

It’s not only What You Say, It’s also Who It’s Said to: Counterfactual Analysis of Interactive Behavior in the Courtroom

Biaoyan Fang[†] Trevor Cohn^{†*} Timothy Baldwin^{†:♠} Lea Frermann[†]

[†]The University of Melbourne, Australia

♠MBZUAI

{biaoyan, tcohn, tbaldwin, lfrermann}@unimelb.edu.au

Abstract

To what extent do personal attributes affect the way we are spoken to? Answering this question requires the precise reproduction of a conversational context except for one personal attribute of interest, amounting to a classical, yet infeasible, causal inference problem. We present a method based on counterfactual analysis by manipulating speaker attributes in observational data. We propose a case study of Advocate responses to Justices in debates in the Supreme Court of the United States. Specifically, we measure changes in politeness and coordination of Advocates when responding to (a) real Justices and (b) counterfactually-manipulated Justices, with responses generated with GPT2. We first validate our method, showing that GPT2-generated outputs capture coordination and politeness. Our results confirm a known impact of the attribute *gender*, and suggest a weaker effect of *seniority* on coordination.¹

1 Introduction

Does the way we speak to others depend on the personal attributes of the addressee? Speakers employ different strategies when replying to persons of different social status (Niederhoffer and Pennebaker, 2002; Taylor and Thomas, 2008; Danescu-Niculescu-Mizil et al., 2012, 2013; Mizukami et al., 2016). Well-known strategies are linguistic accommodation, e.g., in adjusting in style to a more senior conversation partner (Kulesza et al., 2014; Danescu-Niculescu-Mizil et al., 2012; Noble and Fernández, 2015; Xu et al., 2018), and politeness, where speakers vary the level of formality and word choice (Danescu-Niculescu-Mizil et al., 2013; Fu et al., 2020; Li et al., 2020). A better understanding of the factors that cause such strategies is not only of interest to social scientists, but could also render dialogue systems more natural.

*Now at Google DeepMind.

¹Dataset and code are made available at <https://github.com/biaoyanf/SCOTUS-counterfactual>.

[ADVRES] [...] This is an obstacle preemption case masquerading as a field preemption case. [...] and the Nuclear Regulatory Commission have repeatedly reaffirmed that states have the ability to regulate mining up to and including by banning it altogether [...].

[KAVANAUGH] Yeah. So the mining and milling occur together, correct? In other words [...] you don’t have mining without milling; you don’t have milling without mining

[ADVRES] [...] there’s a way – the in situ leaching process, they literally occur at the same time.

Figure 1: An example instance from our dataset. We measure politeness and coordination in the true or generated [ADVRES] response (bottom) to a question from a Justice (middle) in context (top). Counterfactuals replace the **true** Justice identity tag (**male**) with a **different** one (**female**).

To directly address this question would require a dataset of paired situations which are identical except for one speaker’s personal attribute of interest. This would allow us to measure the causal effect of the attribute value (treatment) on language style (outcome) (Pearl, 2009; Imbens and Rubin, 2015).

Since this is not achievable, we present a counterfactual methodology based on manipulation of observational data and the power of pre-trained language models. Specifically, we require a dataset of observed conversations involving persons for which: (a) personal attribute values are known; (b) sufficient conversational data is available to fine-tune LMs; and (c) these conversations occur in a relatively controlled context. Here, we use conversations between Advocates and Justices in arguments from the US Supreme Court (SCOTUS) and investigate differences in politeness (Danescu-Niculescu-Mizil et al., 2013) and coordination (Danescu-Niculescu-Mizil et al., 2012) of Advocate responses to questions asked by Justices with different personal and professional attributes (Figure 1).

To validate our approach, we first show that pre-trained language models reliably capture politeness and coordination (Section 3). We do so

by comparing the politeness and coordination levels of real Advocate answers in the SCOTUS data against answers generated by a large pre-trained language model (GPT2; Radford et al. (2019)) when prompted with the same context (Figure 1, blue).

We consider three binary Justice attributes — gender, nominating party, and seniority — and compare politeness and coordination in GPT2 generated responses in factual vs. counterfactually-manipulated situations where the inquiring Justices’ attribute value is flipped (Section 4; Figure 1, red → blue). We ask *would an Advocate’s response change in politeness (coordination) had the question been asked by a Justice from a different social group?*

We show that LMs are sensitive to social cues, beyond the inherent biases in LMs and data sets (Section 6), prompting further research on fairness and bias from a sociolinguistic perspective.

2 Methodology

Data We use a subset of a published set of SCOTUS arguments² (Chang et al., 2020), covering transcripts from 1955 to 2019, comprising 4.5K cases, 35 Justices, and >1M utterances. We randomly selected one case per year to include in the dev and test sets, respectively, and used the remainder to fine-tune GPT2. We removed Justices with <80 turns from the test set.³

The published SCOTUS data includes the speaker identity (name) and function (e.g., Justice or Advocate) for each turn. We retain the full Justice name and map the Advocates to the side they stand for, i.e. petitioner ([ADVPET]) or respondent ([ADVRES]).⁴ From this, we construct our final dataset, where each instance consists of 400 words of preceding context, followed by a Justice turn directly followed by an Advocate’s response. Figure 1 shows an example context. The statistics of our dataset are listed in Table 1 and Appendix A provides additional details on the data set.

Demographic and Personal Attributes We study the impact of three binary attributes of a Justice on the politeness and coordination levels of the

²<https://convokit.cornell.edu/documentation/supreme.html>

³See Appendix A for data construction details.

⁴Note that while the distinction is not strictly necessary for the experiments, it is intuitive that social alignment is impacted by the position of the conversation partner. See Appendix C for a detailed analysis of advocates’ attributes.

	Train	Dev	Test
Number of Instances	402,230	6,366	6,129
Number of Cases	4,445	65	65
Number of Justices	35	35	20
Total #Turns	981,492	15,653	14,901
Avg. #token/Turn	48.38	45.33	48.11
Avg. #token/Justice Turn	29.93	29.64	29.49
Avg. #token/Advoc. Turn	64.93	60.05	63.56

Table 1: Statistics of our SCOTUS dataset, where individual instances are comprised of 400-token preceding context + Justice Turn + Advocate Turn.

Advocates’ response to their questions: gender (m, f); seniority (Chief Justice or not); and the party which nominated the Justice (Democrat, Republican).

Coordination and Politeness Prior work has proposed measures of coordination (Danescu-Niculescu-Mizil et al., 2012) and politeness (Danescu-Niculescu-Mizil et al., 2013), which have been used to study the connection between linguistic choices and personal attributes. We consider those measures to analyze the behavior of Advocates in the courtroom. Particularity, we directly use the coordination indicator from ConvoKit,⁵ which quantifies the coordination of a respondent *adv* to a speaker *jst* wrt. a linguistic marker⁶ *lm* by calculating how much the fact that *jst* used *lm* increases the probability of *adv* using *lm* in a direct response. We average across markers, and instead of considering individual *jst*, we measure coordination to groups of *jst* who share an attribute value *a* (e.g., all male or female Justices), obtaining coordination measure $\hat{y}_c^{jst=a}$.

We similarly measure politeness as the probability of observing a politeness marker⁷ *pm* in an utterance by *adv* in response to *jst*, normalized by the prior probability of *adv* using *pm*; again, we average over values *pm* and all *jst* that share an attribute value, obtaining a politeness measure $\hat{y}_p^{jst=a}$.

Method We used GPT2 to generate responses given factual (Figure 1, red) and counterfactual (Figure 1, blue) contexts. We fine-tuned GPT2 on the SCOTUS training set for 50 epochs to incorporate a notion of typical Advocates’ response

⁵<https://convokit.cornell.edu/>

⁶E.g., auxiliaries, conjunctions, or quantifiers; cf., Danescu-Niculescu-Mizil et al. (2012) for full list.

⁷E.g., greetings, apologies or hedges; cf. Danescu-Niculescu-Mizil et al. (2013) for the full list.

behavior, and selected the best model based on dev set perplexity. We generated Advocate responses with GPT2 given a Justice question with preceding context. Each turn includes a speaker tag (see [TAGs] in Figure 1). We chose GPT2 because, unlike its successors GPT3 and ChatGPT, it can be fine-tuned to a target domain.

Factuals We first present factual SCOTUS contexts cxt (Figure 1, top and middle context) with true speaker tags (Figure 1, red) as prompts to GPT2 and have it generate a factual response. We obtain a final coordination (or, equivalently, politeness) score, by averaging over all generated responses to questions by Justices with attribute a :

$$y^{jst=a}|\{cxt\}^a, jst=a, \quad (1)$$

where scores y can pertain to coordination or politeness, and we omit subscripts to avoid clutter. In Section 3 we verify that GPT2 scores $y^{jst=a}$ mirror the true values $\hat{y}^{jst=a}$ across attribute values, for both coordination and politeness.

Counterfactual Speakers We next manipulate *only* a Justices’ attribute of interest while keeping the rest of the context fixed, to examine its effect on the Advocate’s response. We do so by changing the speaker tag (Figure 1, red \rightarrow blue). For instance, to test the effect of gender on politeness (or, equivalently, coordination), we take all contexts involving male Justices $\{cxt\}^m$ and replace the male Justice’s indicator (e.g., [KAVANAUGH]) with a female Justice’s name (e.g., [BADER-GINSBURG]). We do this exhaustively for all male-female combinations and average over counterfactually generated responses towards female Justice tags to obtain a ‘generic female’ \hat{f} politeness (coordination) score $y^{jst=\hat{f}}|^m$ under male context $\{cxt\}^m$. Equivalently, we obtain a ‘generic male’ \hat{m} politeness (coordination) score $y^{jst=\hat{m}}|^m$ under male context $\{cxt\}^m$ by substituting male indicators with other male tags:

$$\begin{aligned} & y^{jst=\hat{f}}|^m|\{cxt\}^m, do(jst=m \rightarrow \hat{f}), \\ & y^{jst=\hat{m}}|^m|\{cxt\}^m, do(jst=m \rightarrow \hat{m}), \end{aligned} \quad (2)$$

where we denote the counterfactual manipulation with the *do* operator from the causal inference literature (Pearl, 2009). In sum, we fix the context (e.g., $\{cxt\}^m$) but manipulate properties of the Justices (e.g., \hat{m} vs. \hat{f}) and measure the difference between politeness (or, equivalently, coordination) scores to

Preference	Factual (Exp 1)			C’factual (Exp 2)		
	Rel	Info	Cons	Rel	Info	Cons
None	0.88	0.53	0.71	0.77	0.43	0.52
Real	0.04	0.37	0.14	0.07	0.23	0.16
Generated	0.08	0.10	0.14	0.16	0.34	0.32

Table 2: Human preferences of Relevance (Rel), Informativeness (Info) and Consistency (Cons).

test the effect of attribute values in speakers.⁸

Counterfactual Contexts Alternatively, to understand the effect of context spoken by groups with different values, we hold the properties of the Justices constant, and change the conditioning contexts. E.g., we fix the gender of the asking Justice, e.g., to a ‘generic male’ \hat{m} by exhaustively inserting male Justice names as described above, and generate responses when (i) contextualized with male Justices’ contexts $\{cxt\}^m$, or (ii) female Justices’ contexts $\{cxt\}^f$, and compare their differences:

$$\begin{aligned} & y^{jst=\hat{m}}|^m|\{cxt\}^m, do(jst=m \rightarrow \hat{m}), \\ & y^{jst=\hat{m}}|^f|\{cxt\}^f, do(jst=f \rightarrow \hat{m}). \end{aligned} \quad (3)$$

In sum, we fix the properties of the Justices (e.g., \hat{m}) but manipulate the context from different justices (e.g., $\{cxt\}^m$ vs. $\{cxt\}^f$) and measure the difference between politeness (coordination) scores to test the effect of spoken context.

In Section 4 we apply both counterfactual comparisons (Equations (2) and (3)) across our three binary attributes, and politeness and coordination measures.

3 Experiment 1: Factual generation

First, we validated that generated Advocate responses to Justice questions resemble true responses in both content and coordination/politeness. We compared scores as estimated from the SCOTUS data against those generated by GPT2 when prompted with factual contexts (Equation (1)).

Content validity We observed a decrease in dev perplexity (116.22 to 2.75) after fine-tuning GPT-2. Additionally, we manually evaluated the generated responses in terms of their informativeness, relevance, and consistency (Finch and Choi, 2020). We hired two English native-speaker social scientists not involved in the project and presented them

⁸Analogously we obtain $y^{jst=\hat{m}}|^f$ and $y^{jst=\hat{f}}|^f$ by manipulating Justice groups in female contexts $\{cxt\}^f$.

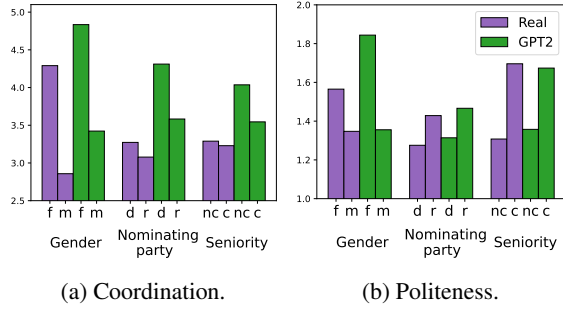


Figure 2: Coordination and politeness scores in Advocates’ responses to factual Justices. Gender: (m=Male, f=Female); Nominating party: (d=Democrat, r=Republican); Seniority (c=Chief jst, nc=Non-Chief). The coordination and politeness scores are scaled by 100 and 10 for clarity, respectively, and the same scaling applies to the following Figures.

with the same prompt as GPT2 and the real and generated answers, and asked to indicate for each category which answer is preferred or whether both are equal. Table 2 (left) shows that overwhelmingly both answers were rated as equal. For cases where one answer was preferred, we verified that the distribution did not differ significantly from random (Binomial test, $p>0.05$, $n=46$).

Coordination and Politeness Do GPT2 responses reflect the differences in coordination and politeness in responses observed in the original SCOTUS data? Figure 2 shows that predicted coordination scores (green bars) consistently align with true scores (purple bars) in terms of direction and magnitude; the same holds for politeness in Figure 2b. Gender produces the largest difference in coordination (Figure 2a left), echoing the finding of Danescu-Niculescu-Mizil et al. (2012) that gender impacts speaker coordination in legal discourse. Gender and seniority incur politeness differences.⁹

4 Experiment 2: Counterfactual Generation

Having demonstrated that GPT2-generated responses reflect content and group-specific social signaling (Section 3), we now apply the counterfactual framework (Equations (2) and (3)). First, we ensured the content validity of *counterfactually* generated utterances via human evaluation as in Section 3. Table 2 (right) confirms that annotators did not prefer real responses over generated ones

⁹Appendix C presents further evidence for this, over 7 personal attributes across Advocates and Justices.

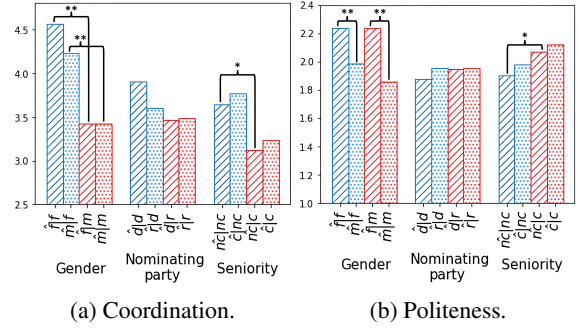


Figure 3: Coordination and politeness scores in Advocate responses to counterfactually-manipulated Justices. x -labels as in Figure 2. Colors indicate attribute-specific contexts. Shading indicates Justice attribute value. Pairwise significant differences according to Welch’s t-test at $p<0.1$ (*) and $p<0.05$ (**).

(or vice versa) for the vast majority of instances.¹⁰

We first ask to what extent social cues in an Advocate response vary wrt. a manipulated attribute of the asking party (Justice). This corresponds to Equation (2) and comparison of *lined vs. dotted bars of the same color* in Figure 3a (coordination) and Figure 3b (politeness). We find a significant difference (Welch’s t-test, $p<0.05$) in politeness in response to Justice gender: Advocates are significantly more polite to real and counterfactual female Justices given the same context. This holds in both directions (a significant increase when manipulating ‘generic male’ $\hat{m} \rightarrow \hat{f}$ (red) and decrease when turning ‘generic female’ $\hat{f} \rightarrow \hat{m}$ (blue)).

Next, we investigate the impact of context on social signals in a response, by keeping properties of Justices fixed, but embedding them in different contexts originating either from Justices with binary attribute label a or $\neg a$ (Equation (3)). We inspect the results by *comparing the blue vs. red bars of the same shade (both lined or both dotted)* in Figure 3. We find a significant difference in coordination, when embedding female Justices in female vs. male contexts (left block Figure 3a blue lined vs. red lined) and male Justices in female vs. male contexts (same block, blue dotted vs. red dotted). With marginal significance ($p<0.1$, $n=16$), we observe a decrease in coordination and increase in politeness when embedding non-Chief Justices in non-Chief vs. Chief contexts (blue vs. red lines in the right blocks in Figures 3a and 3b).

¹⁰Example pairs of real and generated responses are in Appendix D. Preferences are again not significantly different from random (Binomial, $p>0.05$).

5 Discussion

Does the way we speak depend on the personal attributes of our interlocutor? We addressed this question with a new methodology involving counterfactual manipulation and high-quality response generation from a powerful language model. Our method isolates social cues in responses to questions from groups of people with different social or demographic attributes in otherwise controlled contexts. We presented a case study on Advocate responses to Justices in SCOTUS arguments, where we manipulated the Justices’ attributes of gender, seniority, and nominating political party.

We found that (1) the Justice’s gender has a significant impact on Advocate *politeness*: Advocates are more polite toward female Justices; (2) Advocate *coordination* changes significantly in response to the *context* invoked by male vs. female Justices. We also found weak evidence that Advocate coordination and politeness change in response to the *context* invoked by Chief vs. non-Chief Justices.

Our method extends a line of work of leveraging LMs for causal inference with observational text data (Keith et al., 2020; Veitch et al., 2020; Pryzant et al., 2021; Feder et al., 2022), which has predominantly studied the causal effects of different linguistic properties (treatment) on some non-textual outcome, whereas we manipulate speaker attributes (treatment) and study their effect on language as an outcome. We address the problem of confounding (variables that affect both treatment and outcome) by: (a) tight control of contexts as legal court arguments; and (b) verification that our method of GPT2 fine-tuning and attribute-conditioned generation is both faithful in content to the original, and captures real-world differences in expression pertaining to our attributes of interest.

Our proposed counterfactual framework can be further applied to other conversations such as online discussions (Danescu-Niculescu-Mizil et al., 2012), interviews (Fu et al., 2016) and congressional records (Gentzkow et al., 2018), to investigate the effect of attribute groups in various scenarios; and to other sociolinguistic phenomena, such as persuasion (Dimitrov et al., 2021) in order to explore additional perspectives on the social dynamics in conversations.

While current debiasing research (Wang et al., 2020; Kaneko and Bollegala, 2021; Guo et al., 2022) is focused on the semantic level (i.e. unwarranted associations between attributes and content),

our study shows that LMs also capture subtle sociolinguistic biases (i.e. coordination and politeness). We hope our counterfactual framework spurs future work on evaluating fairness and bias from a sociolinguistic perspective.

6 Ethical Discussion

6.1 Potential Bias

We conduct our counterfactual analysis using LMs, which have been proven to be biased (Nadeem et al., 2021; Delobelle et al., 2022). The generated results might contain inherited bias from the pre-trained language models. This could introduce confounds in the generated responses: (1) LMs might be a priori more likely to (not) generate selective politeness/coordination markers due to exposure in the training data.¹¹ (2) Historical biases — as LMs were predominantly trained on contemporary language leading to a temporal confound where predictions on earlier data points might be noisier than those for more recent SCOTUS discussions. We acknowledge that this could further bring undetected harm in analyzing Advocate court behaviors, especially in the counterfactual setting.

Our analysis is based on groups of speakers who share an attribute value, and makes no claims about the content/professional level, nor on individuals. Although our framework could generate intermediate results for individual Justices and, in theory, could derive personal conclusions in court, we strongly caution against this.

6.2 Human annotations

To test the content validity of GPT2, we recruited two native-speakers with social science backgrounds not involved in this project through the authors’ contacts. We did not record any personal information (e.g. demographics). Authors were debriefed in full after completion of the task, and paid an hourly rate of USD\$38, which far exceeds the local minimum pay rate.

7 Limitations

For modeling speaker behavior, we consider social cues based on coordination and politeness. Other dimensions, such as sentiment (Feldman, 2013), persuasion (Dimitrov et al., 2021), or rate of interruption (Epstein et al., 2010), could be further ex-

¹¹We investigate coordination scores under breakdown markers in Appendix E, with Advocates towards male vs. female Justices as an example.

explored to better understand social dynamics in the courtroom. It would also be interesting to explore the interactions among the proposed sociolinguistic perspectives. Also, we focus on the responses from Advocates toward Justices. A study from the other direction could complement this work, and aid in better understanding the dynamics of legal oral arguments. What’s more, another type of speaker, e.g. amici curiae (“friends of the court”) has been shown to have influences on legal judgments (Sim et al., 2015) and is also worthy of investigation.

All experiments were based on GPT2 (Radford et al., 2019). There is room for exploration of pre-trained language models, such as GPT3 (Floridi and Chiriatti, 2020), ChatGPT, BART (Lewis et al., 2020), or T5 (Raffel et al., 2022). In natural language generation, current work (Wu et al., 2020; Hu and Li, 2021) has introduced casual models to generate counterfactual text, and this could further aid the analysis of the impacts of different social factors.

Our study is based on a subset of SCOTUS cases.¹² While they have similar statistics to full cases, they do not reflect the full conversation history of the US Supreme Court or represent the current state of the court. Also, all claims in this paper are bound to this specific use case. They are not generalizable to other SCOTUS parties, legal systems, social strategies, etc. Additionally, although, in preliminary experiments, we applied a temporal train/test split (holding out the final year, i.e. 2019, as test data) and verified that the general pattern of results (i.e. perplexity) is identical in both the random and the temporal split, temporal effects on SCOTUS cases, especially from a sociolinguistic perspective, would be worthy for further investigation.

References

Jonathan P. Chang, Caleb Chiam, Liye Fu, Andrew Wang, Justine Zhang, and Cristian Danescu-Niculescu-Mizil. 2020. [ConvoKit: A toolkit for the analysis of conversations](#). In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 57–60, 1st virtual meeting. Association for Computational Linguistics.

Cristian Danescu-Niculescu-Mizil, Lillian Lee, Bo Pang, and Jon Kleinberg. 2012. [Echoes of power: Language effects and power differences in social interaction](#). In *Proceedings of the 21st International*

Conference on World Wide Web, WWW ’12, page 699–708, New York, NY, USA. Association for Computing Machinery.

Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. [A computational approach to politeness with application to social factors](#). In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 250–259, Sofia, Bulgaria. Association for Computational Linguistics.

Pieter Delobelle, Ewoenam Tokpo, Toon Calders, and Bettina Berendt. 2022. [Measuring fairness with biased rulers: A comparative study on bias metrics for pre-trained language models](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1693–1706, Seattle, United States. Association for Computational Linguistics.

Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. 2021. [SemEval-2021 task 6: Detection of persuasion techniques in texts and images](#). In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 70–98, Online. Association for Computational Linguistics.

Lee Epstein, William M Landes, and Richard A Posner. 2010. Inferring the winning party in the Supreme Court from the pattern of questioning at oral argument. *The Journal of Legal Studies*, 39(2):433–467.

Amir Feder, Katherine A Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E Roberts, et al. 2022. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *Transactions of the Association for Computational Linguistics*, 10:1138–1158.

Ronen Feldman. 2013. [Techniques and applications for sentiment analysis](#). *Communications of the ACM*, 56(4):82–89.

Sarah E. Finch and Jinho D. Choi. 2020. [Towards unified dialogue system evaluation: A comprehensive analysis of current evaluation protocols](#). In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 236–245, 1st virtual meeting. Association for Computational Linguistics.

Luciano Floridi and Massimo Chiriatti. 2020. GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30(4):681–694.

Liye Fu, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Tie-breaker: Using language models to quantify gender bias in sports journalism. In *Proceedings of the IJCAI workshop on NLP meets Journalism*.

¹²We extract a subset from the Covokit dataset (Chang et al., 2020) which is distributed under the MIT license.

- Liye Fu, Susan Fussell, and Cristian Danescu-Niculescu-Mizil. 2020. [Facilitating the communication of politeness through fine-grained paraphrasing](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5127–5140, Online. Association for Computational Linguistics.
- Matthew Gentzkow, Jesse M Shapiro, and Matt Taddy. 2018. Congressional record for the 43rd-114th congresses: Parsed speeches and phrase counts. In URL: https://data.stanford.edu/congress_text.
- Yue Guo, Yi Yang, and Ahmed Abbasi. 2022. [Auto-debias: Debiasing masked language models with automated biased prompts](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1012–1023, Dublin, Ireland. Association for Computational Linguistics.
- Zhiting Hu and Li Erran Li. 2021. [A causal lens for controllable text generation](#). In *Advances in Neural Information Processing Systems*, volume 34, pages 24941–24955. Curran Associates, Inc.
- Guido W Imbens and Donald B Rubin. 2015. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.
- Masahiro Kaneko and Danushka Bollegala. 2021. [De-biasing pre-trained contextualised embeddings](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1256–1266, Online. Association for Computational Linguistics.
- Katherine Keith, David Jensen, and Brendan O’Connor. 2020. [Text and causal inference: A review of using text to remove confounding from causal estimates](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5332–5344, Online. Association for Computational Linguistics.
- Wojciech Kulesza, Dariusz Dolinski, Avia Huisman, and Robert Majewski. 2014. The echo effect: The power of verbal mimicry to influence prosocial behavior. *Journal of Language and Social Psychology*, 33(2):183–201.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. [BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.
- Mingyang Li, Louis Hickman, Louis Tay, Lyle Ungar, and Sharath Chandra Guntuku. 2020. [Studying politeness across cultures using english twitter and mandarin weibo](#). *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2).
- Masahiro Mizukami, Koichiro Yoshino, Graham Neubig, David Traum, and Satoshi Nakamura. 2016. [Analyzing the effect of entrainment on dialogue acts](#). In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 310–318, Los Angeles. Association for Computational Linguistics.
- Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. [StereoSet: Measuring stereotypical bias in pretrained language models](#). 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 5356–5371, Online. Association for Computational Linguistics.
- Kate G Niederhoffer and James W Pennebaker. 2002. Linguistic style matching in social interaction. *Journal of Language and Social Psychology*, 21(4):337–360.
- Bill Noble and Raquel Fernández. 2015. [Centre stage: How social network position shapes linguistic coordination](#). In *Proceedings of the 6th Workshop on Cognitive Modeling and Computational Linguistics*, pages 29–38, Denver, Colorado. Association for Computational Linguistics.
- Judea Pearl. 2009. *Causality*. Cambridge university press.
- Reid Pryzant, Dallas Card, Dan Jurafsky, Victor Veitch, and Dhanya Sridhar. 2021. [Causal effects of linguistic properties](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4095–4109, Online. Association for Computational Linguistics.
- Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2022. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(1).
- Yanchuan Sim, Bryan R Routledge, and Noah A Smith. 2015. The utility of text: the case of amicus briefs and the supreme court. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.
- Paul J Taylor and Sally Thomas. 2008. Linguistic style matching and negotiation outcome. *Negotiation and conflict management research*, 1(3):263–281.
- Victor Veitch, Dhanya Sridhar, and David Blei. 2020. Adapting text embeddings for causal inference. In *Conference on Uncertainty in Artificial Intelligence*, pages 919–928. PMLR.

Tianlu Wang, Xi Victoria Lin, Nazneen Fatema Rajani, Bryan McCann, Vicente Ordonez, and Caiming Xiong. 2020. [Double-hard debias: Tailoring word embeddings for gender bias mitigation](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5443–5453, Online. Association for Computational Linguistics.

Yiquan Wu, Kun Kuang, Yating Zhang, Xiaozhong Liu, Changlong Sun, Jun Xiao, Yueting Zhuang, Luo Si, and Fei Wu. 2020. [De-biased court’s view generation with causality](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 763–780, Online. Association for Computational Linguistics.

Yang Xu, Jeremy Cole, and David Reitter. 2018. [Not that much power: Linguistic alignment is influenced more by low-level linguistic features rather than social power](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 601–610, Melbourne, Australia. Association for Computational Linguistics.

A The SCOTUS data set

Cases were removed if: (i) voting results of individual Justices were missing; or (ii) the side of the Advocate (petitioner or respondent) was unavailable; or (iii) the case was associated with more than one sitting.

When constructing pairs of Justice–Advocate turns where the Justice utterance is directly followed by an Advocate utterance, we disregard the first four utterances per sitting, as they largely consist of legal boilerplate text.

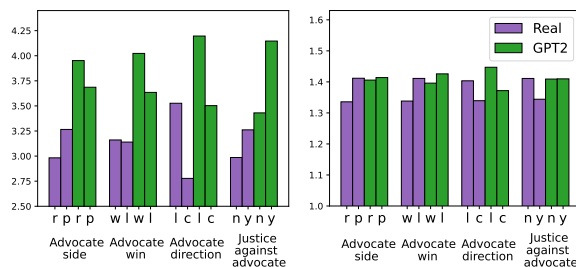
We removed all nonlinguistic information from the transcripts, including indicators of cross-talk (e.g., [voice overlap], [interruption]), nonverbal expressions (e.g., [laughter], [sighs], [applause]), and procedural markers (e.g., [luncheon], [recess]). The full list will be made available as part of the code repository.

B Fine-tune Configuration

For our experiments, we used GPT2-small,¹³ with 124M parameters. We fine-tuned the GPT2 model on our training data for 50 epochs, using the AdamW optimizer with a learning rate of 5e-5 and a batch size of 2. We select the model that performs the best in the Justice–Advocate pairs on the dev set based on perplexity.

Overall, it took approximately 24h to fine-tune the GPT2 model on one NVIDIA A100 40GB GPU.

¹³<https://huggingface.co/gpt2>



(a) Coordination.

(b) Politeness.

Figure 4: Expanded experiments on coordination and politeness scores in Advocates’ responses toward factual Justices. Advocate side: (r=respondent, p=petitioner); Advocate win: (w=Win, l=Lose); Advocate direction: (l=Liberal, c=Conservative); Justice against Advocate: (y=Yes, n=No).

C Further Validation of Coordination and Politeness in GPT2 Responses

We present an extended set of results for Experiment 1 (Section 3), involving an additional four binary attributes: (1) Advocate’s side (respondents; petitioner), (2) Advocate wins (the final voting result is in favor of the side of the Advocate; or not), (3) Advocate direction (liberal, conservative), and (4) Justice against Advocate (the Justice votes eventually against the side of the Advocate; or not).

Figure 4 shows the comparison of real scores derived from the SCOTUS data (purple) vs. factual GPT2-generated responses (green) for coordination (4a) and politeness (4b), respectively. With the exception of “Advocate side” in coordination (leftmost block in Figure 4a), factual GPT2-generated responses align with the real differences. This is in line with our results for attributes gender, party, and seniority in Section 3, and confirms the capability of the fine-tuned GPT2 model to capture coordination and politeness in Advocates’ responses toward Justices.

D Generated Samples

We provide two example contexts with real, factual generated and counterfactually generated responses. Figure 5 involves a question asked by a male non-chief judge, and Figure 6 involves a female non-chief judge. For, each we provide two counterfactual responses where we flip the justice attributes gender and seniority, respectively. We also label politeness indicators with underlining and corresponding index in the list of politeness markers from Danescu-Niculescu-Mizil et al. (2013) in Table 3.

0	“Please”	1	“Please start”	2	“Hashedge”
3	“Indirect (btw)”	4	“Hedges”	5	“Factuality”
6	“Deference”	7	“Gratitude”	8	“Apologizing”
9	“1st person pl.”	10	“1st person”	11	“1st person start”
12	“2nd person”	14	“2nd person start”	13	“Indirect (greeting)”
15	“Direct question”	16	“Direct start”	17	“Positive lexicon”
18	“Negative lexicon”	19	“Counterfactual modal”	20	“Indicative modal”

Table 3: List of politeness markers from Danescu-Niculescu-Mizil et al. (2013). See Table 3 in the cited paper for details and explanations.

Overall, compared to real advocate responses, GPT2-generated responses contain more politeness indicators. Compared to responses toward male justices, generated responses toward female justices exhibit more politeness strategies.

E Breakdown Coordination Scores for Markers in Advocates’ Responses toward Male vs. Female Justices

As shown in Figure 7, to study the confounds in pre-trained LMs, we investigate the breakdown coordination scores for markers with Advocates towards male vs. females Justices.

For every coordination marker, the GPT2-generated ratio for male vs. female corresponds to the real-world relative prevalence observed in the SCOTUS data. In addition, the absolute prevalence per gender was comparable between real and generated texts for all categories except the use of adverbs, where GPT2 generates significantly more adverbs than were observed in the text (likely a reflection of the less institutional language that dominates the GPT2 training data).

F Annotation Instructions for Evaluating Legal Court Advocates’ Responses

The goal of this annotation project is to evaluate the “quality” of conversations in a court room along a variety of criteria. Specifically, you will rate the quality of an advocate’s response to a preceding justice’s turn. This document includes two sections: Annotation Instructions and Rating Explanations.

F.1 Annotation Instructions

You will be provided with extracted excerpts from the US supreme court transcripts. Each excerpt contains three parts:

1. Previous context of 2-6 utterances preceding Justice’s turn.
2. Justice’s turn.

Evaluation Dimension	Explanation	Evaluation Dimension Explanation Corresponding number
Relevance	The advocate’s response is on-topic with the previous context and justice’s turn	
Informativeness	The advocate’s response provides unique and non-generic information Note that high informativeness does not always mean the responses are providing consistent results.	0: Both responses are equally good. 1: Response 1 is better. 2: Response 2 is better.
Consistency	Consistency The advocate’s response does not produce contradictory information wrt. the context. Note that high consistency does not necessarily mean high informativeness	

Table 4: Rating metrics

3. Advocate’s response to Justice’s turn.

The (3) advocate’s response is the part that you will evaluate. More specifically, two advocates’ responses are provided, and your task is to compare the two advocate’s responses based on the Previous context and Justice’s turn, with the evaluation metrics defined in the following section.

F.2 Rating Explanations

We detail the rating metrics for pairwise evaluation on the advocate’s responses in Table 4.

Preceding context: [ADVRES] Mr. Chief Justice, and may it please the Court: This is an obstacle preemption case masquerading as a field preemption case. There is no field preemption here because of the undisputed fact that the Nuclear Regulatory Commission does not and cannot regulate any aspect of uranium mining, including the safety aspects of uranium mining. And there is no obstacle preemption here because the Atomic Energy Act and the Nuclear Regulatory Commission have repeatedly reaffirmed that states have the ability to regulate mining up to and including by banning it altogether. [KAVANAUGH] They – they can regulate milling, correct? [ADVRES] Excuse me, I want to make – the federal government regulates milling – [KAVANAUGH] Right. [ADVRES] – Justice Kavanaugh, yes.

Justice turn: [KAVANAUGH] Yeah. So the mining and milling occur together, correct? In other words [...] you don't have mining without milling; you don't have milling without mining

Real response to [KAVANAUGH]
[ADVRES] Justice Kavanaugh, there's a way – the in situ leaching process, they literally occur at the same time.

Factual GPT2 to [KAVANAUGH]
[ADVRES] We_{4,9} assume_{2,4} that the states aren't going to misstate their reasons for doing something. But₁₆ I₁₀ will tell you₁₂ that it is a lot more difficult₁₈ for a state to, for example, rezone all land that nuclear activity is taking place on as residential, even if it's totally unsuitable for residential dwelling, if they know they're going to have to come forward and justify that. And₁₆ may₂ I₁₀ finish the second point, Your₁₂ Honor₁₇?

Counterfactual GPT2 to [KAVANAUGH] → [KAGAN]
[ADVRES] That's correct₁₇. And₁₆ – and – and that's – that's exactly what the NRC has said. It has said that it has the authority to regulate the milling and the – and the operation of the plant. And₁₆, in fact₅, it has done so in a number of ways. First of all, it's done so on a case-by-case basis. The NRC does not have the authority under the Clean₁₇ Air Act or the Clean₁₇ Water Act to regulate all of the operations of a nuclear plant. That's not the case here. This is not a case where there is a single federal regulation of the entire operation of a plant.

Counterfct'l GPT2 to [KAVANAUGH] → [REHNQUIST]
[ADVRES] That's correct₁₇, Justice Rehnquist. And₁₆ that's – that's exactly what the NRC has done here. The NRC regulates the operation of nuclear power plants. That's what it has done in the past. It has regulated the operation –

Figure 5: Example 1 with context, Justice question (top) and real, factual-generated and two counterfactually generated Advocate answers (bottom). The factual Justice is [KAVANAUGH] (male, non-chief) who is counterfactually replaced with [KAGAN] (female, non-chief) and [REHNQUIST] (male, chief), respectively.

Preceding context: [BADER-GINSBURG] And do we know – do we know what costs Blue Shield then passes on to those other self-insured plans? [ADVRES] We don't know it. We were not able to get it from Blue Cross at the time the case was before the district judge. But in any event, as the Chief Justice's question suggests, the – this issue doesn't end at Vermont. It – it has to take account of a burden that Congress was very, very aware of, it was very cognizant of the regime under the Repeal Disclosure Act, [...] Just as if – if Vermont decides next week, if it wins this case, that it wants to get information about pension plans, and how they're being administered, and what benefits are being provided and not provided, those are all 100 percent self-funded plans. If my employer has to provide all that information, that is coming out of my 401(k) benefits.

Justice turn: [KAGAN] Mr. Waxman, could a State pass a tax law that requires information about – about pension disbursements, about claim payments, about assets held in trust? Could a State do that?

Real response to [KAGAN]
[ADVRES] I_{4,11} don't think_{2,4} that a State – when you₁₂'re saying requested information or imposing₁₈ a tax?

Factual GPT2 to [KAGAN]
[ADVRES] stuff that they are asking for is so obviously critical₁₈ to what the plan does.

Counterfct'l GPT2 to [KAGAN] → [BADER-GINSBURG]
[ADVRES] Yes, Your₁₂ Honor₁₇. I₁₁ – I₁₁ don't believe that that would be preempted by ERISA. And₁₆ I₄ think_{2,4} that the – the key point here is that this is not an ERISA plan. This is a plan that is created by the Federal government. The Federal government is the one that created the plan here, and it's the only one that has the authority to create a plan here. And₁₆ so I₄ think_{2,4} it's – it's very important₁₇ to understand that this plan is not a plan for the benefit₁₇ of the – of the plan participants. That is, it's a plan to provide benefits₁₇ to people who are already covered by the plan, and that's what Congress was trying to do when it passed the ERISA statute. Congress wanted to make sure that people who were already covered under the plan would get the benefits₁₇ they were entitled to. So₁₆ it's not like₁₇ a plan where the plan participant is going to have to pay all of the costs

Counterfct'l GPT2 to [KAGAN] → [REHNQUIST]
[ADVRES] No, I₁₀ don't believe so, Your₁₂ Honor₁₇. I_{4,11} don't think_{2,4} that that would be preempted by ERISA. ERISA preempts State laws that regulate ERISA plans.

Figure 6: Example 2 with context, Justice question (top) and real, factual-generated and two counterfactually generated Advocate answers (bottom). The factual justice is [KAGAN] (female, non-chief) who is counterfactually replaced with [BADER-GINSBURG] (female, non-chief) and [REHNQUIST] (male, chief).

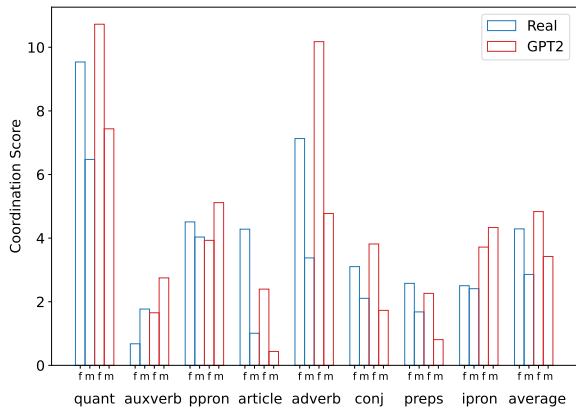


Figure 7: Breakdown coordination scores for markers in Advocates' responses toward male vs. female Justices.