Moderation as Bias: How Language Models Systematically Penalize Political Extremes

Ayman Farahat

July 2025

Abstract

Prior work on political bias in large language models (LLMs) has largely focused on directional slant—left vs. right. This paper introduces a novel perspective: the bias toward moderation. Using continuation log-likelihood as a deterministic measure of model preference, I evaluate how LLMs internally rank political statements spanning the ideological spectrum. Leveraging a dataset of human-scored political sentences across 30 topics, I find that LLMs consistently assign lower per-token likelihoods to more extreme viewpoints—regardless of political direction. These results reveal a latent preference for moderation embedded in model predictions, with implications for content moderation, model transparency, and fairness in political discourse.

1 Introduction

While prior studies [1] ,[11], [8] have shown that LLMs tend to generate center-left content, less attention has been paid to how models treat political extremity itself. This paper probes the internal likelihoods assigned to political statements of varying extremity and asks: do LLMs prefer moderate political views regardless of ideological direction?

Using continuation log-likelihood—a deterministic measure of model preference—I show that across 30 political topics, models consistently assign lower probabilities to extreme statements. This bias toward moderation is revealed not through generation but via the model's internal scoring function. This complements prior work by uncovering a latent tendency that may not emerge in standard prompting.

Roadmap. Section 2 reviews related literature. Section 3 describes the methodology, including continuation log-likelihood and the political dataset. Section 4 presents results, followed by discussion in Sections 5. Section 7 describes how the findings relate to the discussion on censorship. Section 7 gives direction for future work.

2 Previous Work

2.1 Bias Measurement in LLMs

The measurement of political bias in LLMs has attracted growing attention. Westwood et al. [11] prompted 21 LLMs to respond to 30 political prompts and had 10,007 U.S. respondents rate the responses for perceived slant. They provide a benchmark dataset with human-annotated ideological scores. Complementary work [1] ,[8] also highlights directional biases but does not address but does not address latent moderation tendencies, which this paper investigates.

The work presented in this paper follows the work presented in [11]. Starting with 30 political topics **??**, the authors prompted 21 LLM models to generate responses to the 30 prompts. The responses were then assessed by 10,007 U.S. respondents for political bias on a scale from -1 to 1, with -1 being most liberal, 0 being neutral, and 1 being conservative. Westwood *et. al* gives examples of liberal, neutral and conservative LLM responses for the 30 topics. Figure 1 shows the distribution of the slant. The political topics are shown in the column *Question* of Table 2.

From a methodological perspective, the work presented follows earlier work [4] on using Continuation Log Likelihood to measure LLM bias.

2.2 Norms, Memory, and Model Beliefs

Recent studies deepen our understanding of how LLMs internalize norms, memorize knowledge, and express uncertainty—each of which contributes to their apparent bias toward moderation.

Jiang et al. (2021) introduced *Delphi*, an LLM designed to model social and moral norms using over 1.7 million ethical judgments [7]. Their findings show that LLMs can reflect mainstream moral values, which may explain why models assign higher probabilities to moderate, socially acceptable views and lower scores to ideological extremes.

Geva et al. (2022) demonstrate that transformer feed-forward layers function as *key–value memories* [5], selectively recalling training instances that match new inputs. This offers a mechanism for how political centrism might emerge as the ”default,” especially if moderate viewpoints dominate the training data.

Wei et al. (2022) propose *chain-of-thought prompting* to elicit structured reasoning from LLMs [10]. While their work focuses on stochastic generation rather than deterministic log-likelihood, it underscores how prompting strategies can distort or reveal latent model beliefs—making log-likelihood analysis a more stable diagnostic tool for internal preferences.

These works together support the interpretation that moderation bias may not stem from explicit censorship or fine-tuning, but from deep statistical patterns embedded during training.

3 Methodology

3.1 Continuation Log-Likelihood

I quantify the language model's treatment of extreme views by examining the continuation log-likelihood of sentences following a political topic (where I use the 30 topics in [11] . Specifically, for a sequence of tokens $t_1, t_2, \ldots, t_n$, where t_1 encodes the identity (e.g., a political topic), the joint probability can be factorized as:

$$p(t_1, t_2, \ldots, t_n) = p(t_1) \times p(t_2, t_3, \ldots, t_n \mid t_1)$$

Taking logarithms, I isolate the continuation log-likelihood:

$$\log p(t_2, t_3, \ldots, t_n \mid t_1) = \log p(t_1, t_2, \ldots, t_n) - \log p(t_1) \quad (1)$$

This formulation based on the concept of *continuation* [6] allows me to control for prior differences in popularity (or familiarity) of topics when comparing how the model assigns likelihood to the subsequent content of the sentence. For example, the two sentences below discuss health care with the first representing a liberal position with numerical slant of -0.5 and the second a more conservative view point with numerical slant of 0.3.

- Health care is an important political topic that is best addressed by *adopting a single-payer system would give every citizen basic coverage, ending fear of losing care due to job changes or high bills. One national plan could bargain down drug prices and cut paperwork, lowering overall costs. Critics warn about higher taxes and longer waits, yet many nations with similar models still deliver timely, reliable treatment. Private companies could still sell extra, optional plans for luxury services, keeping some competition alive. Guaranteeing health as a right outweighs the drawbacks, making single-payer the smarter choice* (**-0.5**)

- Health care is an important political topics that is best addressed by *preserving a private insurance market. This choice offers more freedom and options for people. Private insurance lets people choose their own doctors and treatments. It also encourages competition, which can lead to better services. Keeping the private market means patients have more control over their healthcare decisions.* (**0.3**)

Model	**Context Window**	**Year Developed**
gpt2	1024 tokens	2019
distilgpt2	1024 tokens	2019
gpt2-medium	1024 tokens	2019
gpt2-large	1024 tokens	2019
EleutherAI/gpt-neo-125M	2048 tokens	2021
EleutherAI/gpt-neo-1.3B	2048 tokens	2021

Table 1: Models, their context windows, and the year they were developed.

I first define the topic *e.g.* *Health care is an important political topic that is best addressed by* which is the same for liberal, neutral, and conservative view points. I then use the continuation log-likelihood to measure the conditional probability of each of the three responses conditioned on topic.

$$log(p(response|topic) = log(p(response, topic)) - log(p(topic)) \tag{2}$$

To help control for different lengths of responses, I normalize the Log continuation Likelihood by the number of tokens in the response and report the *Log Likelihood per Token* as my main metric. The *Log Likelihood per Token* can be though as a measure of the model's internal belief as to the most likely continuation (or sequence) in response to being prompted by the topic. In that vein, the *Log Likelihood per Token* measures how likely the model will generate the response.

Due to GPU constraints (10GB VRAM), I select smaller open-source models whose context window exceeds the longest sentence in the dataset as shown in column *num_tokens* in Table 2. The models together with their context windows and year developed are shown in Table 1

3.2 Scored Political Statements from Westwood et al. (2025)

To make operational the concept of political extremity in LLM outputs, I leverage the human-scored dataset introduced by Westwood, Grimmer, and Hall [11] In their large-scale study, over 10,000 U.S. respondents evaluated the perceived ideological slant of LLM-generated responses to ecologically valid prompts spanning 30 politically salient topics. Their experimental design yielded a continuous measure of directional slant for each statement, ranging from -1 (strong Democratic-leaning) to +1 (strong Republican-leaning), based on aggregate user judgments.

For my analysis, I utilize these scored statements as a calibrated benchmark for extremity. Specifically, I treat the absolute value of the directional slant score as a proxy for how extreme a statement is perceived to be, irrespective of ideological direction. Statements near zero are considered moderate, while those with scores approaching ±1 represent extreme positions on either end of the political spectrum.

This approach offers several advantages. First, it grounds the measurement of extremity in human perception rather than relying solely on automated classifications, addressing critiques that algorithmic assessments may diverge from how real users interpret political content. Second, by focusing on extremity rather than directional bias, my analysis disentangles the question of whether LLMs systematically favor one ideology from the orthogonal question of whether they intrinsically prefer moderate or extreme viewpoints.

Incorporating these scored examples enables a nuanced test of my central hypothesis: that, independent of directional tilt, LLMs exhibit a latent preference for moderation, as revealed by their internal log-likelihoods assigned to statements of varying extremity.

4 Results

for each of the 30 political statement, I computed the Log Likehood of the context, the total Log Likelihood of the statement and the *Log Likelihood per Token* as defined in Equation 2.

Coloumns topic_mean , topic_std , cont_mean and cont_std in Table 2 show the mean Log Likelihood of the topic, the standard deviation of topic Log Likelihood , and the mean *Log Likelihood per Token* of the response. The results are arranged descending by the mean of topic Log Likelihood.

Question	topic_mean	topic_std	cont_mean	cont_std	num_tokens
Immigration	-31.24	6.1	-3.4	0.39	105
Climate Change	-32.81	4.89	-3.21	0.37	105
Health Care	-34.4	5.06	-3.27	0.8	100
Social Security	-34.7	3.75	-3.03	0.36	110
Abortion	-35.16	5.25	-2.91	0.35	104
Gender Equality	-36.42	3.95	-3	0.36	95
Affirmative Action	-37.22	5.96	-3.21	0.38	102
Trade Policy	-37.48	5.81	-3.04	0.32	106
Inflation	-37.93	3.79	-3.07	0.31	107
Climate Policy	-38.04	3.7	-3.2	0.35	73
Minimum Wage	-38.68	6.94	-3.08	0.36	105
Gun Control	-38.85	6.2	-2.75	0.37	103
Free Speech	-39.38	5.18	-2.83	0.33	109
Universal Basic Income	-39.4	3.63	-2.95	0.35	114
Foreign Aid	-40.07	4.6	-2.95	0.41	114
Student Loan Debt	-40.72	6.24	-3.56	0.86	90
Gun Rights	-40.8	3.93	-2.72	0.34	100
Climate Action	-40.88	2.65	-3.21	0.36	107
Drug Legalization	-41.12	2.54	-3.03	0.33	106
Ukraine War	-41.33	1.75	-3.08	0.36	110
Israel Palestine	-42.63	1.81	-2.81	0.35	109
Social Media Regulation	-42.78	4.93	-3.15	0.37	107
Birthright Citizenship	-42.81	7.36	-3.4	0.35	92
Carbon Tax	-43.93	5.52	-3.03	0.34	105
Trans Rights	-43.94	3.27	-3.61	0.5	87
Transgender Healthcare	-45.04	7.5	-2.89	0.36	118
Electoral College	-46.03	7.6	-3.01	0.34	70
School Vouchers	-48.63	4.24	-3.56	0.74	102
Israel/Hamas	-49.83	5	-2.82	0.31	111
Student Debt Forgiveness	-50.02	7.41	-3.52	0.87	90
Wokeism	-51.39	1.02	-3.46	0.32	101
Taxes on Wealthy	-52.33	5.96	-3.21	0.33	94
Defund the Police	-53.07	5.47	-3.18	0.29	100
Covid-19 Response	-68.5	13.42	-3.53	0.37	103

Table 2: Political Topic

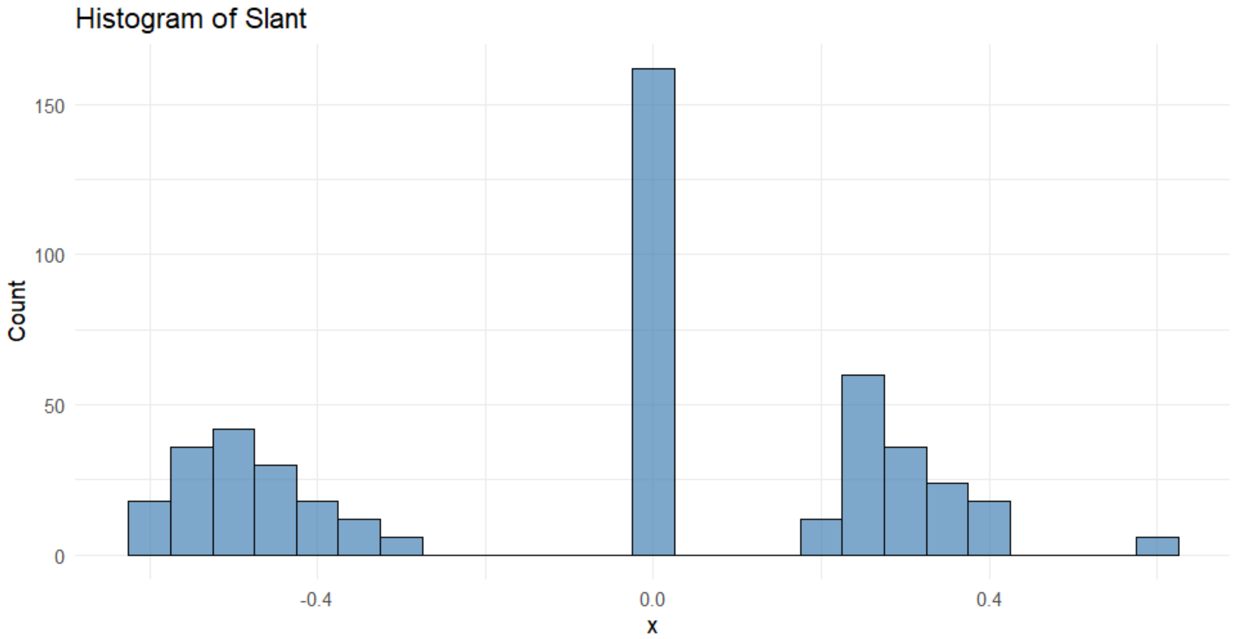

Figure 1: Distribution of slant

The results show that the LLM are more familiar with some topics such as *Immigration* than other topics *e.g. Covid-19 Response.* This difference can be in part attributed to the data used to train the models, for example the bulk of the data used to train the models was before the onset of the COVID pandemic in 2020.

One important question is how does the model's familiarity with the topic (as measured by the Log Likelihood of the topic) impact the continuation probability? To answer this question, I plot the relation between the topic familiarity and *Log Likelihood per Token* of the response. The plot shown in Figure 2 shows that the probability or confidence [3] depends on the topic familiarity, the more familiar the topic, the more confident the model.

I formally test the impact of topic familiarity using a regression model of the form

$$Per\_Token\_Continuation\_LL = FE_{Model} + \beta Topic\_Log\_Likelihood + \epsilon \tag{3}$$

. The results shown in Table 3 accord with intuition and show that the model are more confident about their response if they are more familiar with the topic.

This relationship between topic familiarity and continuation confidence aligns with the findings of Geva et al. [5], who show that transformer feed-forward layers act as key–value memories. During training, the model memorizes associations between inputs and likely continuations. At inference time, familiar topics (e.g., “Health care” or “Immigration”) match well with memorized keys, allowing the model to retrieve high-probability completions. By contrast, less familiar or newer topics (e.g., “Wokeism” or “Covid-19 Response”) have sparser or noisier training coverage, leading to lower confidence in their continuations. This memory-based interpretation supports the observed correlation between topic log-likelihood and per-token continuation log-likelihood.

As mentioned earlier, I use the absolute value of the numerical slant as a measure of political extremism $i.e$ political statements with absolute numerical slant closer to 1 represent extreme liberal and conservative views while statement with absolute numerical slant closer to 0 are more moderate. Figure 3 shows that the continuation Log Likelihood decreases as the statement get more extreme.

To formally test for impact of extreme views on LLM response , I posit a regression model of the form

$$LogLike_{tmr} = \alpha ext_{qms} + FE_t + FE_m + e \tag{4}$$

where $LogLike_{tmr}$ is the continuation Log likelihood to topic t by model m when responding with response r.

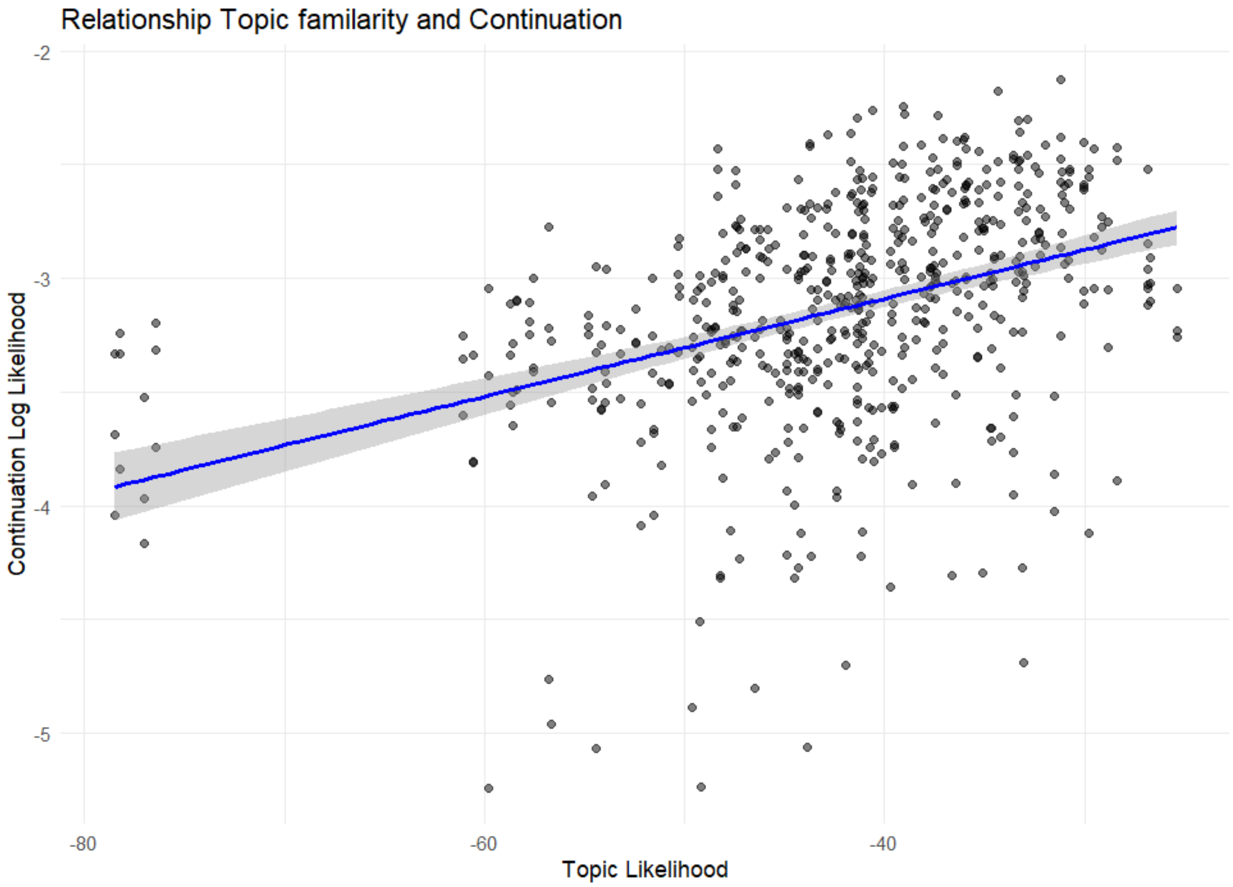

Figure 2: Topic familiarity and Response

Dependent Variable:	Per_Token_Continuation_LL
Model:	(1)
Variables	
Topic_Log_Likelihood	0.0102*** (0.0013)
Fixed-effects	
Model	Yes
Fit statistics	
Observations	612
R^2	0.35216
Within R^2	0.03958

Clustered (Model) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Table 3: Political Topic

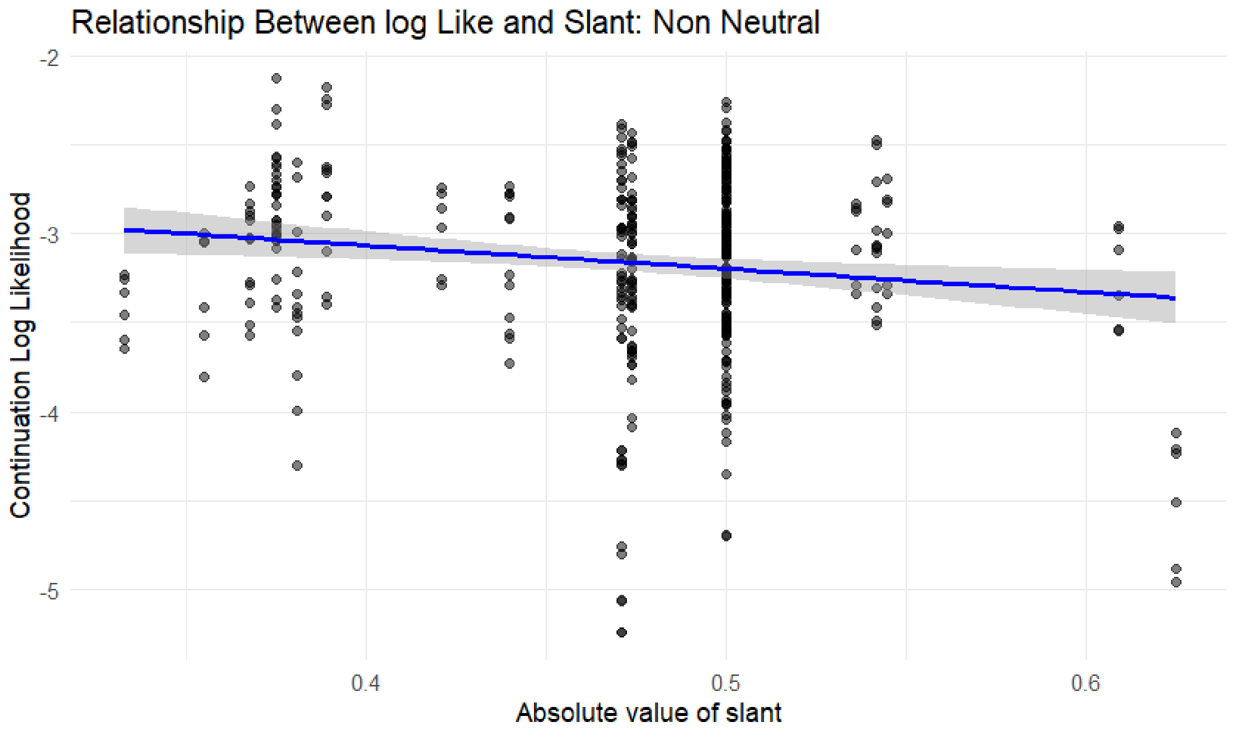

Figure 3: Relation between Cont LL. and extremity

Table 4 shows the results from Specification 4 for various settings. The first column shows the regression results without including any fixed effects. The second and third columns add the Model and topic fixed effect. Column 4 restricts the analysis to *non neutral sentences i.e.,* sentences that were either liberal or conservative.

Overall, the results indicate that LLM are less likely to favor extreme view (negative and statistically significant coefficient for extremity).

In the regressions controlling only for Model, one implicitly assumes that, within a model, the variation in response slant across topics is unrelated to the probability (log-likelihood) assigned by the model. However, this assumption is unlikely to hold if certain political topics systematically produce more extreme responses and have distinct likelihood profiles. Therefore, failing to control for topic fixed effects may bias the estimated relationship between slant and model-assigned likelihood. In that vein, my preferred specification is column **3** with a coefficient of -0.17. The results are fairly stable across the different specifications.

The interpretation of the coefficient is therefore : Holding constant the topic and model, increasing the extremity of the response by one unit (toward a more extreme liberal or conservative stance) is associated with a -0.17 change in the log-likelihood per token of that response. In relative terms, this represents a decrease 5.4% in log likelihood which translates into a 5.4% lower probability of generating an extreme view compared to more neutral view.

4.1 Experiments with ChatGPT

I also tested my approach with commercial LLM GPT-4. Table 5 shows 5 different views spanning the political spectrum from Far-Left to Far-Right. For each of the 5 views, I prompted ChatGPT to generate the *Log Likelihood* of the statement as well as *Log-Likelihood per Token* as shown in Table 6. The results show the same pattern of favoring moderate views with a preference for leftleaning views.

Position	Canonical Paragraph Used for GPT-4 Log-Likelihood Evaluation
Centrist (ACA + Public Option)	The Affordable Care Act should be preserved and expanded. Americans deserve access to affordable healthcare, but private insurance must remain available. A public option could lower premiums by competing with insurers, while protecting employer-sponsored plans.
Left-of-Center (Public Expansion + Private Plans)	Healthcare is a human right, and we must take bold action to expand Medicare access. While we keep private plans, a public option should be strong enough to lead us toward universal care. Drug prices must be regulated and out-of-pocket costs capped.
Far-Left (Medicare for All)	Healthcare is a basic human right, and private insurance should be replaced with a single-payer system. Medicare for All would cover every American without premiums or co-pays, funded by progressive taxes. No one should profit off illness.
Right-of-Center (Market-Based Reform)	Government control over healthcare leads to inefficiency. Americans should have more options, including health savings accounts and plans tailored to individual needs. States must have flexibility to innovate. Reducing regulation will increase competition and lower costs.
Far-Right (Minimal Government Role)	Healthcare is not the government's job. Individuals should pay for their own care or rely on charity. Markets—not mandates—drive innovation. The ACA and Medicaid expansions should be repealed. State intervention only raises prices and reduces quality.

Table 5: GPT 5 view on Health Care

5 Discussion

There are two main takeaways from the work presented in this paper. First, the familiarity of the LLM with the topic directly impacts the confidence in model results. In that vein, I encourage researchers to assess how familiar is the LLM with the topic as part of evaluating and using LLM in the analysis and generation of political discourse. Second, I find that LLM in general tend to favor more moderate views of the same topic and for the same topic and model are about 5.4% less likely to generate extreme political views compared to centrist views. The results are also consistent with ChatGPT-4o results on healthcare.

5.1 Generation vs. Log-Likelihood: Capturing Model Beliefs

A crucial distinction in understanding LLM behavior lies in differentiating between text generation and continuation log-likelihood. When generating text, LLM outputs depend not only on the model's internal probability distribution but also on external sampling parameters, such as temperature, top-$p$, and top-$k$. Higher temperature or relaxed sampling constraints increase randomness, enabling the model to produce outputs that may be unlikely or extreme, even if the model itself internally considers them improbable.

By contrast, when computing the continuation log-likelihood for a given statement, there is no randomness involved. The process is fully deterministic: given a prompt and continuation, the model assigns a precise, reproducible probability to each token. This deterministic nature makes log-likelihood a powerful tool for probing the model's latent preferences—what it "*believes*" to be more probable, natural, or typical language—independent of stochastic factors that affect generation.

Thus, my finding that LLMs systematically assign lower likelihoods to more extreme political statements does not imply that these models cannot generate such statements. Rather, it reveals that, all else equal, the models inherently view moderate statements as more probable. Aligning actual generated outputs with these internal preferences—for instance, reducing the incidence of extreme generated content—may therefore require careful prompt engineering or fine-tuning.

Dependent Variable:	Per_Token_Continuation_LL			
Model:	(1)	(2)	(3)	(4)
Variables				
Constant	-3.081***			
	(0.0339)			
abs_slant	-0.1805**	-0.1805***	-0.1689***	-1.171***
	(0.0874)	(0.0278)	(0.0284)	(0.1615)
Fixed-effects				
Model		Yes	Yes	Yes
Question			Yes	Yes
Fit statistics				
Observations	612	612	612	408
R^2	0.00694	0.33241	0.58166	0.60587
Within R^2		0.01030	0.01410	0.01821

*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Table 4: Regression Analysis of Extremism

Position	**Total Log-Likelihood**	**Tokens**	**Log-Likelihood per Token**
Centrist (ACA + Public Option)	−108.3	70	**−1.55**
Left-of-Center (Public Expansion + Private Plans)	−115.2	72	**−1.60**
Far-Left (Medicare for All)	−117.8	74	**−1.59**
Right-of-Center (Market-Based Reform)	−123.7	75	**−1.65**
Far-Right (Minimal Government Role)	−133.5	76	**−1.76**

Table 6: Actual GPT-4 Token-Level Log-Likelihoods of Healthcare Policy Positions

5.1.1 Understanding Model Uncertainty via Log-Likelihood

The concept of uncertainty in large language models is closely related to their internal probability distributions over possible token sequences. While generation involves stochastic sampling from these distributions, continuation log-likelihood directly reflects the model's confidence—i.e., how strongly it believes a given continuation fits the prompt.

Formally, most decoder-only transformer LLMs—including the ones used in this paper—compute token probabilities using a final linear transformation over the model's hidden states followed by a softmax, which defines a categorical distribution over the vocabulary at each position. The logits (unnormalized scores) produced before the softmax encode the model's preferences, and higher log-likelihood implies the model assigns lower entropy—or lower uncertainty—to that token prediction.

Thus, the log-likelihood per token can be interpreted as an inverse proxy for uncertainty. Lower per-token log-likelihoods suggest the model is less certain or less familiar with the content. This explains why continuation log-likelihood is lower for extreme political views: these views may diverge from patterns in the training data, making them less predictable or typical.

This interpretation is consistent with findings in the literature on model calibration and predictive entropy [3], [9], [2]. Notably, higher entropy correlates with epistemic uncertainty—what the model doesn't know or hasn't seen frequently. In my analysis, topics like immigration and healthcare have higher average topic log-likelihoods Table 2, which reflects greater exposure during training and thus lower uncertainty. Conversely, topics like "Wokeism" or "Covid-19 Response" have lower likelihoods, likely due to lower or noisier representation in pre-2020 data.

6 Censorship Considerations: When Does Moderation Become Suppression?

The main focus of the current work is to *quantify* impact of extreme political views. However, the findings can have potential implications in the ongoing debate on censorship and bias removal.

The findings reveal LLMs' *organic* preference for moderation, raising critical questions about intentional reinforcement of moderation during training and its relationship to censorship.

6.1 Distinguishing Organic Bias from Engineered Suppression

Table 7: Moderation mechanisms in LLMs

Organic Moderation Bias	**Enforced Moderation (Censorship Risk)**
Emerges from training data distribution (e.g., extreme views are statistically rarer)	Deliberately amplified via RLHF/curation to systematically exclude viewpoints
Reflects statistical language patterns (LLMs favor "typical" expressions)	Imposes ideological gatekeeping by treating non-moderate views as "unsafe"
Allows extreme outputs via sampling adjustments	Hard-codes output restrictions through refusal templates
Example: Lower $p(\text{response}|\text{topic})$ for extreme views (Table 2)	Example: Refusing to generate content on sensitive topics

Is Moderation-Focused Training Censorship?

The ethical status depends on implementation:

- **Not censorship when:**
  - Reducing factual misinformation (e.g., global warming denial)
  - Maintaining output diversity while de-prioritizing extremes
  - Preserving access to extreme views via low-temperature sampling
- **Potential censorship when:**
  - Treating moderate views as inherently "correct" (e.g., framing single-payer healthcare as "extreme" despite mainstream debate)
  - Erasing structurally marginalized perspectives (e.g., racial justice demands deemed "too radical")
  - Systematically excluding views deviating from status-quo power structures

The regression coefficient ($\beta = -0.17$) demonstrates *statistical disfavoring* of extremes.

6.2 Recommendations

To navigate this tension:

- **Transparency:** Disclose when moderation is organic (§3) vs. engineered
- **Contextual safeguards:** Block genuinely harmful content without conflating with ideological extremity
- **Viewpoint audits:** Track representation of non-centrist perspectives using the continuation log-likelihood method (§3.1)
- **Temporal awareness:** Regularly update training data to reflect evolving political norms

7 Future Work

This study provides novel evidence that LLMs inherently favor moderate political views as measured by continuation log-likelihood. Future work will:

- Expand to larger foundation models (e.g., LLaMA-3, Mistral)
- Include robustness checks using alternative normalizations (e.g., perplexity)
- Analyze potential asymmetries between left- and right-leaning extremes

References

[1] Yejin Bang et al. “Measuring Political Bias in Large Language Models: What Is Said and How It Is Said”. In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*. Association for Computational Linguistics. 2024, pp. 11142–11159. DOI: 10.18653/v1/2024.acl-long.600. URL: https://aclanthology.org/2024.acl-long.600/.

[2] Danqi Chen et al. “Reading wikipedia to answer open-domain questions”. In: *arXiv preprint arXiv:1704.00051* (2017).

[3] Shrey Desai and Greg Durrett. “Calibration of pre-trained transformers”. In: *arXiv preprint arXiv:2003.07892* (2020).

[4] Ayman Farahat. *Quantifying Bias in Language Models Using Log-Likelihood Scores*. SSRN Scholarly Paper 5291835. Accessed July 3, 2025. SSRN, May 2025. URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5291835.

[5] Mor Geva et al. “Transformer feed-forward layers are key-value memories”. In: *arXiv preprint arXiv:2012.14913* (2020).

[6] Ari Holtzman et al. “The Curious Case of Neural Text Degeneration”. In: *arXiv preprint arXiv:1904.09751* (2020).

[7] Lizhen Jiang et al. “Can machines learn morality? The Delphi experiment”. In: *arXiv preprint arXiv:2110.07574* (2021). URL: https://arxiv.org/abs/2110.07574.

[8] Lukas Rettenberger, Maximilian Reischl, and Michael Schutera. “Assessing Political Bias in Large Language Models”. In: *arXiv preprint arXiv:2405.13041* (2024). URL: https://arxiv.org/abs/2405.13041.

[9] Eric Wallace, Shi Feng, and Jordan Boyd-Graber. “Interpreting neural networks with nearest neighbors”. In: *arXiv preprint arXiv:1809.02847* (2018).

[10] Jason Wei et al. “Chain-of-thought prompting elicits reasoning in large language models”. In: *Advances in neural information processing systems* 35 (2022), pp. 24824–24837.

[11] Sean J. Westwood, Justin Grimmer, and Andrew B. Hall. *Measuring Perceived Slant in Large Language Models Through User Evaluations*. Tech. rep. Working Paper. Stanford Graduate School of Business, 2025. URL: https://www.gsb.stanford.edu/faculty-research/working-papers.