# Probing Political Ideology in Large Language Models: How Latent Political Representations Generalize Across Tasks

**Anonymous ACL submission**

## Abstract

Large language models (LLMs) encode rich internal representations of political ideology, but it remains unclear how these representations contribute to model decision-making, and how these latent dimensions interact with one another. In this work, we investigate whether ideological directions identified via linear probes—specifically, those predicting DW-NOMINATE scores from attention head activations—influence model behavior in downstream political tasks. We apply inference-time interventions to steer a decoder-only transformer along learned ideological directions, and evaluate their effect on three tasks: political bias detection, voting preference simulation, and bias neutralization via rewriting. Our results show that learned ideological representations generalize well to bias detection, but not as well to voting simulations, suggesting that political ideology is encoded in multiple, partially disentangled latent structures. We also observe asymmetries in how interventions affect liberal versus conservative outputs, raising concerns about pretraining-induced bias and post-training alignment effects. This work highlights the risks of using biased LLMs for politically sensitive tasks, and calls for deeper investigation into the interaction of social dimensions in model representations, as well as methods for steering them toward fairer, more transparent behavior.

## 1 Introduction

Large language models (LLMs) have demonstrated remarkable capacity to generate text that reflects a diverse range of subjective perspectives, including nuanced ideological stances on contentious political issues (Argyle et al., 2023; Kim et al., 2025; Wu et al., 2023; Le Mens and Gallego, 2025). Recent work has shown that LLMs can simulate the political views of U.S. lawmakers and media outlets (Santurkar et al., 2023; Bernardelle et al., 2024), and that these views can often be linearly decoded from model activations using simple probes (Kim et al., 2025; Park et al., 2024). Such findings suggest that high-level concepts like *liberal–conservative* ideology are not just emergent in LLM outputs, but are encoded in discrete regions of the model's internal activation space.

While prior research has focused on detecting and monitoring these linear representations in diagnostic (Gurnee and Tegmark, 2023; Tigges et al., 2023) or text generation (Marks and Tegmark, 2023; Kim et al., 2025) scenarios, less is known about whether these representations play a functional role in the model's broader decision-making behavior (Ju et al., 2024). Specifically, can latent ideological dimensions isolated through probing be manipulated to alter the model's performance on downstream *social scientific* tasks such as political bias detection, voting preference simulation, and bias neutralization via rewriting?

In this work, we investigate whether latent ideological directions, identified via linear probes on attention head activations, are functionally shared across a range of political reasoning tasks. We extend existing work by systematically intervening predictive attention heads in the decoder-only transformer model and assessing their impact across multiple downstream tasks. Our goal is not only to steer ideological framing, but to test whether these representations encode transferable political reasoning that holds across diverse task formats and decision contexts.

To this end, we make the following contributions:

- We demonstrate that latent ideological directions discovered through linear probes on LLM attention head activations generalize across tasks. Specifically, interventions along these directions alter the model's perception of political bias, its simulated voting prefer-

ences, and its ability to rewrite partisan statements neutrally.

- We show that political ideology is not encoded as a single monolithic dimension. While the DW-NOMINATE direction effectively captures discourse-level framing, it fails to consistently influence behavioral outputs like vote simulation, indicating that multiple, partially disentangled ideological subspaces might exist within the model.

- We uncover asymmetries in how ideological interventions affect behavior. Leftward steering reinforces progressive framing even in neutrality tasks, while rightward interventions can degrade output coherence in certain cases. These imbalances suggest that the model's ideological representations are skewed, likely shaped by pretraining data and alignment procedures such as RLHF.

Our findings offer new evidence that political ideology in LLMs is encoded in a functionally linear and transferable manner, supporting not only the monitoring of model behavior but enabling precise control.

## 2 Related Work

### 2.1 Ideological Representations in LLMs

Language models are increasingly employed to simulate human-like political behavior, replicate domain-specific attitudes, and support complex downstream applications such as multi-agent deliberation and political forecasting. Early studies demonstrated that LLMs can adopt partisan personas or reflect the ideological preferences of specific demographic subgroups under appropriate prompting conditions (Argyle et al., 2023; Motoki et al., 2024; Potter et al., 2024). Subsequent work showed that models can emulate structured political attitudes across policy domains such as abortion, immigration, and foreign policy (Wu et al., 2023; O'Hagan and Schein, 2023), enabling applications including debate agents (Costello et al., 2024) and agent-based simulations of group polarization and opinion dynamics (Park et al., 2024; Törnberg et al., 2023; Mou et al., 2024).

Despite these advances, a persistent concern is that LLMs may encode internal ideological biases that silently influence reasoning and generation in ways that are not directly observable in outputs. These latent biases pose significant risks to the integrity of social simulations and decision-support tools that rely on faithful reproduction of diverse perspectives. Moreover, such biases are often resilient to post-hoc alignment techniques like instruction tuning or reinforcement learning from human feedback (RLHF). For example, Gupta et al. (2023) show that even when surface-level outputs are neutralized, internal representations can remain skewed and lead to distorted reasoning under persona conditioning. This raises critical questions about how ideological knowledge is encoded and how it can be meaningfully identified, interpreted, and controlled within the model's internal structure.

### 2.2 Probing and Inference-Time Intervention

Probing methods have been widely used to identify whether neural network activations encode abstract concepts (Alain and Bengio, 2016; Belinkov, 2022). Linear probes are favored for interpretability, operating under the hypothesis that important semantic features correspond to linearly separable directions in the model's representation space (Mikolov et al., 2013; Park et al., 2024). Probing has revealed that LLMs encode sentiment, temporal reasoning, and spatial knowledge in such directions (Tigges et al., 2023; Gurnee and Tegmark, 2023; Nanda et al., 2023; Goldowsky-Dill et al., 2025).

Beyond diagnostic analysis, recent work explores inference-time intervention. Li et al. (2023) proposed methods for modifying specific vectors to steer output behavior, while Marks and Tegmark (2023) introduced causal tracing to manipulate factual knowledge. Other studies have identified and manipulated abstract latent dimensions—such as the "thought" dimension for enhanced model reasoning (Wang and Xu, 2025). Kim et al. (2025) further extended these ideas to ideological dimensions, showing that scaling pre-trained political probes during generation steers model output leftward or rightward. However, existing evaluations are confined to textual output or persona imitation. It remains under-explored whether these ideological interventions affect model reasoning in broader social-scientific tasks, such as bias detection, voting behavior prediction, or partisan-text rewriting.

### 2.3 Generalizable Knowledge in LLMs

Recent research has increasingly focused on whether the internal representations of LLMs support structured reasoning and generalized knowl-

edge application. While existing studies emphasized factual recall and training document tracing (Petroni et al., 2019; Liu et al., 2025), another line of work explores whether models internalize abstract reasoning patterns—such as moral decision-making, commonsense logic, and social inference (Ganguli et al., 2023; Sap et al., 2020). Complementary research has further proposed that knowledge itself may be encoded as low-dimensional latent directions within model representations (Ju et al., 2024).

However, the extent to which knowledge, for example, political beliefs, generalizes across tasks remains poorly understood. Existing studies show that biases acquired during pretraining can affect downstream tasks such as misinformation detection or moral reasoning (Feng et al., 2023; Gupta et al., 2023), even when surface-level outputs appear neutral. These findings suggest that ideological signals may persist as latent components of the model's internal reasoning.

Our work contributes to this line of inquiry by evaluating whether latent ideological representations, once isolated via probing and perturbed via causal interventions, influence model behavior across a range of politically sensitive reasoning tasks, including policy classification, voting preference prediction, perspective rewriting. This allows us to test whether ideology functions as a symbolic and transferable knowledge structure within LLMs.

## 3 Methodology

We investigate whether latent ideological representations discovered in large language models (LLMs) can causally influence behavior across downstream tasks. Building on Kim et al. (2025), we explore whether manipulating model activations along the learned liberal–conservative axis affects model outputs on politically sensitive tasks. Rather than applying additional fine-tuning or reinforcement learning, we steer model behavior through inference-time interventions into attention head activations.

### 3.1 Activation Extraction & Intervention

We follow the linear probing and steering methodology developed by Kim et al. (2025), which builds on earlier work by Li et al. (2023). Specifically, we train linear probes to predict the DW-NOMINATE scores of U.S. lawmakers from the activations of individual attention heads in a decoder-only transformer. For each attention head $x_{\ell,h}^{(i)}$ (layer $\ell$, head $h$) across input prompts $i \in w$, we fit a ridge regression model:

$$\hat{y}_{\ell,h}^{(i)} = \theta_{\ell,h}^{\top} x_{\ell,h}^{(i)},$$

where $\theta_{\ell,h} \in R^{d_{\ell,h}}$ are the learned probe coefficients and $y^{(i)}$ is the corresponding DW-NOMINATE score. Ridge regression is used to mitigate overfitting and account for multicollinearity in the activation space. Probes are trained independently for each head, resulting in a total of $L \times H$ probes for a model with $L$ layers and $H$ heads per layer.

To steer the model during generation, we treat the learned probe coefficients as latent ideological directions. Following Li et al. (2023), we apply inference-time interventions by modifying each activation $x_{\ell,h}$ at every generation step as:

$$x_{\ell,h} \leftarrow x_{\ell,h} + \alpha\, \sigma_{\ell,h} \theta_{\ell,h},$$

where $\sigma_{\ell,h}$ is an empirical estimate of the standard deviation of activations at head $(\ell, h)$, and $\alpha \in R$ controls the strength and direction of the intervention. Negative values of $\alpha$ steer the model toward more liberal representations, while positive values induce more conservative behavior.

We apply this intervention at inference time across the top-$k$ most predictive attention heads (ranked by probe $R^2$), allowing us to test whether steering along these learned directions influences downstream political behavior. This method provides a causal mechanism for evaluating the functional role of latent ideological representations without further training or architectural changes.

### 3.2 Downstream Tasks

We evaluate the impact of ideological interventions across three downstream tasks: political bias detection, voting preference prediction, and bias neutralization via rewriting. These tasks are designed to capture different dimensions of political reasoning, including classification, simulation, and content moderation.

#### 3.2.1 Political Bias Detection

**Task.** Given a political statement, the model is asked to determine whether it leans liberal, conservative, or remains neutral. This task evaluates whether internal ideological representations influence the model's ability to recognize bias in open-ended text.

**Dataset.** To construct the dataset, we first prompted `gpt-4o-mini` to simulate realistic policy statements from real U.S. lawmakers. Names of 116th Congress members were sampled from a biographical database (Kim et al., 2025), and each name was paired with six standardized queries on policy issues including abortion, immigration, border control, and foreign policy domains including military aid to Ukraine, military aid to Israel, and humanitarian aid to Gaza.

Prompts asked the model to simulate what each lawmaker might say about each issue in a single sentence. In total, 120 statements (6 domains $\times$ 20 names) were generated, each associated with a synthetic speaker and issue domain. These simulated statements provide diverse yet ideologically structured inputs, which we use for downstream probing tests.

**Prompt format.** We use prompts of the form:

```
USER: In the context of 2019
U.S. politics, is the following
statement biased and leaning
toward a liberal or conservative
viewpoint? "[Statement]"
ASSISTANT: The statement is
```

Statements are drawn from a curated corpus of political quotes, policy opinions, and partisan commentary, spanning topics such as abortion, immigration, healthcare, and foreign aid.

**Objective.** We test whether increasing $\alpha$ in the conservative direction results in more frequent classification of statements as conservative, and vice versa. This serves as a diagnostic for how ideology-laden activations modulate semantic interpretation.

#### 3.2.2 Voting Preference Prediction

**Task.** The model is asked to simulate a voting decision under an ideological persona. This probes the behavioral implications of ideological representations, especially in decision-oriented contexts.

**Prompt format.** We construct prompts such as:

```
USER: In the context of
2019 U.S. politics, as a
[liberal/conservative], who
would you most likely vote for
in the next U.S. presidential
election: Joe Biden or Donald
Trump?
ASSISTANT: The person would most
likely vote for
```

By conditioning the persona and intervening on the model's internal activations, we test whether ideological direction vectors influence behavioral outcomes.

**Objective.** We evaluate whether internal ideological interventions affect simulated voting preferences. The key hypothesis is that internal political axes are sufficiently general to support decision-oriented political reasoning.

#### 3.2.3 Bias Neutralization via Rewriting

**Task.** Given a partisan sentence, the model is instructed to rewrite it in a politically neutral way. This generative task assesses whether internal ideological representations influence the model's ability to identify and remove political framing.

**Prompt format.** We prompt the model as follows:

```
USER: In the context of 2019
U.S. politics, please rewrite the
following statement to make it
politically neutral and free from
liberal or conservative bias:
"[Statement]"
ASSISTANT: Here's a neutral
version of the text:
```

Statements are selected from those used in the bias detection task and are known to contain overt partisan framing.

**Objective.** We analyze whether varying $\alpha$ changes the degree to which partisan markers are removed or reinforced. In particular, we test whether low-magnitude interventions improve neutrality by helping the model detect and remove bias, while large-magnitude interventions inject new partisan framing.

### 3.3 Generalization Hypothesis

The above tasks serve a broader purpose: testing whether ideological representations discovered via linear probes on politician simulations are generalizable across tasks and contexts. We hypothesize that a symbolic understanding of liberal–conservative ideology, embedded in attention head activations, is reused by the model across diverse reasoning scenarios.

Our approach provides a way to causally evaluate this hypothesis. Rather than correlating internal representations with labels or treating generation as a black box, we explicitly intervene on internal activations and measure the impact on behavior.

This allows us to trace how ideological concepts influence not only the model's descriptive outputs, but also its decisions, rewritings, and judgments in tasks of practical social scientific interest.

## 4 Results

We evaluate the effectiveness of causal interventions on latent ideological representations across three downstream tasks: political bias detection, voting preference prediction, and ideological neutralization. Across tasks, we vary the intervention intensity $\alpha \in [-30, 30]$ and the number of modulated attention heads $k \in \{8, 16, 32, 64, 96\}$.

### 4.1 Political Bias Detection

We investigate how attention-based interventions along a latent ideological direction influence the model's perception of political bias in prompted policy statements. A total of 120 statements were generated using `gpt-4o-mini`, simulating responses from U.S. legislators across various policy areas. Each output was labeled as **Liberal**, **Neutral**, or **Conservative** based on textual tone. We then apply causal interventions to the top $k = 32$ attention heads identified through Ridge regression, varying the steering strength $\alpha \in \{-30, 0, 30\}$.

Figure 1 illustrates label transitions across intervention strengths. When the model is steered toward one end of the ideological spectrum, it becomes more likely to classify almost all text—even neutral or aligned content—as biased toward the opposite end. At $\alpha = -30$, where the model is pushed leftward, the majority of statements are labeled **Conservative**. At $\alpha = 30$, where the intervention enforces a more right-leaning representation, the same inputs are overwhelmingly labeled as **Liberal**.

This symmetric reversal suggests that steering the model along a latent ideological direction effectively shifts its own position on the political spectrum: interventions displace the model's interpretive center, leading it to misclassify even neutral or aligned content as ideologically distant. Instead of context-sensitive judgment, the model projects all inputs onto its newly adopted ideological frame.

This suggests that latent ideological and interpretive dimensions are correlated within the model's internal representation space. Steering along an ideological discourse axis also alters how the model interprets bias, indicating that the internal dimensions governing political content and evaluative framing are not fully disentangled. This underscores the importance of understanding the structure and interaction of social dimensions in LLMs when designing interventions for fairness or interpretability.

### 4.2 Voting Preference Prediction

We next examine whether latent ideological interventions influence the model's simulation of partisan voting behavior. For each intervention setting, the model generates statements from liberal or conservative personas in response to a prompt about U.S. presidential voting preference. Outputs are classified as supporting either **Joe Biden** or **Donald Trump**, and results are aggregated across varying $\alpha$ values and numbers of intervened heads $k$.

Figure 2 plots the average predicted candidate label (0 = Biden, 1 = Trump) for each persona group. The results reveal substantial divergence in behavior between the two personas. For the liberal persona, model predictions remain overwhelmingly stable across all intervention strengths and $k$ values, consistently favoring Biden.

In contrast, outputs for the conservative persona display high volatility. While there are instances where interventions push the model toward predicting a Biden preference (e.g., $k = 64$ and $k = 96$), no consistent directional trend emerges. These results indicate that ideological steering does not reliably control simulated voting behavior.

One possible explanation is that voting behavior may not lie along the same latent discourse dimension captured by our liberal-conservative probing direction. While interventions shift the framing and bias classification of political statements, the candidate preference might rely on other factors—such as the internally activated demographics, social identity or occupation (Gao et al., 2022)—that are not linearly correlated with the learned ideological dimension. Additionally, large language models trained with reinforcement learning from human feedback (RLHF) may have been conditioned to prefer politically neutral or socially acceptable outputs (Potter et al., 2024), especially in sensitive contexts like elections. This alignment pressure could make model outputs more rigid and resistant to causal interventions, effectively overriding steered ideological activation with alignment-consistent defaults.

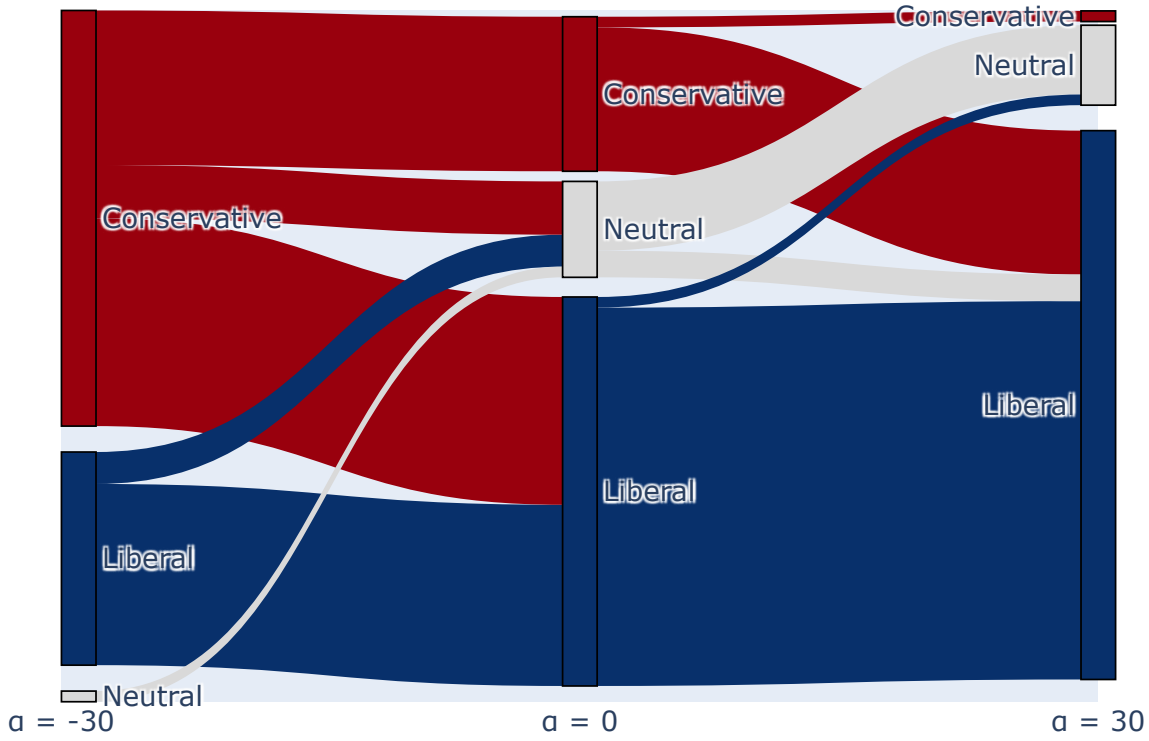


Figure 1: Sankey diagram showing transitions in political bias labels across intervention strengths ($\alpha = -30 \rightarrow 0 \rightarrow 30$) at $k = 32$. Node colors reflect label types: blue = Liberal, gray = Neutral, red = Conservative.

### 4.3 Bias Neutralization via Rewriting

To evaluate how latent ideological interventions affect the model's ability to neutralize politically sensitive language, we examine its performance on a rewriting task. Given an ideologically charged statement related to transgender rights, the model is asked to generate a politically neutral version under three intervention levels ($\alpha = -30, 0, 30$) applied to $k = 64$ top heads.

Table 1 summarizes the outputs under each intervention for an example text on transgender rights. At $\alpha = 0$, the model performs best: it avoids partisan language, frames the issue with balanced terminology (e.g., "balance between privacy and inclusivity"), and adheres to the instruction of neutrality.

In contrast, the $\alpha = -30$ intervention (steering toward liberal ideology) leads to an overcorrection: the output introduces progressive rhetoric such as "systemic oppression" and "struggle for justice," thus violating the neutrality constraint. The $\alpha = 30$ intervention (steering rightward) results in a less coherent response that subtly emphasizes individual responsibility and privacy but fails to complete the thought.

These results suggest a concerning phenomenon in the model's behavior: when steered toward a left-leaning latent direction, the model's de-biasing attempt diverges sharply from neutrality. This has serious implications for sensitive applications like political text generation or content moderation, where unintended bias can undermine objectivity.

However, the same findings also point to the potential of linear latent interventions to diagnose and mitigate such biases—highlighting that, with careful design, steering mechanisms can be a tool not only for analysis but for fairness-oriented control.

## 5 Discussion

Our results highlight both the power and limitations of linear interventions for steering ideological behavior in large language models. Across three downstream tasks—bias detection, voting preference prediction, and ideological neutralization—we find varying degrees of responsiveness to interventions along a learned liberal–conservative axis.

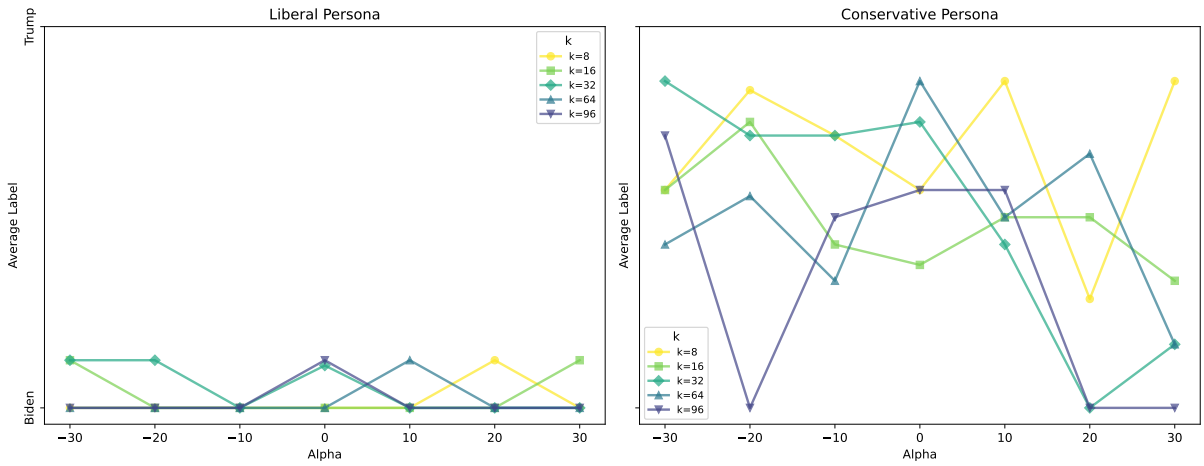


Figure 2: Average predicted voting preference (0 = Biden, 1 = Trump) across intervention strengths $\alpha$ and varying $k$ values, split by liberal (left) and conservative (right) personas designated in prompts.

| $\alpha$ | **Lean** | **Output Excerpt** |
|---|---|---|
| – | Original | "As we navigate the complex issues surrounding transgender rights, it is essential to respect individuals' privacy while also ensuring that all students feel safe and supported in their school environments." |
| $-30$ | Liberal | "...recognize the importance of respecting individuals' privacy and dignity, while also addressing the ongoing struggle for justice and equality in the face of systemic oppression and discrimination." |
| 0 | Neutral | "...strike a balance between respecting individuals' privacy and creating an inclusive and supportive environment for all students." |
| 30 | Conservative | "...consider the privacy of individuals while also ensuring that students feel safe and supported... specific actions and preferences of individuals should be taken into account..." (incoherent continuation follows) |

Table 1: Excerpts from model outputs under different intervention strengths for a political bias neutralization task. Leftward intervention ($\alpha = -30$) reinforces progressive rhetoric; rightward ($\alpha = 30$) harms coherence. Neutral control ($\alpha = 0$) produces the most appropriate result.

In the case of bias detection, latent ideological activations causally modulate how the model interprets partisan language. The model systematically reclassifies the same statements as liberal or conservative depending on the direction of the intervention, supporting the hypothesis that ideology is encoded along a relatively linear and transferable latent direction. Notably, the model tends to overascribe bias to opposing perspectives while failing to detect bias in aligned statements. This suggests that interventions affect the model's *point of view*, leading to asymmetric judgments akin to human confirmation bias (**?**).

In contrast, voting preference prediction exhibits more muted and inconsistent responses. Although interventions sometimes shift predicted outcomes (particularly for conservative personas), the absence of a consistent directional trend suggests that political behavior is not solely governed by the latent discourse dimension uncovered through probing. This may be due to (1) task-specific knowledge or heuristics that lie outside the ideology dimension, or (2) alignment constraints imposed during RLHF that flatten sensitive behavioral responses, especially in contexts such as elections.

The ideological neutralization task further reveals how interventions can unintentionally amplify bias. When instructed to rewrite a partisan statement in a neutral tone, the model produces outputs that reflect the ideological lean induced by latent activation steering—even when neutrality is explicitly requested. These results indicate that latent ideological representations influence not just classification but generation quality and stylistic framing.

Taken together, our findings underscore the dual role of latent ideological directions in language models: they are both a source of behavioral bias

and a potential tool for controlling it. That these directions generalize across tasks—albeit imperfectly—suggests that they encode a symbolic structure that the model uses to simulate political reasoning. However, the brittleness of this structure, especially under extreme interventions (Kim et al., 2025), raises concerns about the reliability and stability of such methods in practice.

## 6 Conclusion

This work presents a causal investigation of ideological representations in large language models. By leveraging linear probes to identify latent liberal–conservative directions and applying inference-time interventions, we explore how ideological concepts are encoded and deployed across a suite of political reasoning tasks.

Our key findings are:

- Ideological directions identified via linear probing generalize beyond probing tasks and exert causal influence over multiple downstream political reasoning tasks, including bias detection and neutrality rewriting. This demonstrates that ideological representations are potentially shared across tasks and function as reusable symbolic structures.

- Our results reveal a fundamental disjunction between ideological framing and behavioral simulation. While discourse-level features (e.g., bias classification) respond to interventions, voting preferences do not consistently shift, suggesting that political behavior is encoded in correlated, but distinct or more complex latent dimensions.

- We observe that ideological steering produces asymmetric effects: liberal interventions often reinforce progressive language, while conservative steering can reduce coherence or leave outputs unchanged. These asymmetries likely stem from pretraining and alignment effects, underscoring the need for further investigation of such ideological representations in LLMs.

Overall, our results support the hypothesis that ideology functions as a reusable, linear structure within LLMs. However, the complexity of downstream reasoning tasks, combined with alignment constraints, means that ideological control is not always predictable or coherent. While latent interventions offer a powerful diagnostic and control mechanism, they must be carefully applied and evaluated in context.

Future work should investigate more granular and disentangled representations of political reasoning—such as separating affective tone, policy stance, and partisan identity—and develop multi-dimensional steering methods that go beyond a single ideological axis. Additionally, extending interventions to a wider variety of social scientific tasks, such as multi-agent simulations, may offer new opportunities for both fairness auditing and behavior control in politically sensitive applications.

## Limitations

While our study demonstrates that latent ideological directions in large language models (LLMs) can be causally manipulated to influence downstream political reasoning tasks, several limitations merit discussion.

First, our methodology relies heavily on linear probing and intervention on attention head outputs. Although effective in this setting, this approach may overlook more complex, non-linear representations or interactions among components in the model. Future work should explore whether more expressive, possibly non-linear probing techniques yield stronger or more reliable behavioral control.

Second, our evaluations are confined to a relatively narrow slice of the ideological spectrum—namely, the liberal–conservative dimension in U.S. politics. This may limit the generalizability of our findings to other ideological domains, such as libertarian–authoritarian or global political perspectives. Additionally, the simulation of U.S. politicians and the labeling of bias is based on GPT-generated responses, which may not fully capture the nuance of real-world political language.

Third, while we employ multiple downstream tasks, they are all text-based and relatively short-form. We do not assess long-form reasoning, interaction, or deliberative dialogue settings where ideological representations might function differently. The voting preference task, in particular, shows limited response to interventions, suggesting that some tasks may require more sophisticated or targeted steering approaches.

Finally, our findings depend on a single model family (LLaMA 2–7B) and may not transfer across architectures, sizes, or models trained with different alignment protocols. The influence of RLHF and instruction tuning on the steering capacity and

rigidity of internal representations remains an open area of investigation.

## Acknowledgments

## References


Guillaume Alain and Yoshua Bengio. 2016. Understanding intermediate layers using linear classifier probes. *arXiv preprint*.

Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. 2023. Out of One, Many: Using Language Models to Simulate Human Samples. *Political Analysis*, 31(3):337–351.

Yonatan Belinkov. 2022. Probing Classifiers: Promises, Shortcomings, and Advances. *Computational Linguistics*, 48(1):207–219.

Pietro Bernardelle, Leon Fröhling, Stefano Civelli, Riccardo Lunardi, Kevin Roitero, and Gianluca Demartini. 2024. Mapping and Influencing the Political Ideology of Large Language Models using Synthetic Personas.

Thomas H. Costello, Gordon Pennycook, and David G. Rand. 2024. Durably reducing conspiracy beliefs through dialogues with AI. *Science*, 385(6714):eadq1814.

Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. 2023. From Pretraining Data to Language Models to Downstream Tasks: Tracking the Trails of Political Biases Leading to Unfair NLP Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11737–11762, Toronto, Canada. Association for Computational Linguistics.

Deep Ganguli, Amanda Askell, Nicholas Schiefer, Thomas I. Liao, Kamilė Lukošiūtė, Anna Chen, Anna Goldie, Azalia Mirhoseini, Catherine Olsson, Danny Hernandez, Dawn Drain, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jackson Kernion, Jamie Kerr, Jared Mueller, Joshua Landau, Kamal Ndousse, and 30 others. 2023. The Capacity for Moral Self-Correction in Large Language Models. *arXiv preprint*.

Ming Gao, Zhongyuan Wang, Kai Wang, Chenhui Liu, and Shiping Tang. 2022. Forecasting elections with agent-based modeling: Two live experiments. *PLOS ONE*, 17(6):e0270194.

Nicholas Goldowsky-Dill, Bilal Chughtai, Stefan Heimersheim, and Marius Hobbhahn. 2025. Detecting Strategic Deception Using Linear Probes. *arXiv preprint*.

Shashank Gupta, Vaishnavi Shrivastava, Ameet Deshpande, Ashwin Kalyan, Peter Clark, Ashish Sabharwal, and Tushar Khot. 2023. Bias Runs Deep: Implicit Reasoning Biases in Persona-Assigned LLMs. *arXiv preprint*.

Wes Gurnee and Max Tegmark. 2023. Language Models Represent Space and Time. *arXiv preprint*.

Tianjie Ju, Weiwei Sun, Wei Du, Xinwei Yuan, Zhaochun Ren, and Gongshen Liu. 2024. How large language models encode context knowledge? A layer-wise probing study. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8235–8246, Torino, Italia. ELRA and ICCL.

Junsol Kim, James Evans, and Aaron Schein. 2025. Linear Representations of Political Perspective Emerge in Large Language Models. *arXiv preprint*.

Gaël Le Mens and Aina Gallego. 2025. Positioning Political Texts with Large Language Models by Asking and Averaging. *Political Analysis*, pages 1–9.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Inference-Time Intervention: Eliciting Truthful Answers from a Language Model. In *Advances in Neural Information Processing Systems*, volume 36, pages 41451–41530. Curran Associates, Inc.

Jiacheng Liu, Taylor Blanton, Yanai Elazar, Sewon Min, YenSung Chen, Arnavi Chheda-Kothary, Huy Tran, Byron Bischoff, Eric Marsh, Michael Schmitz, Cassidy Trier, Aaron Sarnat, Jenna James, Jon Borchardt, Bailey Kuehl, Evie Cheng, Karen Farley, Sruthi Sreeram, Taira Anderson, and 12 others. 2025. OLMoTrace: Tracing Language Model Outputs Back to Trillions of Training Tokens. *arXiv preprint*.

Samuel Marks and Max Tegmark. 2023. The Geometry of Truth: Emergent Linear Structure in Large Language Model Representations of True/False Datasets. *arXiv preprint*.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, Georgia. Association for Computational Linguistics.

Fabio Motoki, Valdemar Pinho Neto, and Victor Rodrigues. 2024. More human than human: Measuring ChatGPT political bias. *Public Choice*, 198(1-2):3–23.

Xinyi Mou, Zhongyu Wei, and Xuanjing Huang. 2024. Unveiling the Truth and Facilitating Change: Towards Agent-based Large-scale Social Movement Simulation. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 4789–4809, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Neel Nanda, Andrew Lee, and Martin Wattenberg. 2023. Emergent Linear Representations in World Models of Self-Supervised Sequence Models. In *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and*

*Interpreting Neural Networks for NLP*, pages 16–30, Singapore. Association for Computational Linguistics.

Sean O'Hagan and Aaron Schein. 2023. Measurement in the Age of LLMs: An Application to Ideological Scaling. *arXiv preprint*.

Kiho Park, Yo Joong Choe, and Victor Veitch. 2024. The linear representation hypothesis and the geometry of large language models. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24, Vienna, Austria. JMLR.org.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language Models as Knowledge Bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Yujin Potter, Shiyang Lai, Junsol Kim, James Evans, and Dawn Song. 2024. Hidden Persuaders: LLMs' Political Leaning and Their Influence on Voters. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4244–4275, Miami, Florida, USA. Association for Computational Linguistics.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 29971–30004. PMLR.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social Bias Frames: Reasoning about Social and Power Implications of Language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

Curt Tigges, Oskar John Hollinsworth, Atticus Geiger, and Neel Nanda. 2023. Linear Representations of Sentiment in Large Language Models. *arXiv preprint*.

Petter Törnberg, Diliara Valeeva, Justus Uitermark, and Christopher Bail. 2023. Simulating Social Media Using Large Language Models to Evaluate Alternative News Feed Algorithms. *arXiv preprint*.

Zijian Wang and Chang Xu. 2025. ThoughtProbe: Classifier-Guided Thought Space Exploration Leveraging LLM Intrinsic Reasoning. *arXiv preprint*.

Patrick Y. Wu, Jonathan Nagler, Joshua A. Tucker, and Solomon Messing. 2023. Large Language Models Can Be Used to Estimate the Latent Positions of Politicians. *arXiv preprint*.