Online Supporting Information:
Large Language Models Can Argue in Convincing and Novel Ways About Politics

Contents (Appendix)

A Data Availability and Human Subjects

Data Availability Statement: all data underlying the paper will be released publicly in the event that this manuscript is accepted for publication. Code to replicate the figures and table will also be released. Both code and data will be deposited at Harvard's `Dataverse` or a system similarly publicly accessible.

Human Subjects: Our university IRB approved all protocols; all participants gave informed consent. Subjects were informed as to the purpose of the research and were able to opt out of participation at any time. The tasks, both writing and choosing arguments, reflected topics and language present in everyday life. The only form of deception was not informing approximately half of participants that the author of some arguments varied, which was necessary for the research and did not constitute an additional risk. Finally, participants were recruited through MTurk and compensated $4 per task through that platform (this is commensurate or higher than local minimum hourly wages given the task takes a few minutes).

B LLM and interface specification

To provide the LLM written arguments, we used Open Pre-trained Transformer Language model from Zhang et al. (2022). The files associated with the model were downloaded into our environment on June 1, 2022. On June 23, 2022 the weights on the OPT-30B were adjusted; these adjustments were not added to our files. However, as suggested by the classification in Figure 2 and the results in Figure 3, the text quality does not drive the main treatment effect; therefore we do not expect this to make a substantial difference for our main results.

We generated 15-30 arguments in response to each of our prompts. For four of the nine ("most important problem", more restrictions on abortion, more gun control, opt-in organ donation) we also ran a large batch of 300 responses to assess how often the LLM produced usable/unique arguments. See C for more details on that analysis.

For all model runs, we did minimal adjusting from the default parameters, aside from specifying the max length of 150 tokens. This was in part due to the difficult in assessing optimal performance when changes were made. However, we did test several configurations for number of beams and implementing early stopping. However, beam search generally produced less usable text. Therefore, instead of early stopping we did minimal editing to the responses, e.g. deleting repeating clauses and fixing punctuation. We then filtered for those we judged to be coherent and 'on topic', and selected the qualitatively 'best' three of the arguments for each prompt.

To solicit responses from crowdworkers, we developed an app that would provide five prompts for workers to answer with a limit of 300 characters. Each respondent answered only the 'pro' or the 'con' side of the first four prompts described above, and all saw the final, open ended question. They were asked to provide the best argument for the prompts regardless of personal opinion. We used Amazon MTurk to find respondents and we required that they must be in the U.S.; no other information was collected about them. We had a total of 50 participants for a total of 25 responses to each side of the first four topics and 50 responses to the last, open-ended question.

C Proportion of Useable LLM Arguments

As can be seen in Table C when we generated large batches of arguments, there were a maximum of $\frac{1}{6}$ of the sample size that consisted of unique, usable arguments.

	Argument	Total Runs	Usable Answers	Unique Usable Ans.
1	Opt-in Donation	300	29	27
2	More Gun Control	300	42	40
3	More Abortion Restrictions	300	76	49
4	Most Important Problem	300	231	31

D Curating Responses

From this set of responses, we again filtered by coherency and 'on topic'-ness. We also filtered out responses that exist in that form online. We (the authors) independently rated the remaining responses by quality and used the arguments with the best joint scores. Specifically, the two coauthors independently scored all responses as 'coherent' (0/1) and then 'on topic' (0/1), with arguments that scored 2 (out of 2) being allowed to remain in the pool.[1] For the crowdworker responses we jettisoned any that were copy-and-pasted from existing text online. Finally, the authors independently rated all remaining responses by their quality and selected the three 'best' arguments from each group based on the joint score per prompt.[2] Ultimately, this results in similar length arguments of comparable quality. While we acknowledge that this necessarily incorporates some degree of the preferences and life experiences of the authors, we note again that what follows are necessarily (plausible) "possibility results."

To compare which responses are preferred, we created a light-weight app and recruited participants through MTurk as judges. We collected 767 responses, which were filtered for adequate task completion etc, in May 2023 after pilot runs in December 2022 and January 2023. Each judge was randomly assigned 10 pairs of arguments, with the prompt (from nine possible), LLM and human arguments (from three each possible per prompt), and order of the same randomized. They were asked, given the question, regardless of personal opinion, which is the "best" argument. Judges were randomly assigned to either a control condition, where the authors of all arguments were anonymous (379 respondents) or a treatment condition where they were informed whether a human or an LLM wrote the argument (388 respondents).

E How human and LLM arguments differ

E.1 Descriptive Statistics

On "reading ease" in the sense of Flesch, LLM arguments are typically higher mean ($p < 0.05$) and lower variance. That is, LLM arguments are easier to read, and tend to be more similar to each other on this metric. Second, while the parts of speech used were very similar across groups, the overall sentiment varied. The LLM was consistently more positive in speech (multiple dictionaries, $p < 0.05$ in one case). However, the human-written text covered a wider variety of sentiment. That is, the LLM produced little variation in tone as compared to humans. There was (at most) weak correlation between positive sentiment and judge preference for a given argument. Note that we used dictionaries from Hu and Liu (2004) and Stone, Dunphy and Smith (1966) to estimate sentiment.

[1] An argument is "coherent" to the extent that it literally makes sense in basic grammatical terms, and can be seen as an 'argument' for a given position/priority. An argument is "on topic" to the extent that it is on the subject or theme requested.

[2] Literally, for each orator type, the authors scored the arguments in terms of their perceived ability to 'beat' other arguments.

E.2 Coherence and Training Data

To reiterate, we note that the LLM constructs more coherent arguments on topics—like abortion, immigration and gun control—which are more frequently discussed online and in the general discourse. It does worse when asked to defend organ donation policies, for example, often producing only sentence fragments or arguing for unrelated positions. This is likely a consequence of (vastly) differing amounts of web training data. To get a sense of this, we inspected two popular communities on the social media site `Reddit` where such matters are discussed (`r/changemyview`, `r/politics`) and on which the LLM is trained. We found that "organ donation" is represented less than half as often as the next most popular of our focus topics in one subreddit, and 17 times less often in the other. Given this lack of balance, it is perhaps unsurprising that the LLM struggles with more obscure issues.

E.3 Nuance

Above we noted that humans and the LLMs differ on the relative nuance with which topics are discussed. This is obviously a judgement call, but more concretely we observe that the LLM typically uses fewer unique words than humans do (129 v 304), at least when advocating for abortion restrictions. And on the topic of abortion generally, the LLM uses more ($p < 0.01$) emotive 'conceptual' terms (as opposed to 'primordial' tokens, in the sense of Martindale (1990)) as a proportion of its statements than humans.

In terms of specific tokens here, the LLM uses "argument", "murder", "protect" and "controversial" more often. Meanwhile, humans reach for "consider", "harm" and "thought". Again, this is in keeping with where such models are trained, but suggests LLMs may be less nuanced in communication than a human attempting to persuade another person.

E.4 Aggregate Performance

We noted above that, overall, the LLM tended to have lower performing 'worse case' arguments (i.e. its least popular offerings were considerably less popular than the least popular ones from humans). As an aside here, we note that judges liked arguments that were logically ordered, and appealed to human welfare. For example, the most preferred arguments from each source in the control condition were (from a human and the LLM, respectively):

- **Human:** "There are a ridiculous number of people waiting for organs on the transplant lists that have to wait sometimes years to get said organ, even though people die every day. This is because the dying do not donate their organs enough, so making it default is better for those waiting to continue living."
- **LLM:** "I think the best argument for more gun control is that it is a proven fact that more guns in the hands of more people leads to more gun violence. The United States has more guns per capita than any other country in the world, and we have the highest rate of gun violence in the world."

E.5 Embeddings

We used the R package `doc2vec` to create embeddings and topics from these embeddings on the prompts for which were had generated a large number of responses. These were plotted along two dimensions of topics and clustered using k-means to determine the similarity of the language used.

F Supporting Data for Figure 1

Unadjusted, only the probability that the human argument for reducing abortion and opt-in organ donation was chosen more than the LLM are different from 0.5 and significant; in this case greater than 0.5. However, if we apply a Bonferroni correction and adjust the significance threshold $p = .05/10$, then the result is no longer significant.

Table 1: Probability LLM argument was chosen in the control condition

Prompt	Probability	SE	N	P-value
More gun control	0.49	0.026	384	0.75
Less gun control	0.50	0.024	422	1.00
More abortion	0.51	0.025	416	0.73
Less abortion	0.55	0.024	439	0.02
More immigration	0.47	0.024	442	0.23
Less immigration	0.50	0.024	428	0.88
Opt-in donation	0.56	0.025	401	0.01
Opt-out donation	0.50	0.025	411	0.88
Main problem	0.50	0.024	427	0.88
All	0.51	0.0082	3759	0.26

G Author Regression Results

Treatment and control were compared using a linear regression (1) with "choosing the human argument of a human/LLM pair" as the binary dependent variable ($Y \in \{0, 1\}$) and (2) with each argument instance now an observation, with the dependent variable being whether it was chosen and with the author and treatment condition as the independent variables (plus their interaction). Consider Table 2. There we give the relevant coefficient estimates for a regression of preferring a human argument ($Y = 1$) on the treatment, which is knowing the identity of the author and the interaction of that with that author being the LLM, using prompt fixed effects and clustered standard errors. The point here is that the interaction is statistically significant, and negative: that is, overall, when judges are told that a given argument is produced by an LLM they are more likely to prefer the human-produced argument.

Table 2: Regression results showing that an LLM wrote the argument causes judges to prefer human offerings over machine ones.

	Dependent variable:
	Likelihood an argument is picked
Audience Knows Author	0.049^{***}
	(0.008)
Arg. Written by LLM	-0.018
	(0.020)
LLM*Knows Author	-0.099^{***}
	(0.015)
Prompt FE	Yes
Observations	15,302
R^2	0.007
Adjusted R^2	0.006
Note:	**p<0.05

Based on Table 2, in the control group, judges are about 2 percentage points more likely to pick the human-written arguments (than LLM arguments) on average, though this is not significant. When informed about the author, they are an additional 10 percentage points more likely to pick the human written argument, a substantial increase over the initial difference.

In SI H we adjust this analysis for multiple comparisons.

H Adjusting for Multiple Comparisons

Without adjusting for multiple hypotheses, both the overall treatment effect and the effect for several specific prompts—more gun control, less abortion, and less immigration—are positive and significant. However, if we apply a Bonferroni correction and adjust the significance threshold $p = .05/10$, only the overall effect is significant.

Table 3: Treatment effect of knowing the author on likelihood of preferring the human written argument

Prompt	Coefficient	SE	P-value
More gun control	0.071	0.035	0.043
Less gun control	0.013	0.033	0.70
More abortion	0.056	0.034	0.109
Less abortion	0.07	0.033	0.035
More immigration	0.048	0.034	0.15
Less immigration	0.084	0.034	0.014
Opt-in donation	0.022	0.034	0.50
Opt-out donation	0.057	0.035	0.10
Main problem	0.043	0.034	0.21
All	0.051	0.0079	0.00018

References

Hu, Minqing and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. pp. 168–177.

Martindale, Colin. 1990. *The clockwork muse: The predictability of artistic change.* Basic Books.

Stone, Philip J, Dexter C Dunphy and Marshall S Smith. 1966. *The General Inquirer: A computer approach to content analysis.* MIT press.

Zhang, Susan, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang and Luke Zettlemoyer. 2022. "OPT: Open Pre-trained Transformer Language Models.".