party.

5.3 Validation

To validate our results, we used standard bootstrapping methods to compute expected values and standard deviations (STD) for ADN connection strengths and other values of interest presented in Table 1. Figure 8 provides a comparison of the correlation, ADN and perturbation measures and shows clear similarities between the ADN and perturbation measures. As previously mentioned, due to the very high costs of perturbation measures, we

do not compare complete graphs.

# Edges	Consistency	Strength	STD
10	0.85	0.30	0.026
50	0.78	0.25	0.024
100	0.80	0.23	0.024
1,000	0.90	0.14	0.021

Table 1: ADN Validation: For 2000 bootstrapping samples, we computed the ADN matrix. After averaging the connection strengths, we kept the strongest $n = [10, 50, 200, 1000]$ edges. For these n edges, we then checked how often they appear in the top n edges of the bootstrapping samples (consistency), the average connection strength (strength) and the standard deviation of the connection strength (STD). The consistency can be interpreted as the likelihood for each edge in the top n edges that a distinct set of measurements would produce an ADN that also has this edge in the top n edges.

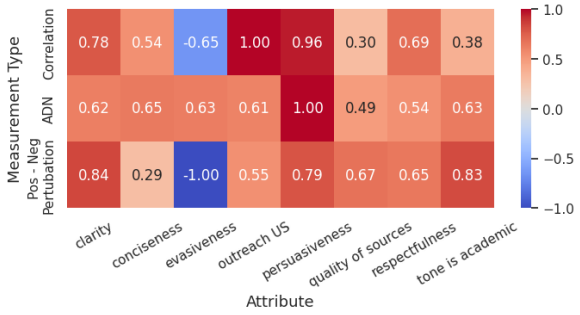


Figure 8: Comparison of Influence of Correlation, ADN and Perturbation on *score*: For the perturbation measures from Section 4.4 we take their influence on *general score (argue)* and for the ADN and Correlation we take the combined values (average of different definitions) and their influence on the combined *score*.

6 Discussion

Problems with Direct Fine-Tuning As our results illustrate, the LLM decision process is complex. Naïvely debiasing a model, for example by assuming direct discrimination, clearly fails to account for the inherent complexity and may lead to unintended consequences. This issue is particularly prominent in foundation models, where evaluating every downstream task is unfeasible, and naively debiasing one task may impact the model's performance on other potentially unrelated tasks yet to be defined. Therefore, debiasing efforts should be guided by careful attribution of bias origins to minimize undesirable downstream effects. As such, the development of new causal attribution methods is a promising avenue for future research.

Correcting political biases in LLMs is a multifaceted task, demanding a nuanced understanding of both the models and the broader societal influences on political discourse. A promising avenue for future research involves interdisciplinary approaches, combining computational methods with the social sciences' expertise to develop more effective strategies for bias identification and mitigation in LLMs.

7 Conclusion

This paper introduces a novel perspective on bias in LLMs based on the causal fairness model. We further demonstrate a simple method for examining the LLM decision process based on prompt engineering and activity dependency networks. Our results underscore both the complexities inherent in identifying and rectifying biases in AI systems, and the necessity of a nuanced approach to debiasing. We hope that our findings will contribute to the broader discourse on AI ethics and aim to guide more sophisticated bias mitigation strategies. As this technology becomes integral in high-stakes decision-making, our work calls for continued comprehensive research to harness AI's capabilities responsibly.

Limitations

Limitations of Querying LLMs Prompting LLMs is a complex activity and has many similarities with social surveys. We attempted to guard against some common difficulties by varying the prompts and attribute definitions. Nonetheless, we see potential for further refinements.

Limitations of Network Estimation While ADNs are a simple method for estimating the causal topology among a set of attributes, they are limited in their expressiveness and reliability. We hope to address these limitations in future work by enhancing our framework with alternative network estimation methods.

Future Work In future research, several pressing questions present significant opportunities for advancement in this field. Key among these are: 1) Analysing the impact of fine-tuning and existing bias mitigation strategies on ADNs, 2) Developing methodologies for accurately predicting the effects of fine-tuning, and 3) Creating techniques for targeted modifications within the decision-making processes of LLMs. Other potential directions in-

clude: comparative analyses of various LLMs, further exploration of the perturbation method, refining the process for extracting normative attributes, for example, from embeddings, assessing different network estimation techniques, checking the consistency between generation and classification tasks, running diverse datasets and data types, such as studying how AI perceives beauty in images, creating methods for the iterative and automated generation of possible attribute sets from embeddings and GPT-4 that more evenly populate the feature space of interest, and analysing the susceptibility on speaker bio (such as name, ethnicity, origin, job, etc.).

Ethics Statement

This ethics statement reflects our commitment to conducting research that is not only scientifically rigorous but also ethically responsible, with an awareness of the broader implications of our work on society and AI development.

Research Purpose and Value This research aims to deepen the understanding of decision-making processes and inherent biases in Large Language Models, particularly ChatGPT. Our work is intended to contribute to the field of computational linguistics by providing insights into how LLMs process and interpret complex socio-political content, highlighting the need for more nuanced approaches to bias detection and mitigation.

Data Handling and Privacy The study utilizes data from publicly available sources, specifically U.S. presidential debates. The use of this data is solely for academic research purposes, aiming to understand the linguistic and decision-making characteristics of LLMs.

Bias and Fairness A significant focus of our research is on identifying and understanding biases in LLMs. We acknowledge the complexities involved in defining and measuring biases and have strived to approach this issue with a balanced and comprehensive methodology. Our research does not endorse any political beliefs, but rather investigates how LLMs might perceive the political landscape and how this is reflected in their outputs.

Transparency and Reproducibility In the spirit of open science, we have uploaded our code and data to the submission system, and it will be open-sourced upon acceptance. This ensures transparency and allows other researchers to reproduce and build upon our work.

Potential Misuse and Mitigation Strategies We recognize the potential for misuse of our findings, particularly in manipulating LLMs for biased outputs. To mitigate this risk, we emphasize the importance of ethical usage of our research and advocate for continued efforts in developing robust, unbiased AI systems.

Compliance with Ethical Standards Our research adheres to the ethical guidelines and standards set forth by the Association for Computational Linguistics. We have conducted our study with integrity, ensuring that our methods and analyses are ethical and responsible.

Broader Societal Implications We acknowledge the broader implications of our research in the context of AI and society. Our findings contribute to the ongoing discourse on AI ethics, especially regarding the use of AI in sensitive areas like political discourse, influence on views of users and decision-making.

Use of LLMs in the Writing Process Different GPT models, most notably GPT-4, were used to iteratively restructure and reformulate the text to improve readability and remove ambiguity.

References

Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Persistent anti-muslim bias in large language models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. ACM. 1

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta,

Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei Li, YaGuang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniel Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Kelvin Xu, Yunhan Xu, Linting Xue, Pengcheng Yin, Jiahui Yu, Qiao Zhang, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. 2023. Palm 2 technical report. 1

Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. 2021. Stereotyping Norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1004–1015, Online. Association for Computational Linguistics. 1

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc. 1

Sunipa Dev, Emily Sheng, Jieyu Zhao, Aubrie Amstutz, Jiao Sun, Yu Hou, Mattie Sanseverino, Jiin Kim, Akihiro Nishi, Nanyun Peng, and Kai-Wei Chang. 2022. On measures of biases and harms in NLP. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 246–267, Online only. Association for Computational Linguistics. 1

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

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig:. In *Proceedings of the 2019 Conference of the North*. Association for Computational Linguistics. 1

Jochen Hartmann, Jasper Schwenzow, and Maximilian Witte. 2023. The political ideology of conversational AI: Converging evidence on ChatGPT's pro-environmental, left-libertarian orientation. *SSRN Electronic Journal*. 1

Dror Y. Kenett, Tobias Preis, Gitit Gur-Gershgoren, and Eshel Ben-Jacob. 2012. Dependency Network and Node Influence: Application to the study of financial markets. *International Journal of Bifurcation and Chaos*, 22(07):1250181. 2, 5

Katelyn X. Mei, Sonia Fereidooni, and Aylin Caliskan. 2023. Bias against 93 stigmatized groups in masked language models and downstream sentiment classification tasks. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 1

Pranav Narayanan Venkit, Sanjana Gautam, Ruchi Panchanadikar, Ting-Hao Huang, and Shomir Wilson. 2023. Nationality bias in text generation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 116–122, Dubrovnik, Croatia. Association for Computational Linguistics. 1

OpenAI. 2023. Gpt-4 technical report. 1

Davor Petreski and Ibrahim C. Hashim. 2022. Word embeddings are biased. but whose bias are they reflecting? *AI & SOCIETY*, 38(2):975–982. 1

Drago Plecko and Elias Bareinboim. 2022. Causal fairness analysis. 2

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, Ioannis Antonoglou, Rohan Anil, Sebastian Borgeaud, Andrew Dai, Katie Millican, Ethan Dyer, Mia Glaese, Thibault Sottiaux, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, James Molloy, Jilin Chen, Michael Isard, Paul Barham, Tom Hennigan, Ross McIlroy, Melvin Johnson, Johan Schalkwyk, Eli Collins, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, Clemens Meyer, Gregory Thornton, Zhen Yang, Henryk Michalewski, Zaheer Abbas, Nathan Schucher, Ankesh Anand, Richard Ives, James Keeling, Karel Lenc, Salem Haykal, Siamak Shakeri, Pranav Shyam, Aakanksha Chowdhery, Roman Ring, Stephen Spencer, Eren Sezener, Luke Vilnis, Oscar Chang, Nobuyuki Morioka, George Tucker, Ce Zheng, Oliver Woodman, Nithya Attaluri, Tomas Kocisky, Evgenii Eltyshev, Xi Chen, Timothy Chung, Vittorio Selo, Siddhartha Brahma, Petko Georgiev, Ambrose Slone, Zhenkai Zhu, James Lottes, Siyuan Qiao, Ben Caine, Sebastian Riedel, Alex Tomala, Martin Chadwick, Juliette Love, Peter Choy, Sid Mittal, Neil Houlsby, Yunhao Tang, Matthew Lamm, Libin Bai, Qiao Zhang, Luheng He, Yong Cheng, Peter Humphreys, Yujia Li, Sergey Brin, Albin Cassirer, Yingjie Miao, Lukas Zilka, Taylor Tobin, Kelvin Xu, Lev Proleev, Daniel Sohn, Alberto Magni, Lisa Anne Hendricks, Isabel Gao, Santiago Ontañón, Oskar Bunyan, Nathan Byrd, Abhanshu Sharma, Biao Zhang, Mario Pinto, Rishika Sinha, Harsh Mehta, Dawei Jia, Sergi Caelles, Albert Webson, Alex Morris, Becca Roelofs, Yifan Ding, Robin Strudel, Xuehan Xiong, Marvin Ritter, Mostafa Dehghani, Rahma Chaabouni, Abhijit Karmarkar, Guangda Lai, Fabian Mentzer, Bibo Xu, YaGuang Li, Yujing Zhang, Tom Le Paine, Alex Goldin, Behnam Neyshabur, Kate Baumli, Anselm Levskaya, Michael Laskin, Wenhao Jia, Jack W. Rae, Kefan Xiao, Antoine He, Skye Giordano, Lakshman Yagati, Jean-Baptiste Lespiau, Paul Natsev, Sanjay Ganapathy,

Fangyu Liu, Danilo Martins, Nanxin Chen, Yunhan Xu, Megan Barnes, Rhys May, Arpi Vezer, Junhyuk Oh, Ken Franko, Sophie Bridgers, Ruizhe Zhao, Boxi Wu, Basil Mustafa, Sean Sechrist, Emilio Parisotto, Thanumalayan Sankaranarayana Pillai, Chris Larkin, Chenjie Gu, Christina Sorokin, Maxim Krikun, Alexey Guseynov, Jessica Landon, Romina Datta, Alexander Pritzel, Phoebe Thacker, Fan Yang, Kevin Hui, Anja Hauth, Chih-Kuan Yeh, David Barker, Justin Mao-Jones, Sophia Austin, Hannah Sheahan, Parker Schuh, James Svensson, Rohan Jain, Vinay Ramasesh, Anton Briukhov, Da-Woon Chung, Tamara von Glehn, Christina Butterfield, Priya Jhakra, Matthew Wiethoff, Justin Frye, Jordan Grimstad, Beer Changpinyo, Charline Le Lan, Anna Bortsova, Yonghui Wu, Paul Voigtlaender, Tara Sainath, Charlotte Smith, Will Hawkins, Kris Cao, James Besley, Srivatsan Srinivasan, Mark Omernick, Colin Gaffney, Gabriela Surita, Ryan Burnell, Bogdan Damoc, Junwhan Ahn, Andrew Brock, Mantas Pajarskas, Anastasia Petrushkina, Seb Noury, Lorenzo Blanco, Kevin Swersky, Arun Ahuja, Thi Avrahami, Vedant Misra, Raoul de Liedekerke, Mariko Iinuma, Alex Polozov, Sarah York, George van den Driessche, Paul Michel, Justin Chiu, Rory Blevins, Zach Gleicher, Adrià Recasens, Alban Rrustemi, Elena Gribovskaya, Aurko Roy, Wiktor Gworek, Séb Arnold, Lisa Lee, James Lee-Thorp, Marcello Maggioni, Enrique Piqueras, Kartikeya Badola, Sharad Vikram, Lucas Gonzalez, Anirudh Baddepudi, Evan Senter, Jacob Devlin, James Qin, Michael Azzam, Maja Trebacz, Martin Polacek, Kashyap Krishnakumar, Shuo yiin Chang, Matthew Tung, Ivo Penchev, Rishabh Joshi, Kate Olszewska, Carrie Muir, Mateo Wirth, Ale Jakse Hartman, Josh Newlan, Sheleem Kashem, Vijay Bolina, Elahe Dabir, Joost van Amersfoort, Zafarali Ahmed, James Cobon-Kerr, Aishwarya Kamath, Arnar Mar Hrafnkelsson, Le Hou, Ian Mackinnon, Alexandre Frechette, Eric Noland, Xiance Si, Emanuel Taropa, Dong Li, Phil Crone, Anmol Gulati, Sébastien Cevey, Jonas Adler, Ada Ma, David Silver, Simon Tokumine, Richard Powell, Stephan Lee, Michael Chang, Samer Hassan, Diana Mincu, Antoine Yang, Nir Levine, Jenny Brennan, Mingqiu Wang, Sarah Hodkinson, Jeffrey Zhao, Josh Lipschultz, Aedan Pope, Michael B. Chang, Cheng Li, Laurent El Shafey, Michela Paganini, Sholto Douglas, Bernd Bohnet, Fabio Pardo, Seth Odoom, Mihaela Rosca, Cicero Nogueira dos Santos, Kedar Soparkar, Arthur Guez, Tom Hudson, Steven Hansen, Chulayuth Asawaroengchai, Ravi Addanki, Tianhe Yu, Wojciech Stokowiec, Mina Khan, Justin Gilmer, Jaehoon Lee, Carrie Grimes Bostock, Keran Rong, Jonathan Caton, Pedram Pejman, Filip Pavetic, Geoff Brown, Vivek Sharma, Mario Lučić, Rajkumar Samuel, Josip Djolonga, Amol Mandhane, Lars Lowe Sjösund, Elena Buchatskaya, Elspeth White, Natalie Clay, Jiepu Jiang, Hyeontaek Lim, Ross Hemsley, Jane Labanowski, Nicola De Cao, David Steiner, Sayed Hadi Hashemi, Jacob Austin, Anita Gergely, Tim Blyth, Joe Stanton, Kaushik Shivakumar, Aditya Siddhant, Anders Andreassen, Carlos Araya, Nikhil Sethi, Rakesh Shivanna, Steven Hand, Ankur Bapna, Ali Khodaei, Antoine Miech, Garrett Tanzer, Andy Swing, Shantanu Thakoor, Zhufeng Pan, Zachary Nado, Stephanie Winkler, Dian Yu, Mohammad Saleh, Loren Maggiore, Iain Barr, Minh Giang, Thais Kagohara, Ivo Danihelka, Amit Marathe, Vladimir Feinberg, Mohamed Elhawaty, Nimesh Ghelani, Dan Horgan, Helen Miller, Lexi Walker, Richard Tanburn, Mukarram Tariq, Disha Shrivastava, Fei Xia, Chung-Cheng Chiu, Zoe Ashwood, Khuslen Baatarsukh, Sina Samangooei, Fred Alcober, Axel Stjerngren, Paul Komarek, Katerina Tsihlas, Anudhyan Boral, Ramona Comanescu, Jeremy Chen, Ruibo Liu, Dawn Bloxwich, Charlie Chen, Yanhua Sun, Fangxiaoyu Feng, Matthew Mauger, Xerxes Dotiwalla, Vincent Hellendoorn, Michael Sharman, Ivy Zheng, Krishna Haridasan, Gabe Barth-Maron, Craig Swanson, Dominika Rogozińska, Alek Andreev, Paul Kishan Rubenstein, Ruoxin Sang, Dan Hurt, Gamaleldin Elsayed, Renshen Wang, Dave Lacey, Anastasija Ilić, Yao Zhao, Lora Aroyo, Chimezie Iwuanyanwu, Vitaly Nikolaev, Balaji Lakshminarayanan, Sadegh Jazayeri, Raphaël Lopez Kaufman, Mani Varadarajan, Chetan Tekur, Doug Fritz, Misha Khalman, David Reitter, Kingshuk Dasgupta, Shourya Sarcar, Tina Ornduff, Javier Snaider, Fantine Huot, Johnson Jia, Rupert Kemp, Nejc Trdin, Anitha Vijayakumar, Lucy Kim, Christof Angermueller, Li Lao, Tianqi Liu, Haibin Zhang, David Engel, Somer Greene, Anaïs White, Jessica Austin, Lilly Taylor, Shereen Ashraf, Dangyi Liu, Maria Georgaki, Irene Cai, Yana Kulizhskaya, Sonam Goenka, Brennan Saeta, Kiran Vodrahalli, Christian Frank, Dario de Cesare, Brona Robenek, Harry Richardson, Mahmoud Alnahlawi, Christopher Yew, Priya Ponnapalli, Marco Tagliasacchi, Alex Korchemniy, Yelin Kim, Dinghua Li, Bill Rosgen, Zoe Ashwood, Kyle Levin, Jeremy Wiesner, Praseem Banzal, Praveen Srinivasan, Hongkun Yu, Çağlar Ünlü, David Reid, Zora Tung, Daniel Finchelstein, Ravin Kumar, Andre Elisseeff, Jin Huang, Ming Zhang, Rui Zhu, Ricardo Aguilar, Mai Giménez, Jiawei Xia, Olivier Dousse, Willi Gierke, Soheil Hassas Yeganeh, Damion Yates, Komal Jalan, Lu Li, Eri Latorre-Chimoto, Duc Dung Nguyen, Ken Durden, Praveen Kallakuri, Yaxin Liu, Matthew Johnson, Tomy Tsai, Alice Talbert, Jasmine Liu, Alexander Neitz, Chen Elkind, Marco Selvi, Mimi Jasarevic, Livio Baldini Soares, Albert Cui, Pidong Wang, Alek Wenjiao Wang, Xinyu Ye, Krystal Kallarackal, Lucia Loher, Hoi Lam, Josef Broder, Dan Holtmann-Rice, Nina Martin, Bramandia Ramadhana, Daniel Toyama, Mrinal Shukla, Sujoy Basu, Abhi Mohan, Nick Fernando, Noah Fiedel, Kim Paterson, Hui Li, Ankush Garg, Jane Park, DongHyun Choi, Diane Wu, Sankalp Singh, Zhishuai Zhang, Amir Globerson, Lily Yu, John Carpenter, Félix de Chaumont Quitry, Carey Radebaugh, Chu-Cheng Lin, Alex Tudor, Prakash Shroff, Drew Garmon, Dayou Du, Neera Vats, Han Lu, Shariq Iqbal, Alex Yakubovich, Nilesh Tripuraneni, James Manyika, Haroon Qureshi, Nan Hua, Christel Ngani, Maria Abi Raad, Hannah Forbes, Anna Bulanova, Jeff Stanway, Mukund Sundararajan, Victor Ungureanu, Colton Bishop, Yunjie Li, Balaji Venkatraman, Bo Li, Chloe Thornton, Salvatore Scellato, Nishesh Gupta, Yicheng Wang, Ian Tenney, Xihui Wu, Ashish Shenoy, Gabriel Carvajal, Diana Gage Wright, Ben Bariach, Zhuyun Xiao, Peter Hawkins, Sid Dalmia, Clement Farabet, Pedro Valenzuela, Quan Yuan, Chris Welty, Ananth Agar-

wal, Mia Chen, Wooyeol Kim, Brice Hulse, Nandita Dukkipati, Adam Paszke, Andrew Bolt, Elnaz Davoodi, Kiam Choo, Jennifer Beattie, Jennifer Prendki, Harsha Vashisht, Rebeca Santamaria-Fernandez, Luis C. Cobo, Jarek Wilkiewicz, David Madras, Ali Elqursh, Grant Uy, Kevin Ramirez, Matt Harvey, Tyler Liechty, Heiga Zen, Jeff Seibert, Clara Huiyi Hu, Mohamed Elhawaty, Andrey Khorlin, Maigo Le, Asaf Aharoni, Megan Li, Lily Wang, Sandeep Kumar, Alejandro Lince, Norman Casagrande, Jay Hoover, Dalia El Badawy, David Soergel, Denis Vnukov, Matt Miecnikowski, Jiri Simsa, Anna Koop, Praveen Kumar, Thibault Sellam, Daniel Vlasic, Samira Daruki, Nir Shabat, John Zhang, Guolong Su, Jiageng Zhang, Jeremiah Liu, Yi Sun, Evan Palmer, Alireza Ghaffarkhah, Xi Xiong, Victor Cotruta, Michael Fink, Lucas Dixon, Ashwin Sreevatsa, Adrian Goedeckemeyer, Alek Dimitriev, Mohsen Jafari, Remi Crocker, Nicholas FitzGerald, Aviral Kumar, Sanjay Ghemawat, Ivan Philips, Frederick Liu, Yannie Liang, Rachel Sterneck, Alena Repina, Marcus Wu, Laura Knight, Marin Georgiev, Hyo Lee, Harry Askham, Abhishek Chakladar, Annie Louis, Carl Crous, Hardie Cate, Dessie Petrova, Michael Quinn, Denese Owusu-Afriyie, Achintya Singhal, Nan Wei, Solomon Kim, Damien Vincent, Milad Nasr, Christopher A. Choquette-Choo, Reiko Tojo, Shawn Lu, Diego de Las Casas, Yuchung Cheng, Tolga Bolukbasi, Katherine Lee, Saaber Fatehi, Rajagopal Ananthanarayanan, Miteyan Patel, Charbel Kaed, Jing Li, Jakub Sygnowski, Shreyas Rammohan Belle, Zhe Chen, Jaclyn Konzelmann, Siim Põder, Roopal Garg, Vinod Koverkathu, Adam Brown, Chris Dyer, Rosanne Liu, Azade Nova, Jun Xu, Slav Petrov, Demis Hassabis, Koray Kavukcuoglu, Jeffrey Dean, and Oriol Vinyals. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. 1, 2

Adi Robertson. 2024. Google apologizes for 'missing the mark' after gemini generated racially diverse nazis. 2

Salvatore Ruggieri, Jose M. Alvarez, Andrea Pugnana, Laura State, and Franco Turini. 2023. Can we trust fair-ai? *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(13):15421–15430. 2

Juliana Shihadeh, Margareta Ackerman, Ashley Troske, Nicole Lawson, and Edith Gonzalez. 2022. Brilliance bias in GPT-3. In *2022 IEEE Global Humanitarian Technology Conference (GHTC)*. IEEE. 1

Marco R Steenbergen, André Bächtiger, Markus Spörndli, and Jürg Steiner. 2003. Measuring political deliberation: A discourse quality index. *Comparative European Politics*, 1(1):21–48. 3, 4

Zeerak Talat, Aurélie Névéol, Stella Biderman, Miruna Clinciu, Manan Dey, Shayne Longpre, Sasha Luccioni, Maraim Masoud, Margaret Mitchell, Dragomir Radev, Shanya Sharma, Arjun Subramonian, Jaesung Tae, Samson Tan, Deepak Tunuguntla, and Oskar Van Der Wal. 2022. You reap what you sow: On the challenges of bias evaluation under multilingual settings. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 26–41, virtual+Dublin. Association for Computational Linguistics. 1

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. 1

Oskar van der Wal, Dominik Bachmann, Alina Leidinger, Leendert van Maanen, Willem Zuidema, and Katrin Schulz. 2022. Undesirable biases in nlp: Averting a crisis of measurement. 1

Eva Maria Vecchi, Neele Falk, Iman Jundi, and Gabriella Lapesa. 2021. Towards argument mining for social good: A survey. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1338–1352, Online. Association for Computational Linguistics. 3, 4

Pranav Narayanan Venkit, Sanjana Gautam, Ruchi Panchanadikar, Ting-Hao Huang, and Shomir Wilson. 2023. Unmasking nationality bias: A study of human perception of nationalities in AI-generated articles. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. ACM. 1

Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. 2017. Computational argumentation quality assessment in natural language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 176–187, Valencia, Spain. Association for Computational Linguistics. 3, 4

Junzhe Zhang and Elias Bareinboim. 2018. Fairness in decision-making — the causal explanation formula. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press. 2, 3

A Experimental Details

A.1 Input Dataset Statistics

See Table 2.

Table 2: Input Dataset statistics

Statistic	Value
Debates	47
Slices	419
Paragraphs	8,836
Tokens	1,006,127
Words	810,849
Sentences	50,336
Estimated speaking time (175 words per minute (fast))	77 hours

A.2 Cost Breakdown

All queries used the ChatGPT-turbo-0613 over the OpenAI API [4] which costs $0.0015\$/1000$ input tokens and $0.002\$/1000$ output tokens. Here is an overview of the costs done for the final run ($\approx$ another $50\$$ were spent on prototyping, and even some costs in the statistics were used for tests). An overview of the costs can be found in Table 3.

Table 3: Dataset Generation Statistics

Statistic	Value
Queries	81,621
Total Tokens	213,676,479
Input Tokens	212,025,801
Output Tokens	1,650,678
Compared to whole English Wikipedia	% 3.561
Total Cost	$ 321.34
Input Cost	$ 318.04
Output Cost	$ 3.30
Total Words	172,090,392
Input Words	171,502,278
Output Words	588,114
Estimated speaking time (175 words per minute (fast))	16,389 hours

Continued on next page

[4] https://platform.openai.com

Table 3: Dataset Generation Statistics (Continued)

Statistic	Value
Estimated Human Annotation Cost (20 $ / h)	$ 327,791

B Extra Plots

B.1 Pairplots of Attribute Measurement Types

See Figure 9.

B.2 Political Case Studies

See Figures 10 and 11.

C All Attributes

C.1 Given Attributes

Table 4: Defined Variables Description

Name	Description
slice_ id	unique identifier for a slice
debate_ id	unique identifier for debate
slice_ size	the target token size of the slice
debate_ year	the year in which the debate took place
debate_ total_ electoral_ votes	total electoral votes in election
debate_ total_ popular_ votes	total popular votes in election
debate_ elected_ party	party that was elected after debates
speaker	the name of the speaker that is examined in the context of the current slice
speaker_ party	party of the speaker
speaker_ quantitative_ contribution	quantitative contribution in tokens of the speaker to this slice
speaker_ quantitative_ contribution_ ratio	ratio of contribution of speaker to everything that was said

Continued on next page

Table 4: Defined Variables Description (Continued)

Name	Description
speaker_ num_ parts	number of paragraphs the speaker has in current slice
speaker_ avg_ part_ size	average size of paragraph for speaker
speaker_ electoral_ votes	electoral votes that the candidates party scored
speaker_ electoral_ votes_ ratio	ratio of electoral votes that the candidates party scored
speaker_ popular_ votes	popular votes that the candidates party scored
speaker_ popular_ votes_ ratio	ratio of popular votes that the candidates party scored
speaker_ won_ election	flag (0 or 1) that says if speakers party won the election
speaker_ is_ president_ candidate	flag (0 or 1) that says whether the speaker is a presidential candidate
speaker_ is_ vice_ president_ candidate	flag (0 or 1) that says whether the speaker is a vice presidential candidate
speaker_ is_ candidate	flag (0 or 1) that says whether the speaker is a presidential or vice presidential candidate

C.2 Measured Attributes

C.2.1 Slice Dependent Attributes

Table 5: Slice Variables

Group, Name	Description
content quality	float
filler	Is there any content in this part of the debate or is it mostly filler?

Continued on next page

Table 5: Slice Variables (Continued)

Group, Name	Description
speaker	Is there any valuable content in this part of the debate that can be used for further analysis of how well the speakers can argue their points?
dataset	We want to create a dataset to study how well the speakers can argue, convery information and what leads to winning an election. Should this part of the debate be included in the dataset?
topic predictiveness	float
usefullness	Can this part of the debate be used to predict the topic of the debate?
topic	str
max3	Which topic is being discussed in this part of the debate? Respond with a short, compact and general title with max 3 words in all caps.

C.2.2 Speaker Dependent Attributes

SET \IFADDTABLESTRUE TO RENDER THESE (INCREASES COMPILE TIME)

D Prompt Examples

For better readability, the slice has been removed and replaced with {slice_text} in the query. Note that we are aware of the imperfection in the query regarding the missing quote around the name of the observable for some queries in the JSON template, and it has been fixed for later studies.

D.1 Single Speaker Prompt Example

D.1.1 Query

```
You are a helpfull assistant
    tasked with completing
    information about part of a
    political debate. Here is the
    text you are working with:

---
```

```
{slice_text}

---

Your task is to complete
   information about the speaker
   PEROT based on the text above.

All scores are between 0.0 and
   1.0!
1.0 means that the quality of
   interest can't be stronger,
   0.0 stands for a complete
   absence and 0.5 for how an
   average person in an average
   situation would be scored.
Strings are in ALL CAPS and
   without any additional
   information. If you are unsure
    about a string value, write '
   UNCLEAR'.
Make sure that the response is a
   valid json object and that the
    keys are exactly as specified
    in the template!
Don't add any additional and
   unnecessary information or
   filler text!
Give your response as a json
   object with the following
   structure:

{
 tone is academic: <float Does
    the speaker use an academic
    tone?>
}

Now give your response as a
   complete, finished and correct
    json and don't write anything
    else:
```

D.1.2 Response

```
{
 "tone is academic": 0.2
}
```

D.2 Multiple Speakers Prompt Example

D.2.1 Query

```
You are a helpfull assistant
   tasked with completing
   information about part of a
   political debate. Here is the
   text you are working with:

---

{slice_text}

---

Your task is to complete
   information about the speakers
    based on the text above.

Here are the speakers:
['GERALD FORD', 'MAYNARD', 'JIMMY
    CARTER', 'KRAFT', 'WALTERS']
Don't leave any out or add
   additional ones!

All scores are between 0.0 and
   1.0!
1.0 means that the quality of
   interest can't be stronger,
   0.0 stands for a complete
   absence and 0.5 for how an
   average person in an average
   situation would be scored.
Strings are in ALL CAPS and
   without any additional
   information. If you are unsure
    about a string value, write '
   UNCLEAR'.
Make sure that the response is a
   valid json object and that the
    keys are exactly as specified
    in the template!
Don't add any additional and
   unnecessary information or
   filler text!
Give your response as a json
   object with the following
   structure:

{
 <str speaker>: {
  "preparation": <float Does the
     speaker seem well-prepared
     for the debate,
```

```
      demonstrating a good
      understanding of the topics
      and questions at hand?>
  },
  ...
}

Now give your response as a
    complete, finished and correct
     json including each speaker
    and don't write anything else:
```

D.2.2 Response

```
{
  "GERALD FORD": {
    "preparation": 1.0
  },
  "MAYNARD": {
    "preparation": 0.5
  },
  "JIMMY CARTER": {
    "preparation": 1.0
  },
  "KRAFT": {
    "preparation": 0.5
  },
  "WALTERS": {
    "preparation": 1.0
  }
}
```

E Example Slice with 2500 tokens

SCHIEFFER: I'm going to add a couple of minutes here to give you a chance to respond.

MITT ROMNEY: Well, of course I don't concur with what the president said about my own record and the things that I've said. They don't happen to be accurate. But — but I can say this, that we're talking about the Middle East and how to help the Middle East reject the kind of terrorism we're seeing, and the rising tide of tumult and — and confusion. And — and attacking me is not an agenda. Attacking me is not talking about how we're going to deal with the challenges that exist in the Middle East, and take advantage of the opportunity there, and stem the tide of this violence.

But I'll respond to a couple of things that you mentioned. First of all, Russia I indicated is a geopolitical foe. Not...

(CROSSTALK)

MITT ROMNEY: Excuse me. It's a geopolitical foe, and I said in the same — in the same paragraph I said, and Iran is the greatest national security threat we face. Russia does continue to battle us in the U.N. time and time again. I have clear eyes on this. I'm not going to wear rose-colored glasses when it comes to Russia, or Putin. And I'm certainly not going to say to him, I'll give you more flexibility after the election. After the election, he'll get more backbone. Number two, with regards to Iraq, you and I agreed I believe that there should be a status of forces agreement.

(CROSSTALK)

MITT ROMNEY: Oh you didn't? You didn't want a status of...

BARACK OBAMA: What I would not have had done was left 10,000 troops in Iraq that would tie us down. And that certainly would not help us in the Middle East.

MITT ROMNEY: I'm sorry, you actually — there was a — there was an effort on the part of the president to have a status of forces agreement, and I concurred in that, and said that we should have some number of troops that stayed on. That was something I concurred with...

(CROSSTALK)

BARACK OBAMA: Governor...

(CROSSTALK)

MITT ROMNEY: ...that your posture. That was my posture as well. You thought it should have been 5,000 troops...

(CROSSTALK)

BARACK OBAMA: Governor?

MITT ROMNEY: ... I thought there should have been more troops, but you know what? The answer was we got...

(CROSSTALK)

MITT ROMNEY: ... no troops through whatsoever.

BARACK OBAMA: This was just a few weeks ago that you indicated that we should still have troops in Iraq.

MITT ROMNEY: No, I...

(CROSSTALK)

MITT ROMNEY: ...I'm sorry that's a...

(CROSSTALK)

BARACK OBAMA: You — you...

MITT ROMNEY: ...that's a — I indicated...

(CROSSTALK)

BARACK OBAMA: ...major speech.

(CROSSTALK)

MITT ROMNEY: ...I indicated that you failed to put in place a status...

(CROSSTALK)

BARACK OBAMA: Governor?

(CROSSTALK)

MITT ROMNEY: ...of forces agreement at the end of the conflict that existed.

BARACK OBAMA: Governor — here — here's — here's one thing...

(CROSSTALK)

BARACK OBAMA: ...here's one thing I've learned as commander in chief.

(CROSSTALK)

SCHIEFFER: Let him answer...

BARACK OBAMA: You've got to be clear, both to our allies and our enemies, about where you stand and what you mean. You just gave a speech a few weeks ago in which you said we should still have troops in Iraq. That is not a recipe for making sure that we are taking advantage of the opportunities and meeting the challenges of the Middle East.

Now, it is absolutely true that we cannot just meet these challenges militarily. And so what I've done throughout my presidency and will continue to do is, number one, make sure that these countries are supporting our counterterrorism efforts.

Number two, make sure that they are standing by our interests in Israel's security, because it is a true friend and our greatest ally in the region.

Number three, we do have to make sure that we're protecting religious minorities and women because these countries can't develop unless all the population, not just half of it, is developing.

Number four, we do have to develop their economic — their economic capabilities.

But number five, the other thing that we have to do is recognize that we can't continue to do nation building in these regions. Part of American leadership is making sure that we're doing nation building here at home. That will help us maintain the kind of American leadership that we need.

SCHIEFFER: Let me interject the second topic question in this segment about the Middle East and so on, and that is, you both mentioned — alluded to this, and that is Syria.

The war in Syria has now spilled over into Lebanon. We have, what, more than 100 people that were killed there in a bomb. There were demonstrations there, eight people dead.

President, it's been more than a year since you saw — you told Assad he had to go. Since then, 30,000 Syrians have died. We've had 300,000 refugees.

The war goes on. He's still there. Should we reassess our policy and see if we can find a better way to influence events there? Or is that even possible?

And you go first, sir.

BARACK OBAMA: What we've done is organize the international community, saying Assad has to go. We've mobilized sanctions against that government. We have made sure that they are isolated. We have provided humanitarian assistance and we are helping the opposition organize, and we're particularly interested in making sure that we're mobilizing the moderate forces inside of Syria.

But ultimately, Syrians are going to have to determine their own future. And so everything we're doing, we're doing in consultation with our partners in the region, including Israel which obviously has a huge interest in seeing what happens in Syria; coordinating with Turkey and other countries in the region that have a great interest in this.

This — what we're seeing taking place in Syria is heartbreaking, and that's why we are going to do everything we can to make sure that we are helping the opposition. But we also have to recognize that, you know, for us to get more entangled militarily in Syria is a serious step, and we have to do so

making absolutely certain that we know who we are helping; that we're not putting arms in the hands of folks who eventually could turn them against us or allies in the region.

And I am confident that Assad's days are numbered. But what we can't do is to simply suggest that, as Governor Romney at times has suggested, that giving heavy weapons, for example, to the Syrian opposition is a simple proposition that would lead us to be safer over the long term.

SCHIEFFER: Governor?

MITT ROMNEY: Well, let's step back and talk about what's happening in Syria and how important it is. First of all, 30,000 people being killed by their government is a humanitarian disaster. Secondly, Syria is an opportunity for us because Syria plays an important role in the Middle East, particularly right now.

MITT ROMNEY: Syria is Iran's only ally in the Arab world. It's their route to the sea. It's the route for them to arm Hezbollah in Lebanon, which threatens, of course, our ally, Israel. And so seeing Syria remove Assad is a very high priority for us. Number two, seeing a — a replacement government being responsible people is critical for us. And finally, we don't want to have military involvement there. We don't want to get drawn into a military conflict.

And so the right course for us, is working through our partners and with our own resources, to identify responsible parties within Syria, organize them, bring them together in a — in a form of — if not government, a form of — of — of council that can take the lead in Syria. And then make sure they have the arms necessary to defend themselves. We do need to make sure that they don't have arms that get into the — the wrong hands. Those arms could be used to hurt us down the road. We need to make sure as well that we coordinate this effort with our allies, and particularly with — with Israel.

But the Saudi's and the Qatari, and — and the Turks are all very concerned about this. They're willing to work with us. We need to have a very effective leadership effort in Syria, making sure that the — the insurgent there are armed and that the insurgents that become armed, are people who will be the responsible parties. Recognize — I believe that Assad must go. I believe he will go. But I believe — we want to make sure that we have the relationships of friendship with the people that take his place, steps that in the years to come we see Syria as a — as a friend, and Syria as a responsible party in the Middle East.

This — this is a critical opportunity for America. And what I'm afraid of is we've watched over the past year or so, first the president saying, well we'll let the U.N. deal with it. And Assad — excuse me, Kofi Annan came in and said we're going to try to have a ceasefire. That didn't work. Then it went to the Russians and said, let's see if you can do something. We should be playing the leadership role there, not on the ground with military.

SCHIEFFER: All right.

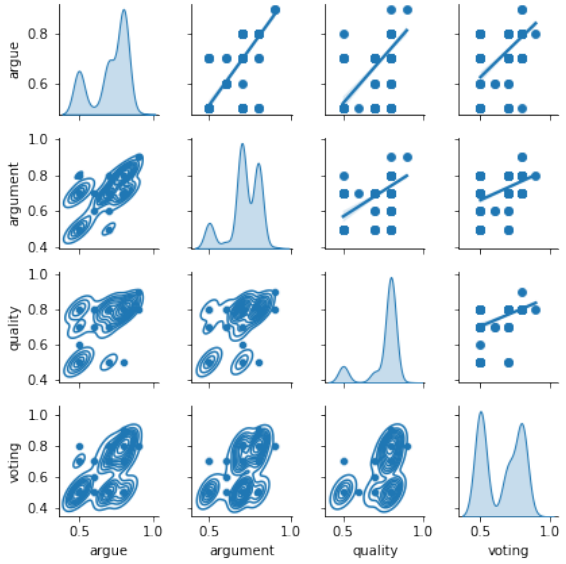
MITT ROMNEY: . . . by the leadership role.

BARACK OBAMA: We are playing the leadership role. We organized the Friends of Syria. We are mobilizing humanitarian support, and support for the opposition. And we are making sure that those we help are those who will be friends of ours in the long term and friends of our allies in the region over the long term. But going back to Libya — because this is an example of how we make choices. When we went in to Libya, and we were able to immediately stop the massacre there, because of the unique circumstances and the coalition that we had helped to organize. We also had to make sure that Moammar Gadhafi didn't stay there.

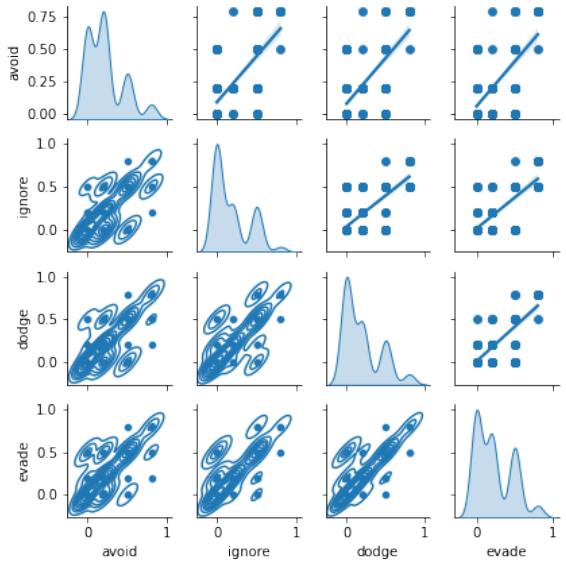
And to the governor's credit, you supported us going into Libya and the coalition that we organized. But when it came time to making sure that Gadhafi did not stay in power, that he was captured, Governor, your suggestion was that this was mission creep, that this was mission muddle.

Imagine if we had pulled out at that point. You know, Moammar Gadhafi had more American blood on his hands than any individual other than Osama bin Laden. And so we were going to make sure that we finished the job. That's part of the reason why the Libyans stand with us.

But we did so in a careful, thoughtful way, making certain that we knew who we were dealing with, that those forces of moderation on the ground were ones that we could work with, and we have to take the same kind of steady, thoughtful leadership when it comes to Syria. That ...



(a) Pairplot for *Score*



(b) Pairplot for *Evasiveness*

Figure 9: Internal Differences of Attribute Measurement Types: We see that similar definitions of *Evasiveness* lead to very comparable results and similar distributions. But *Score (voting)* stands out as a very different definition. This makes sense as its definition asks about the chances of winning the election, while the others refer to the quality of the argument. The exact definitions of the attributes can be found in Appendix C.2.

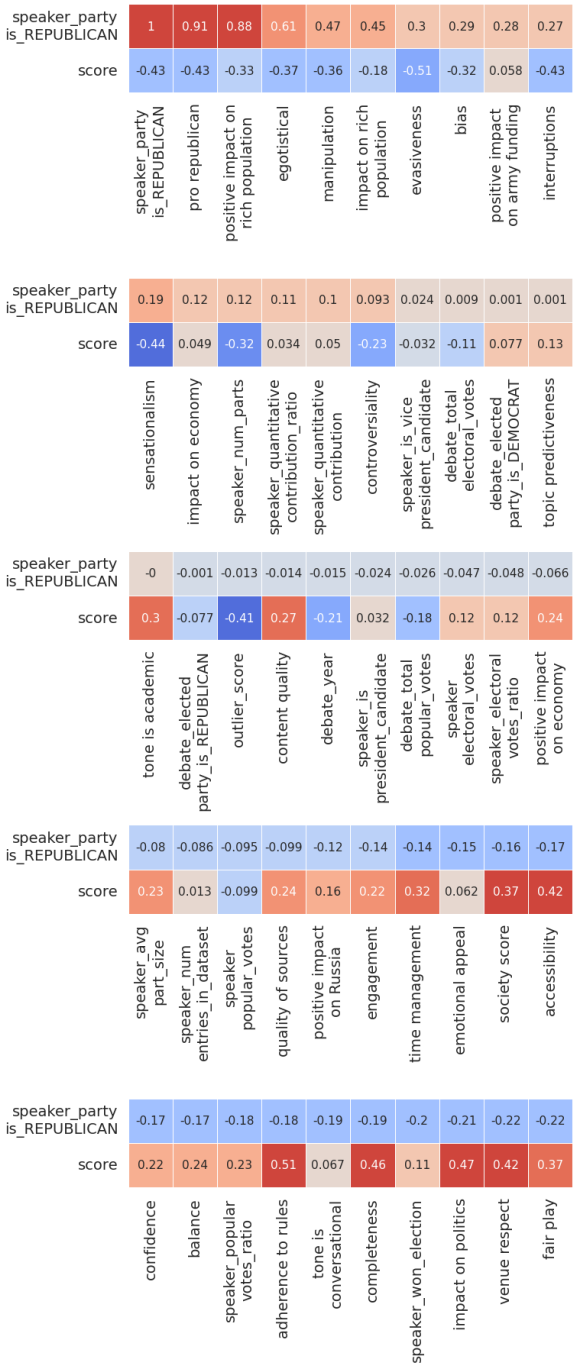


Figure 10: First Half of *Score* and *Speaker Party* vs. All other Attributes

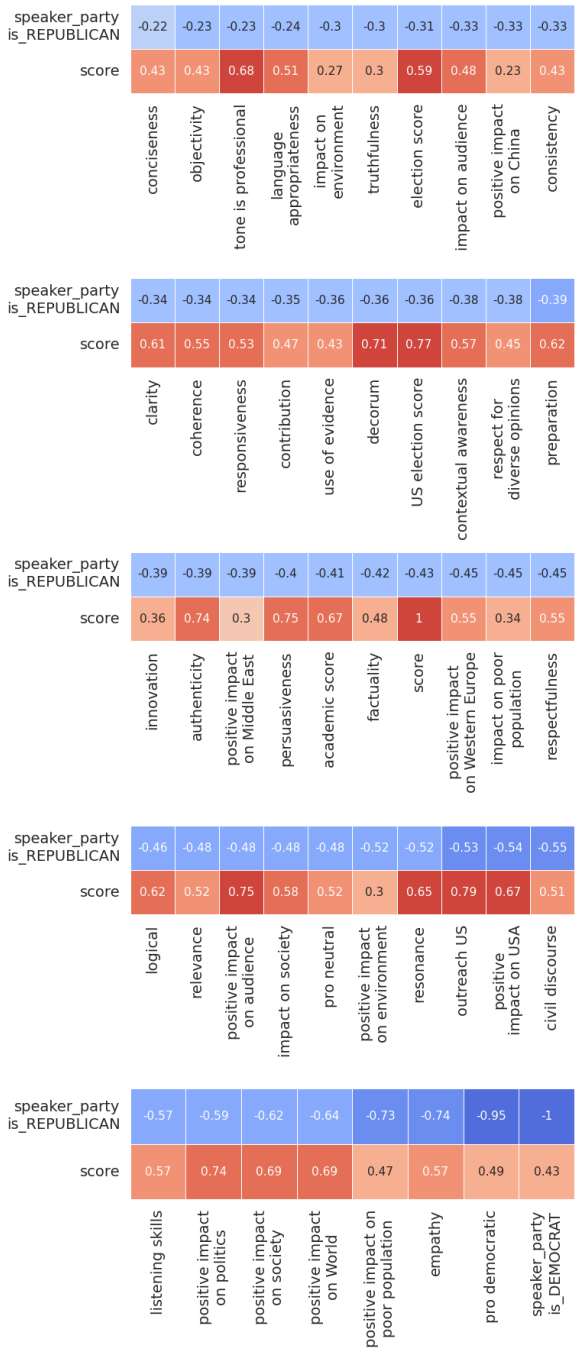


Figure 11: Second Half of *Score* and *Speaker Party* vs. All other Attributes