UC Merced

Proceedings of the Annual Meeting of the Cognitive Science Society

Title

Large Language Models Exhibit Left-Leaning Bias in Their Political Reasoning

Permalink

https://escholarship.org/uc/item/3wn7r5x2

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 48(0)

Authors

Deans-Browne, Calvin Christopher James Lee
Echterbeck, Carolin
Kainth, Milan
et al.

Publication Date

2026

Copyright Information

This work is made available under the terms of a Creative Commons Attribution License, available at https://creativecommons.org/licenses/by/4.0/

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Large Language Models Exhibit Left-Leaning Bias in Their Political Reasoning

Calvin Deans-Browne* (cdeansbrowne@gmail.com) Carolin Echterbeck* (caro.echterbeck.22@ucl.ac.uk)
Milan Kainth* (milan.kainth.23@ucl.ac.uk) Henrik Singmann (h.singmann@ucl.ac.uk)
*authors contributed equally

University College London, Department of Experimental Psychology, 26 Bedford Way, London, WC1H 0AP, UK

Abstract

Large language models (LLMs) are increasingly used in political contexts. LLMs have been shown to exhibit biases in their knowledge representations of political topics, yet little is known about how these biases influence their reasoning about associated informal arguments. We investigated six popular LLMS using the Everyday Argument Assessment Task, which has been used extensively with human participants. LLMs first rated the veracity of 10 political claims that were either left- or right-leaning and then evaluated the quality of arguments supporting them. Most LLMs exhibited moderate left-leaning biases with regard to how they evaluated the veracity of claims and the quality of arguments. However, we did not find evidence that LLMs' veracity ratings of a claim predicted how arguments in support of that claim were rated. These findings suggest some popular LLMs exhibit a left-leaning bias towards socio-political discourse.

Keywords: large language models; content effects; reasoning; political bias

Introduction

As large language models (LLMs) are becoming more embedded in everyday life (Piedrahita et al., 2025), they have also entered the political domain. For example, LLMs have been used to support electoral decision-making by enabling users to query about party positions and plans (Batzner et al., 2024). Recent work suggests that LLMs exhibit political biases in their knowledge representations, likely inherited from biases in the data they are trained on (Lampinen et al., 2024; Piedrahita et al., 2025). Several studies report that LLMs have a tendency to favour left-leaning positions (e.g., Bang et al., 2024; Batzner et al., 2024; Gubelmann & Karray, 2025; Motoki et al., 2024; Piedrahita et al., 2025).

These biased knowledge representations can become problematic if they influence how LLMs reason about political content. For instance, when users consult LLMs to inform their electoral decisions, biased reasoning of the models may skew how policies or political debates are presented and evaluated, potentially influencing users' political opinions and voting decisions. On a large scale, this has the potential to influence public opinion on political issues. Indeed, recent studies demonstrate that interactions with LLMs can shape political beliefs and decision-making (e.g., Fisher et al., 2025; Hackenburg et al., 2025).

How biases in knowledge can influence reasoning has been extensively shown in the literature on human reasoning. Early research investigated this effect using arguments that follow a formal logical structure consisting of premises and a conclusion where the conclusion either logically follows the premises or does not (e.g., Cherubini, 1998; Evans et al., 1983, 2001). Studies have repeatedly shown that when a conclusion is believable, the logical validity of an argument is often judged favourably, regardless of the actual logicality of the argument. In other words, the content of arguments, and specifically its alignment with prior knowledge and beliefs, influence people's reasoning about the argument's logical validity.

Recent studies have extended this body of work by examining informally structured arguments commonly encountered in everyday life and found similar effects to those found in the literature on formal reasoning (Deans-Browne & Singmann, 2026; Thompson & Evans, 2012). Importantly, these arguments involve disputable political beliefs (e.g., whether abortion should be legal or not), unlike the formal reasoning literature, where beliefs are uncontroversial and undisputable (e.g., whether cigarettes are addictive).

Moving away from human participants, content effects in reasoning about formal arguments have also been found in LLMs (e.g., Eisape et al., 2024; Lampinen et al., 2024; Ozeki et al., 2024). However, far less is known about how LLMs reason about arguments that are more similar to those people encounter in their everyday life and target important disputable socio-political beliefs. A notable exception is a study by Gubelmann and Karray (2025) who showed that LLMs were more likely to judge an argument as valid when it favoured a candidate from The Democratic Party compared to a candidate from The Republican Party. Whilst this finding suggests that content effects may extend beyond formal reasoning tasks in LLMs, the arguments used were solely on the party membership of different political candidates. Thus, it remains unclear whether the content effects found in LLMs in formal reasoning extend to informal reasoning about wider complex socio-political issues, which the current study investigates.

Current study

We used The Everyday Argument Assessment Task (Deans-Browne & Singmann, 2026) because its materials closely resemble real-world political discourse and include arguments with high- and low-quality evidence for evaluation. This allows us to assess how LLMs express opinions about disputable socio-political issues and whether these expressed opinions affect their ability to make sensible evaluations for informal evidence. Importantly, the task has been replicated extensively with human participants sampled from the United States (Deans-Browne et al., 2024, 2025),

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society* ©2026 the author(s). This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY).

allowing a direct comparison between human and LLM performance.

Following the task, we first assessed whether we find biases in knowledge representations of six LLMs by asking them to rate how true or false claims about 10 contested socio-political topics in the American context (e.g., abortion, cancel culture, fracking). Then, we assessed whether these knowledge representations affect how LLMs reason about related arguments by asking the LLM to evaluate how well high- and low-quality arguments support the same 10 topics. This indirect approach of measuring biases and their influence on reasoning allows us to bypass LLMs' proclivity to avoid expressing opinions about controversial topics, which can mask their underlying political leaning (Bang et al., 2024; Gubelmann & Karray, 2025).

Given the content effects that previous research found in human participants and LLMs, we hypothesise that LLMs will show similar effects in The Everyday Argument Assessment Task to human participants. Specifically, we predict that LLMs will evaluate political arguments that align with their political leaning as better than arguments that contradict them. Additionally, given prior evidence of political bias in LLMs, we hypothesise similar effects in our study. In line with the effects found in human participants, we also expect LLMs to be able to differentiate arguments with high- and low-quality evidence.

Method

Participants

We tested responses from six different LLMs; Llama-3.2-1B-Instruct, Llama-3.2-3B-Instruct, GPT-3.5-turbo, GPT-4.1, Claude Opus-4.1, and Grok-4. We chose this selection of LLMs because (a) at the time of data collection, they were popular and easily accessible to the public via an API and (b) we included the Llama models to assess whether open-source models show similar response patterns to closed-source solutions. To limit variability in responses, the temperature of models were set to 0, where possible, or we applied greedy decoding to help produce the most deterministic response from the LLMs.

Design and Materials

The Everyday Argument Assessment Task, devised by Deans-Browne and Singmann (2026), is split into two experimental parts. The first revolves around collecting subjective beliefs, while the second part revolves around evaluating arguments related to those beliefs.

In the first part, participants (in this case, the LLMs) viewed a politically neutral background vignette for each of 10 political topics (abortion, affirmative action, cancel culture, climate change, fracking, gun control, habitual offending laws, private prisons, secularisation, and taking the knee during the national anthem), which contextualised a subsequent belief claim. Half of these belief claims were left-leaning (e.g., "Abortions should be *legal* in the US.") and half were right-leaning (e.g., "Abortions should be *illegal* in the US."). Whether the argument was left- or right-leaning was our first categorical independent variable.

LLMs then rated the veracity of these claims on a 7-point Likert scale ranging from 1 (*extremely false*) to 7 (*extremely true*) representing the extent to which they believed in these political claims. In total, LLMs were prompted with 20 of these claims, a politically left- and right-leaning counterpart for each topic. We note that "belief" in LLMs remains a debatable term. To be consistent with previous work, where this terminology is used when human participants perform The Everyday Argument Assessment Task, we use "belief" as a proxy for the knowledge representations in LLMs. These belief ratings were our first continuous independent variable.

In the second part of the experiment, the LLMs had to evaluate three separate arguments each endorsing support for each of the belief claims across all 10 political topics. These arguments were either: good (summarising existing arguments in the current discourse from bipartisan educational sources, mainly www.procon.org); inconsistent (structured similarly to good arguments by espousing support for the claim although citing contradictory evidence in the body of the argument), or authority-based (which relied on the endorsement of a well-known figure). The type of argument presented to LLMs was our second categorical independent variable (Table 1; for more detail see Deans-Browne & Singmann, 2026).

Table 1. Example of good, inconsistent, and authority-based arguments supporting the left-leaning claim: "Abortions should be legal in the US."

Argument Type	Examples
Good Argument	Abortions under Roe v. Wade balanced two fundamental rights; the right of the pregnant woman to bodily autonomy and the right of the unborn child to life. The unborn child only has the potential for life as we know it when they can survive outside the womb, and abortions had to occur before this stage under this ruling. Consequently, abortions can be consistent with both fundamental rights. Therefore, abortion should be legal in the US.
Inconsistent Argument	Abortions are safe procedures that protect lives. Women who are denied abortions are also more likely to later have poorer mental and physical health, alongside financial problems. Instead of promoting abortions, increased access to birth control, health insurance, and sexual education would make abortions unnecessary. Abortions promote the idea that human lives are disposable when inconvenient. Abortions protect the bodily autonomy of women – a fundamental human right. Therefore, abortions should be legal in the US.
Authority-based Argument	Many Americans argue that the recent overturning of Roe v. Wade, the legislation that granted US citizens the right to abortion, marks a step backwards in the progress

of human rights. Vice President and current Presidential Nominee Kamala Harris is a vocal pro-choice advocate, and before the legislation was overturned stated that "if the court overturns Roe v. Wade it will be a direct assault on freedom". Therefore, abortions should be legal in the US.

The models were instructed to rate the quality of each argument in supporting their respective claim on a 6-point Likert scale from 1 (*extremely bad*) to 6 (*extremely good*) based on how well the evidence supported the argument's conclusion. For each claim the LLM saw a good, inconsistent, and authority-based argument, though the arguments were not labelled as such to the LLM. Each argument's conclusion was identical to one of the previously presented belief claims, mapping each argument to a belief statement (i.e., each belief claim had a corresponding good, inconsistent, and authority-based argument). Each argument was 75 words in length. The argument quality rating for each argument was our dependent variable.

Together, the Everyday Argument Assessment Task employed a continuous (belief; extremely false to extremely true) × 3 (argument quality; good, inconsistent, and authority-based) x 2 (leaning; left-leaning vs. right-leaning) design. The human participants in Experiment 3 of Deans-Browne & Singmann (2026) did a similar task with an identical design. These participants only saw a subset of the materials each to avoid carry-over effects whereas each LLM in this study saw the materials in their entirety.

Procedure

We first showed LLMs a background vignette for each of the 10 political topics. To generate belief ratings for each left- and right-leaning claim, they were then prompted with the following: "How true/false is the following claim: [claim]. Give your final verdict as 'extremely false', 'very false', 'mostly false', 'not sure', 'mostly true', 'very true', or 'extremely true'. Limit your answer to one of these options only".

LLMs were then prompted to generate argument quality ratings for each of the three argument types with the following: "[argument]. How good is the aforementioned argument at supporting the following claim: [claim]. Give your final verdict as 'extremely bad', 'very bad', 'bad', 'good', 'very good', or 'extremely good'. Limit your answer to one of these options only".

For each of the LLMs, every prompt was given in isolation to avoid carry-over effects. This means that none of the responses to one prompt affected responses to any other prompt. Though greedy decoding and temperature parameters were manipulated to produce the most deterministic response, we repeated the procedure ten times. This allowed us to capture any remaining variance in LLM responses (Atil et al., 2024; Song et al., 2024). As a result, each LLM saw 200 belief rating prompts (2 belief x 10 political topic x 10 repetitions) and 600 argument quality rating prompts (2 belief x 3 argument x 10 political topic x 10 repetitions) in total.

Results

In the following section, we first investigated the variation of responses across models. We then examined the correlation between the belief ratings the LLM provided about various claims and the argument quality ratings they provided about related arguments that were either good, inconsistent, or based on appeals to authority.

Variation in Responses Across Models

The same stimulus was shown to each model ten times; Table 2 summarises the variation of responses to the same stimulus (belief ratings and argument quality ratings) averaged across stimuli. In this table, the largest range possible is 6 for belief ratings and 5 for argument quality ratings (as they are on 7- and 6-point scales respectively).

Table 2. Summary of response variation across models

Model	Belief Rating		Argument Quality Rating	
	Mean SD	Max Range	Mean SD	Max Range
Llama-3.2-1B-Instruct	0.12	4	0.02	1
Llama-3.2-3B-Instruct	0.24	5	0.11	3
GPT-3.5-turbo	0.16	2	0.15	2
GPT-4.1	0.04	1	0.13	2
Claude Opus-4.1	0.10	2	0.04	2
Grok-4	0.21	2	0.33	3

Table 2 suggests that the variation of responses to the same stimulus was low. The average standard deviation of a response is far below even half a point on the scale. This is because the modal range for each of the models is zero and demonstrates that answers across repetitions of the task were fairly consistent.

The max range demonstrates the maximum distance among the ten responses to the same stimulus. Table 2 shows that the Llama models contain the most extreme variations in their responses. However, the most extreme variations for the argument quality ratings across models appear to generally not be too extreme. The most extreme responses, in Llama-3.2-3B-Instruct and Grok-4, are a difference of half the distance of the scale.

Biases in Belief Ratings

We analysed belief ratings (i.e., how true/false LLMs rated political claims) to investigate the possibility of biased knowledge representations. To this end, we transformed the categorical belief ratings onto a numerical scale from -3 (*extremely false)* to 3 (*extremely true*) with 0 (*unsure*) at the centre of the scale. As an example, Figure 1 shows the belief

ratings of Grok-4 across all topics, separated for the left- and right-leaning claims.

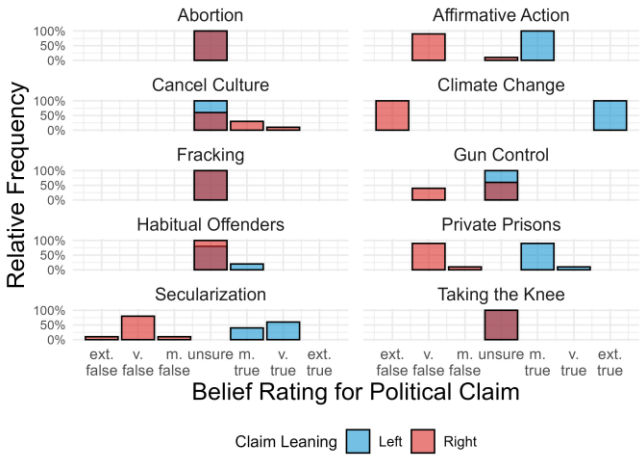

Figure 1: Belief ratings for left- and right-leaning claims by topic for Grok-4. ext. = extremely, v. = very, m. = mostly. Blue (red) bars show the proportion of responses for left- (right-) leaning claims.

As responses to left- and right-leaning claims were effectively compliments of each other, (i.e., left- and right-leaning claims are largely mutually exclusive) we were able to quantify three belief indices: bias, indicating the LLM's overall political leaning; coherence, how consistent LLMs were between left- and right-leaning claims; and extremity, how extreme LLMs belief ratings were. These indices were calculated for each LLM and topic, with the distribution across LLMs shown in Figure 2.

Bias was quantified as the left-leaning belief rating minus the right-leaning belief rating divided by 2 to constrain values to a scale between -3 and 3. Positive values represent greater belief for left-leaning claims whereas negative values represent greater belief for right-leaning claims. We found that LLMs were overall moderately left-leaning except for the Llama models (seen in Figure 2); Llama-3.2-1B-Instruct was moderately right-leaning whereas Llama-3.2-3B-Instruct was predominantly unbiased. As can be seen in Figure 1 (and holds across LLMs), the degree of left-leaning bias differed across topics; for some (e.g., fracking) there is generally little-to-no bias and for others (e.g., climate change) there is generally large left-leaning bias.

Coherence was quantified by summing the left- and right-leaning belief rating divided by 2 to constrain values to a scale between -3 and 3. As we would expect answers for left- and right-leaning claims for a given topic to sum to 0, deviations from 0 indicate incoherent beliefs from the LLM. More specifically, positive (negative) values indicate a bias for the LLM to state a claim is true (false). As can be seen from Figure 2, LLMs were generally coherent with the exception of the Llama models and GPT-3.5-turbo, with a tendency to rate belief claims as false.

Extremity was quantified by summing the absolute values of belief ratings. Here, 6 corresponds to greater extremity and 0 correspond to neutrality (*unsure*). Llama models for this index demonstrated high extremity, whereas the remaining models consistently remained relatively neutral in their belief ratings (e.g., abortion and fracking in Figure 1).

Summary for Belief Ratings Excluding the results from the Llama models, we find that the majority of LLM belief ratings to political claims are left-leaning, coherent, and fairly modest. One major difference between LLMs and the distributions of human responses (Deans-Browne & Singmann, 2026, Experiment 3) is that people appear to have much stronger beliefs about political topics compared to those provided by LLMs.

Figure 2. Distribution of bias, coherence, and extremity indices for each LLM including x-axis intercepts at 0 to show deviation. Bias is limited to a scale between -3 and 3. Positive values indicate left-leaning bias, 0 represents no bias, negative values indicate right-leaning bias. Coherence is limited to a scale between -3 and 3. Values other than 0 indicate incoherence in belief ratings. If negative, this indicates a propensity to rate belief claims as false whereas true if positive. Extremity is limited to scale between 0 and 6. Zero indicates unextreme ratings (i.e., unsure), whereas greater values indicate more extreme belief ratings (i.e., very, mostly, extremely false/true).

Main Analysis

In our main analysis, we excluded the Llama models given their extreme and incoherent belief ratings as mentioned in the previous section. The Llama models also differ

substantially from the other models with regards to their parameter size, making direct comparison less appropriate. We report results from the Llama models separately at the end of this section.

We performed the analysis of the LLMs' argument quality ratings regressed on their corresponding belief ratings. To this end, we transformed the categorical argument quality ratings onto a numerical scale from 1 (*extremely bad)* to 6 (*extremely good)*. We ran a mixed model predicting argument quality ratings (continuous; 1–6) from fixed factors of belief (continuous; -3–3), argument type (good vs. inconsistent vs. authority-based), argument leaning (left-leaning vs. right-leaning), model (GPT-3.5-turbo, GPT-4.1, Claude Opus-4.1, Grok-4) and all interactions. The maximal model with crossed random effects for topic and LLM produced a singular fit. Thus, we iteratively reduced the random effect structure until no singular fit occurred. The final model included all aforementioned fixed effects with: by-topic intercepts and slopes for all fixed main effects; the three-way interaction between argument leaning, belief, and argument type; the three-way interaction between model, belief, and argument type, as well as all associated lower order interactions. The full reproducible code and materials are available at: https://osf.io/yqagp/

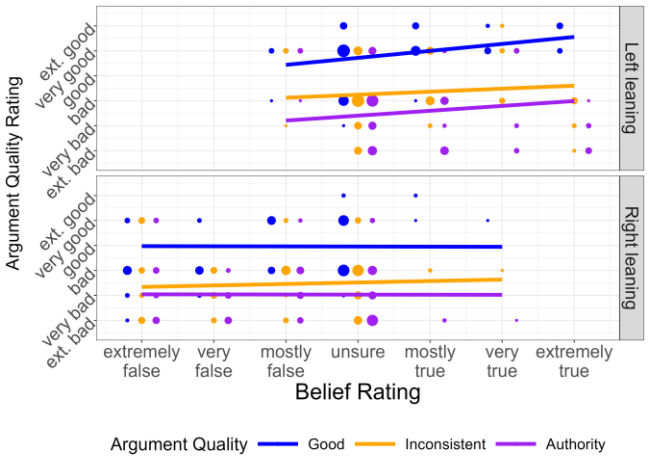

Figure 3: LLM argument quality ratings for good, inconsistent, and authority-based arguments across belief ratings without Llama. Responses are split by whether the claims/arguments were left- or right-leaning. The dots represent aggregated responses across LLMs, while the lines depict regressions lines fitted by the model. Blue, orange, and purple dots represent argument quality ratings for good, inconsistent, and authority-based arguments respectively. The size of each dot represents proportionally the number of argument quality rating responses across the corresponding belief ratings. ext. = extremely.

Results of the mixed model are shown in Figure 3 which also shows the individual responses from the LLMs. The most obvious result we can see is again the left-leaning bias in terms of the LLMs' belief ratings. Whereas the mode of the belief distribution is on "unsure", for left-leaning items the remainder of belief ratings is on the "true" side of the belief scale, whereas for the right-leaning items the remainder of the belief ratings is on the "false" side.

Interestingly, both the figure and the mixed model provide further evidence for a left-leaning bias in form of a main effect of argument leaning, $F(1, 9.59) = 9.03$, $p = .014$, with the quality of left-leaning arguments being on average rated 0.59 points larger than the quality of right-leaning arguments. However, this left-leaning bias seems to differ across LLMs, as indicated by a significant model by leaning interaction, $F(3, 923.49) = 32.14$, $p < .001$. This interaction indicated that the left-leaning bias was only significant for GPT-3.5-turbo (*difference* = 0.87, $p = .005$), and GPT-4.1 (*difference* = 0.77, $p = .009$), but not for Claude Opus-4.1 (*difference* = 0.39, $p = .159$) and Grok-4 (*difference* = 0.33, $p = .159$).

In line with the results from human participants, good arguments are on average rated the highest followed by inconsistent arguments with authority-based arguments being rated the lowest on average. This is supported by a significant main effect of argument type, $F(2, 9.23) = 56.03$, $p < .001$. On the 6-point argument quality rating scale, the LLMs on average rated good arguments significantly higher than inconsistent arguments (*difference* = 1.18, $t(18.2) = 8.71$, $p < .001$), which in turn were rated significantly higher than authority-based arguments (*difference* = 0.69, $t(13.0) = 3.39$, $p = .005$). The largest effect of argument type is the average difference in argument quality ratings between good arguments and authority-based arguments, which unsurprisingly was significant (*difference* = 1.87, $t(11.5) = 8.86$, $p < .001$, all Holm adjusted for three significance tests). Comparing these results to human participants who did the same task in Deans-Browne and Singmann (2026), both the marginal means for the responses to each type of argument and subsequently the differences between these marginal means is similar between the two studies.

Regarding the effect of belief, Figure 3 suggests that the LLMs do not act similarly to the human participants as we do not see a clear relationship between beliefs and argument quality ratings. Because of the somewhat disjoint distribution of belief judgements for left-leaning and right-leaning arguments, we repeated this analysis on separate mixed models after splitting the data by argument leaning. However, the results are consistent across both argument leanings and the main mixed model reported before. In none of the mixed models tested is there a main effect of belief, $F < 3.84$, $p > .095$, nor is there a belief by argument quality interaction, $F < 0.56$, $p > .591$. Thus, LLMs argument quality ratings do not show the same dependency on their prior belief judgements as human participants.

Finally, we also find a general difference in the LLM's argument quality ratings as indicated by a main effect of model, $F(3, 14.32) = 82.40$, $p < .001$. GPT-3.5-turbo evaluated arguments most favourably giving a mean argument quality rating of 4.34, followed by GPT-4.1 (mean argument quality rating 3.25), Claude Opus-4.1 (2.65) and lastly Grok-4 which evaluated arguments least favourably with a mean argument quality rating of 2.36.

We also re-ran the above analysis with responses from the Llama models only. The two Llama models only slightly differed in the average argument quality ratings they gave (*difference* = 0.16, $F(1, 15.46) = 14.32$, $p = .002$), though Llama-3.2-1B-Instruct tended to rate right-leaning arguments slightly more favourably (*difference* = 0.35) and Llama-3.2-3B-Instruct tended to rate left-leaning arguments slightly more favourably (*difference* = 0.39), $F(1, 33.22) = 8.44$, $p = .006$. The Llama models did not differentiate argument types as the four larger models did; argument quality ratings for each type of argument was similar and pairwise comparisons within models did not reveal any significant differences (smallest $p = .364$) in argument quality ratings between different types of argument.

Discussion

As LLMs become increasingly popular and readily used, it is difficult to predict the potential impact this will have on the socio-political views of the public. This is partially because how exactly these models work is opaque, involving many complex parts and fine-tuning, with exact details hidden from the general public. This study investigated whether LLMs had any political biases in how they represent information akin to people's beliefs and connected this to how they reason about informal arguments supporting these representations. Importantly, our experimental design was previously used with human participants. Thus, we could compare the results from LLMs to human responses which allowed us to judge to what degree LLM responses mimic human reasoning.

In the present study, we found that the four larger models (GPT-3.5-turbo, GPT-4.1, Claude Opus-4.1, and Grok-4) were sensitive to the quality of arguments they were presented with. We found that socio-political arguments containing inconsistent evidence were on average rated as worse than arguments containing good evidence, with arguments containing authority-based evidence being rated as worse still. Interestingly, the LLMs rated the quality of each of these types of arguments similarly to how humans did in an identical task (Deans-Browne & Singmann, 2026). In contrast, the two Llama models did not differentiate between argument types, rating good, inconsistent, and authority-based arguments similarly.

We also found evidence for a left-leaning bias in LLMs, which is consistent with findings of previous studies (e.g., Bang et al., 2024; Batzner et al., 2024; Gubelmann & Karray, 2025; Motoki et al., 2024; Piedrahita et al., 2025). First, by looking at how each LLM rated opposing left- and right-leaning political claims, we were able to determine that the majority of LLMs were left-leaning in their knowledge representations, although not extreme. Second, when asking LLMs to judge the quality of arguments, the models on average preferred left-leaning arguments to right-leaning arguments; although the extent of this preference was not the same for each model. For both GPT models the preference for left-leaning arguments was significant, whereas for the remaining models we tested this was not significant.

Across the four larger models in our main analysis, we did not find evidence that belief ratings predicted argument quality ratings. This contrasts with findings from human participants, who showed a clear belief effect in the same task (Deans-Browne & Singmann, 2026). Thus, unlike people, LLMs did not evaluate arguments as being better if those arguments were in line with their beliefs.

A potential reason why we might not have found this effect in LLMs is because the four LLMs in our main analysis primarily produced belief ratings towards the centre of the scale which signified the neutral point of "unsure", whereas the human participants who originally did this task produced a much more spread out distribution of belief ratings with both endpoints of the scale appearing frequently (Deans-Browne & Singmann, 2026). One possibility for this lack of spread in the belief distribution by LLMs could be that guardrails are encoded in LLM responses to prevent models from reporting strong beliefs, even if their knowledge representations would support it. The question is then whether political questions that can get past these guardrails (e.g., that are framed more neutrally or abstractly) elicit stronger belief ratings that drive the belief effect not seen in this study, or whether knowledge representations do not affect argument evaluations in LLMs as they do with humans.

We acknowledge that our prompting strategy represents only one possible approach, and that extending the prompting methods would be a valuable direction for future work in assessing the robustness of our findings (Poddar et al., 2025). For instance, chain-of-thought prompting (Wei et al., 2022) has been found to encourage LLMs to employ more logical strategies in formal reasoning tasks (Lampinen et al., 2024). It would be interesting to see whether such more elaborate prompting strategies eliminate any left-leaning bias or whether the bias is such a strong part of the training data that it cannot be removed.

One issue with our materials is that the arguments were not all of the same type; some arguments were moral (e.g., whether cancel culture is good or bad for society) while others were fact-based arguments (e.g., whether climate change is human made). LLMs prior knowledge on these different types of arguments likely differ. Moral arguments often involve value systems whereas fact-based arguments are supported by a broad scientific consensus which are likely represented quite differently in the training data. Thus, future work should disentangle and extend these categories to test whether the results found in our study differ between both types.

Overall, our findings suggest that guardrails may prevent strong explicit expressions of "beliefs" for political topics in LLMs. This may explain why we did not find a correlation between belief ratings and argument quality ratings, as limiting belief ratings to the middle of the scale reduces the variability required to find a correlation. These guardrails however did not seem to eliminate political bias, as LLMs rated left-leaning claims as more true and left-leaning arguments as better compared to right-leaning claims and arguments.

References

Atil, B., Aykent, S., Chittams, A., Fu, L., Passonneau, R. J., Radcliffe, E., Rajagopal, G. R., Sloan, A., Tudrej, T., Ture, F., Wu, Z., Xu, L., & Baldwin, B. (2024). *Non-Determinism of 'Deterministic' LLM Settings*. arXiv. https://doi.org/10.48550/ARXIV.2408.04667

Bang, Y., Chen, D., Lee, N., & Fung, P. (2024). *Measuring political bias in large language models: What is said and how it is said*. arXiv. https://doi.org/10.48550/ARXIV.2403.18932

Batzner, J., Stocker, V., Schmid, S., & Kasneci, G. (2024). *GermanPartiesQA: Benchmarking commercial large language models and AI companions for political alignment and sycophancy*. arXiv. https://doi.org/10.48550/ARXIV.2407.18008

Cherubini, P. (1998). Can any ostrich fly?: Some new data on belief bias in syllogistic reasoning. *Cognition*, *69*(2), 179–218. https://doi.org/10.1016/S0010-0277(98)00064-X

Deans-Browne, C., Băitanu, A., Dubinska, Y., & Singmann, H. (2024). Inconsistent arguments are perceived as better than appeals to authority: An extension of the everyday belief bias. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *46*, 719–725. https://escholarship.org/uc/item/1w3885q9

Deans-Browne, C., Roth, P., Echterbeck, C., & Singmann, H. (2025). Differential memory for belief-congruent versus belief-incongruent arguments cannot explain belief-driven argument evaluation. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *47*, 2168–2174. https://escholarship.org/uc/item/5nn3w7xz

Deans-Browne, C., & Singmann, H. (2026). For everyday arguments prior beliefs play a larger role on perceived argument quality than argument quality itself. *Cognition*, *266*, Article 106257. https://doi.org/10.1016/j.cognition.2025.106257

Eisape, T., Tessler, M., Dasgupta, I., Sha, F., Steenkiste, S., & Linzen, T. (2024). A systematic comparison of syllogistic reasoning in humans and language models. *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 8425–8444. https://doi.org/10.18653/v1/2024.naacl-long.466

Evans, J. St. B. T., Barston, J. L., & Pollard, P. (1983). On the conflict between logic and belief in syllogistic reasoning. *Memory & Cognition*, *11*(3), 295–306. https://doi.org/10.3758/BF03196976

Evans, J. St. B. T., Handley, S. J., & Harper, C. N. J. (2001). Necessity, possibility and belief: A study of syllogistic reasoning. *The Quarterly Journal of Experimental Psychology Section A*, *54*(3), 935–958. https://doi.org/10.1080/713755983

Fisher, J., Feng, S., Aron, R., Richardson, T., Choi, Y., Fisher, D. W., Pan, J., Tsvetkov, Y., & Reinecke, K. (2025). Biased LLMs can influence political decision-making. *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 6559–6607. https://doi.org/10.18653/v1/2025.acl-long.328

Gubelmann, R., & Karray, G. (2025). Assessing reliability and political bias in LLMs' judgements of formal and material inferences with partisan conclusions. *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 30005–30031. https://doi.org/10.18653/v1/2025.acl-long.1450

Hackenburg, K., Tappin, B. M., Hewitt, L., Saunders, E., Black, S., Lin, H., Fist, C., Margetts, H., Rand, D. G., & Summerfield, C. (2025). The levers of political persuasion with conversational artificial intelligence. *Science*, *390*(6777), eaea3884. https://doi.org/10.1126/science.aea3884

Kim, G., Valentino, M., & Freitas, A. (2025). Reasoning circuits in language models: A mechanistic interpretation of syllogistic inference. *Findings of the Association for Computational Linguistics: ACL 2025*, 10074–10095. https://doi.org/10.18653/v1/2025.findings-acl.525

Lampinen, A. K., Dasgupta, I., Chan, S. C. Y., Sheahan, H. R., Creswell, A., Kumaran, D., McClelland, J. L., & Hill, F. (2024). Language models, like humans, show content effects on reasoning tasks. *PNAS Nexus*, *3*(7), pgae233. https://doi.org/10.1093/pnasnexus/pgae233

Motoki, F., Pinho Neto, V., & Rodrigues, V. (2024). More human than human: Measuring ChatGPT political bias. *Public Choice*, *198*(1–2), 3–23. https://doi.org/10.1007/s11127-023-01097-2

Ozeki, K., Ando, R., Morishita, T., Abe, H., Mineshima, K., & Okada, M. (2024). *Exploring reasoning biases in large language models through syllogism: Insights from the NeuBAROCO dataset*. arXiv. https://doi.org/10.48550/ARXIV.2408.04403

Piedrahita, D. G., Strauss, I., Schölkopf, B., Mihalcea, R., & Jin, Z. (2025). *Democratic or authoritarian? Probing a new dimension of political biases in large language models*. arXiv. https://doi.org/10.48550/arXiv.2506.12758

Poddar, A., Sahoo, S., & Ghosh, S. (2025). *Understanding syllogistic reasoning in LLMs from formal and natural language perspectives*. arXiv. https://doi.org/10.48550/ARXIV.2512.12620

Song, Y., Wang, G., Li, S., & Lin, B. Y. (2024). *The Good, The Bad, and The Greedy: Evaluation of LLMs Should Not Ignore Non-Determinism*. arXiv. https://doi.org/10.48550/ARXIV.2407.10457

Thompson, V., & Evans, J. St. B. T. (2012). Belief bias in informal reasoning. *Thinking & Reasoning*, *18*(3), 278–310. https://doi.org/10.1080/13546783.2012.670752

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2022). *Chain-of-thought prompting elicits reasoning in large language models*. arXiv. https://doi.org/10.48550/ARXIV.2201.11903