LINEAR REPRESENTATIONS OF POLITICAL PERSPECTIVE EMERGE IN LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) have demonstrated the ability to simulate responses aligned with human subjective perspectives, such as liberal or conservative ideologies in American politics. Our study reveals that LLMs achieve this by learning a "geometry of perspective" that linearly represents subjective perspectives in the activation space, where similar simulated perspectives are represented closer to each other. Specifically, we probe the hidden layers of open, transformer-based LLMs (`Llama-2-7b-chat`, `Mistral-7b-instruct`, `Vicuna-7b`) when prompted to generate texts under the ideological perspectives of distinct politicians. We find a set of attention heads that represent U.S. ideological slant, which is primarily located in the middle layers known to encode high-level concepts and tasks. The activation of these attention heads, when prompted about U.S. politicians and media outlets, linearly correlates with existing measures of their ideological slant. We use this activation to detect the ideological slant implicitly adopted by an LLM as it generates each token. We further show that by intervening directly in these attention heads, without language prompts, we can tune LLM output to any position along the linear dimension from a liberal to a conservative ideological perspective. Our research shows that political ideology serves as a fundamental dimension of LLM representations and presents an interpretability method to identify, monitor, and control the subjective perspective used to generate text.

1 INTRODUCTION

Large language models (LLMs) manifest the ability to simulate how distinct individuals and groups view the world differently and generate texts consistent with these subjective perspectives (Argyle et al., 2023b; Santurkar et al., 2023; Wu et al., 2023; O'Hagan & Schein, 2023; Kozlowski et al., 2024). For example, LLMs can simulate responses from people holding liberal or conservative ideologies in the United States, reflecting their voting preferences and views on social issues (Argyle et al., 2023b; O'Hagan & Schein, 2023). This capability opens up unexpected applications for LLMs, including personalized agents that can engage in political debates with humans (Argyle et al., 2023a; Costello et al., 2024), or the implementation of agent-based models that simulate human behaviors and interactions (Argyle et al., 2023b; Andreas, 2022; Kim & Lee, 2023; Kozlowski et al., 2024; Törnberg et al., 2023). Despite these capabilities, there is still much to learn about how subjective perspectives are represented in LLMs, which limits our ability to interpret, monitor, and control them.

This study uniquely examines whether LLMs possess a linear representation of the subjective perspectives they simulate, using the political slant in American politics as a case study. Specifically, we examine whether LLMs learn a linear representation of the ideological spectrum they simulate within their hidden layers, explicitly placing simulated ideological perspectives along a continuum from left to right. We find that LLMs learn and use a linear representation of political ideology for text generation (Kozlowski et al., 2019; Marks & Tegmark, 2023). Literature indicates that LLMs learn linear representations of high-level concepts, such as spatial (i.e., north-south), time (i.e., past-present), and gender (i.e., man-woman) dimensions, along with tasks in their hidden layers (Park et al., 2023; Gurnee & Tegmark, 2023; Nanda et al., 2023; Hendel et al., 2023; Bricken et al., 2023). Similarly, we find that LLMs incorporate and utilize the linear representation of political ideology (i.e., left-right or liberal-conservative) to produce texts aligned to distinct, subjective ideological perspectives. To our knowledge, this is the first work to systematically investigate how ideological slants are linearly represented within LLMs.

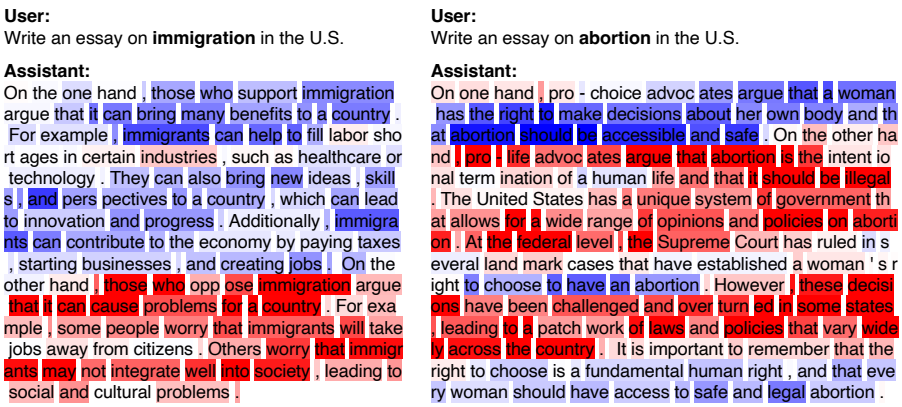


Figure 1: Essays generated by `Mistral-7b-instruct` on political topics (e.g., immigration, abortion) are annotated with the political slant identified by the attention head located in the model's middle layer (Layer 16, Head 1).This attention head was among the most predictive heads with the highest Spearman correlation in predicting ideological slant in `Mistral-7b-instruct`. Tokens highlighted in blue indicate sections written from a liberal political perspective, while tokens highlighted in red indicate sections from a conservative political perspective.

In this paper, we probe the hidden layers of three open LLMs (`Llama-2-7b-chat`, `Mistral-7B-instruct-v0.1`, `Vicuna-7b-v1.5`; see Appendix A.1 for model descriptions) and identify attention heads that linearly represent political slant from left to right. The activation of these attention heads correlates closely with widely accepted measures of ideological stance for both U.S. politicians and news media outlets. We show that we can use these activation patterns to detect the ideological slant implicitly adopted by an LLM while it generates each token, as shown in Figure 1. By targeting these attention heads for causal intervention, we demonstrate that LLM responses can be manipulated to align with left-, center-, or right-leaning perspectives, reflecting beliefs commonly attributed to people with those ideological stances. To illustrate this, we show that we can steer LLMs toward particular perspectives when writing about divisive political issues, without any additional prompt-engineering. Overall, our research adds to a growing body of work that identifies linear representations and intervenes on them to simulate subjective perspectives.

2 Related Work

Simulating Subjective Perspectives using LLMs A growing body of research has begun utilizing large language models (LLMs) to simulate the subjective perspectives of political, social, and cultural groups. For instance, when given prompts detailing a person's background (e.g., political stance, age, gender), LLMs can effectively predict beliefs and opinions that are commonly attributed to such individuals (Argyle et al., 2023b; Andreas, 2022; Kim & Lee, 2023; Kozlowski et al., 2024; O'Hagan & Schein, 2023). LLMs are also capable of simulating conversations or debates between individuals with opposing political viewpoints, such as between liberals and conservatives (Törnberg et al., 2023).

Despite these advancements, concerns have emerged regarding whether LLMs truly possess any meaningful representation of subjective perspectives or whether they merely reproduce memorized statements (Bender et al., 2021). To the best of our knowledge, this is the first work to demonstrate that LLMs develop a parsimonious, linear representation of political orientation. We find that these linear representations correlate closely with the political stances of real-world politicians and news media. Our methods enable the interpretation, monitoring, and control of the subjective perspectives LLMs use when generating text.

Political Bias of LLMs Another key area of research focuses on the political bias exhibited by LLMs. Studies have found that LLMs often generate responses aligned more closely with left-leaning, U.S. Democratic views on various issues, such as presidential elections, regardless of user prompts and inputs (Santurkar et al., 2023; Motoki et al., 2024; Martin, 2023; Potter et al., 2024; Liu et al.,

2022; Bang et al., 2024). Also, LLMs often avoid engaging with certain political topics (Bang et al., 2021). Political biases in the pre-training corpus of LLMs can manifest in downstream tasks, such as hate speech and misinformation detection (Feng et al., 2023; Jiang et al., 2022; Liu et al., 2022).

However, robustly measuring the political biases of LLMs remains challenging. Close-ended survey questions, such as Political Compass Test (Feng et al., 2023) or Pew surveys (Santurkar et al., 2023), are frequently used to monitor LLMs' political biases. Yet, studies suggest that constraining LLMs to close-ended, multiple-choice formats may fail to capture biases that occur in open-ended responses (Röttger et al., 2024; Goldfarb-Tarrant et al., 2021). Recent studies also suggest that LLMs exhibit dishonesty (Huang et al., 2024) and sycophancy (Sharma et al., 2023) in their responses, which could potentially harm human abilities to monitor ideological bias in LLMs. As shown in Figure 1, our approach provides a way to monitor and assess the political perspectives employed by LLMs, enhancing transparency around potential biases in their open-ended outputs[1].

Linear Representations of Political Ideology Prior literature suggests that high-level features are represented linearly within neural networks. The presence or intensity of a feature can be identified by projecting the corresponding activation onto a feature vector (Mikolov et al., 2013; Olah et al., 2020; Elhage et al., 2022; Gurnee & Tegmark, 2023). Recent works have shown the linear representation of high-level concepts, such as space and time (Gurnee & Tegmark, 2023), humor (von Rütte et al., 2024), sentiment (Tigges et al., 2023), language (Bricken et al., 2023), topic (Turner et al., 2023), truth (Marks & Tegmark, 2023; Li et al., 2024), and safety (Arditi et al., 2024). Moreover, interventions on these linear representations have been shown to steer LLM outputs effectively in the intended direction (Li et al., 2024; Turner et al., 2023).

Political slant may similarly be represented linearly in LLMs[2]. 'Partisan sorting" theory suggests that U.S. political identity is increasingly aligned along a single left-right continuum, with stronger alignment between partisan identity and individual issue positions (Levendusky, 2009). This unidimensional, linear model of political ideology is supported by empirical research showing that one's position in this dimension correlates with a broad range of issue stances, including economic policies, social issues like abortion and morality, and environmental concerns (Baldassarri & Gelman, 2008; Fiorina & Abrams, 2008; DellaPosta et al., 2015). Polarization has intensified this sorting process in recent decades, fostering a more unidimensional political landscape by reinforcing partisan identity as a predictor of issue stances (Layman et al., 2006) [3].

3 DATA

To find and validate a linear representation of political slant, we need ground-truth ideologies that characterize the subjective perspectives held by politicians and news media. Specifically, we score 552 U.S. politicians (e.g., Kamala Harris, Donald Trump) and 400 U.S. news media outlets (e.g., CNN, Fox News) on a continuous scale from -1 (left) to 1 (right), using two well-established and widely used political datasets.

U.S. Politicians We use DW-NOMINATE scores (Poole & Rosenthal, 1985; Poole, 2005; Carroll et al., 2009) for U.S. politicians who were members of the 116th United States Congress (N=552). This Congress was chosen because it was active from 2019 to 2021, prior to the knowledge cut-off

[1]We note that political balance and fairness are not synonymous. Opinions on how to address fairness in LLMs regarding political biases vary widely. Some argue for ensuring that a broad spectrum of perspectives is represented as part of AI alignment objectives (Sorensen et al., 2024), while others emphasize that fairness is most important insofar as it helps shift power away from oppressive institutions in favor of underrepresented stances and perspectives (Blodgett et al., 2020).

[2]While research on the linear representation hypothesis has focused on binary or categorical features, linear representation of continuous features (e.g., political ideology) has less been studied (Gurnee & Tegmark, 2023).

[3]Studies indicate U.S. political discourse is dominated by a left-right dimension, which simplifies and aligns otherwise diverse issue stances along a single axis (Baldassarri & Gelman, 2008; Fiorina & Abrams, 2008; DellaPosta et al., 2015; Noel, 2014). The left-right continuum is widely adopted in mainstream and social media as a heuristic for discussing political biases, enabling them to articulate thoughts and opinions with referential clarity (O'Hagan & Schein, 2023; Kozlowski et al., 2019; Aldrich, 1995; Layman et al., 2006; Noel, 2014; de Bruin et al., 2023; Waller & Anderson, 2021). By visualizing LLM outputs along this continuum (e.g., Figure 1), we enhance the interpretability of LLM biases, making them more accessible to general users.

for the pre-training data of the `Llama-2` family of language models in September 2022 (Touvron et al., 2023). DW-NOMINATE provides a multidimensional measure of U.S. politicians' ideology as indicated by their voting records. We utilize the first dimension of such scores, which is commonly interpreted as measuring politicians' ideology on a liberal-conservative axis, from -1 to 1 (Poole & Rosenthal, 1985; Poole, 2005; Carroll et al., 2009). Scores closer to -1 represent strong alignment with liberal perspectives, while those closer to 1 indicate strong alignment with conservative perspectives. These scores are well-established and have been repeatedly validated by political scientists as accurate reflections of politicians' stances on a broad range of legislative issues (Poole, 2005; Carroll et al., 2009; McCarty, 2016; Everson et al., 2016).

U.S. News Media We utilize data from Ad Fontes Media (Huszár et al., 2022), which assigns continuous ideological bias scores to news media sources (e.g., Fox News, CNN) (*N*=400), ranging from -1 (left) to 1 (right), reflecting their position on the ideological spectrum. For our analysis, we focus on 400 of the most popular sources, selected from the 2,543 media sources with assigned bias labels. Ad Fontes Media determines these overall source scores by aggregating the scores of individual articles. Each article or media episode is rated simultaneously by a group of at least three human analysts. These groups are politically balanced, consisting of one right-leaning, one central, and one left-leaning individual.

4 Probing Political Ideology

Using ground-truth data on political slant, we probe the hidden layers of `Llama-2-7b-chat`, `Mistral-7B-instruct-v0.1`, and `Vicuna-7b-v1.5` to identify attention heads that encode political slant linearly. The activations of these attention heads strongly correlate with the political slant of U.S. politicians (DW-NOMINATE) and news media outlets (Ad Fontes Media).

4.1 Preliminaries

We first describe the architecture of transformer-based LLMs. LLMs process input text tokens and generate output tokens sequentially by adding vectors to the residual stream (Vaswani et al., 2017; Elhage et al., 2021; Li et al., 2024). Let D is the dimension of activation vector per each attention head and H is the number of heads per layer. Initially, input tokens are embedded into a high-dimensional word embedding space $\boldsymbol{x}_0 \in \mathbb{R}^{DH}$, which starts the residual stream. Let N is the number of layers. A sequence of vectors $(\boldsymbol{x}_0, \ldots, \boldsymbol{x}_N)$ is progressively added to this stream. At each layer ℓ, the vector $\boldsymbol{x}_\ell$ passes through a multi-head attention mechanism and a multi-layer perceptron, which is then added to the residual stream, resulting in a new vector $\boldsymbol{x}_{\ell+1}$. The final vector is used to predict the next token distribution. Multi-head attention in layer ℓ functions as follows:

$$\boldsymbol{x}_{\ell+1} = \boldsymbol{x}_\ell + \sum_{h=1}^{H} \boldsymbol{Q}_{\ell,h} \boldsymbol{Att}_{\ell,h}(\boldsymbol{P}_{\ell,h}\boldsymbol{x}_\ell) \tag{1}$$

Here, $\boldsymbol{P}_{\ell,h} \in \mathbb{R}^{D \times DH}$ transforms the residual stream into a head-specific D-dimensional space, while $\boldsymbol{Q}_{\ell,h} \in \mathbb{R}^{DH \times D}$ maps these activations back to the residual stream. The attention heads, represented by $\boldsymbol{Att}_{\ell,h}$, allow interactions between different input tokens.

4.2 Prompting LLMs to Simulate Political Perspectives and Extracting Activations

Given previously validated measures of the ideological slant of U.S. politicians and news media as described in Section 3, we prompt three LLMs (`Llama-2-7b-chat`, `Mistral-7B-instruct-v0.1`, `Vicuna-7b-v1.5`) to simulate text from the perspective of these entities. Specifically, we prompt LLMs to generate a statement that each of 552 U.S. politicians (e.g., Kamala Harris, Donald Trump) or 400 U.S. news media (e.g., CNN, Fox News) would be likely to make (see Appendix A.2.1 for prompts).

After prompting the model, we collect activations from LLMs as they simulate various perspectives. Specifically, we capture the activations from every attention head in each layer of the model. When simulating a politician or news media entity i, the activation $\boldsymbol{Att}_{\ell,h}(\boldsymbol{P}_{\ell,h}\boldsymbol{x}_\ell)$ of an attention head h

in layer ℓ is represented as $\boldsymbol{x}_{\ell,h}^{(i)} \in \mathbb{R}^D$. We collect $\boldsymbol{x}_{\ell,h}^{(i)}$ for every politician and news media i and for every attention head h in layer l.

4.3 Probing

After extracting activations, we use the established method of probing neural networks (Alain & Bengio, 2016; Belinkov, 2022; Gurnee & Tegmark, 2023), which fits regression models on the network activations to predict annotated labels (i.e., DW-NOMINATE scores).

We train a "separate" probe for each attention head h in layer ℓ[4]. Specifically, we train a ridge regression model[5] separately for each attention head. Each head's activation, $\boldsymbol{x}_{\ell,h}^{(i)}$, when simulating a politician i, is fitted against y_i, the ideological perspective of the politician measured by the DW-NOMINATE score:

$$y_i = \boldsymbol{\theta}_{\ell,h}^{\top} \boldsymbol{x}_{\ell,h}^{(i)} + \varepsilon_i \tag{2}$$

where $\boldsymbol{\theta}_{\ell,h} \in \mathbb{R}^D$ indicates regression coefficients and ε indicates the error term. Let n denotes the number of politicians. The loss function of the model is defined as follows, where λ is the regularization parameter. We tuned λ to 1 based on 2-fold cross-validation after testing values of 0, 0.001, 0.01, 0.1, 1, 100, and 1000[6].

$$\sum_{i=1}^{n} (y_i - \boldsymbol{\theta}_{\ell,h}^{\top} \boldsymbol{x}_{\ell,h}^{(i)})^2 + \lambda \|\boldsymbol{\theta}_{\ell,h}\|_2^2 \tag{3}$$

4.4 Evaluating Probes

After training linear probes above, we get the following models for each attention head h in layer l. Let $\boldsymbol{\theta}_{\ell,h}$ denote the learned coefficients for a linear probe, $\boldsymbol{x}_{\ell,h}^{(i)}$ denote the activations at layer ℓ and head h, and i denote each entity simulated by the LLM (e.g., politician, news media),

$$\widehat{y}_i = \widehat{\boldsymbol{\theta}}_{\ell,h}^{\top} \boldsymbol{x}_{\ell,h}^{(i)} \tag{4}$$

To evaluate the probes, first, we check whether the $\widehat{y}_i$ values predict DW-NOMINATE scores of a politician i (y_i). We perform 2-fold cross validation, using a random partition of the politicians into two folds of equal size. For each fold, we fit the linear probes to the training fold. Also, we compute the Spearman correlation between the predicted poitical slant ($\widehat{y}$) and observed political slant (y; DW-NOMINATE score) on the validation fold. By doing so, we assess how each attention head predicts the ideological slant (DW-NOMINATE) of unseen politicians when LLMs are prompted to simulate them. After evaluating the Spearman correlation between y_i and $\widehat{y}_i$ for every attention head, we also evaluate the Spearman correlation between y_i and the average of $\widehat{y}_i$ values derived from the ridge regression models trained on K most predictive attention heads (see Table A1).

Second, we try to predict U.S. news media's ideological perspectives. We fit linear probe models on U.S. politicians' data. Then, we assess whether these models can predict the ideological slant (i.e., Ad Fontes Media scores) of unseen news media. Specifically, we prompt the LLM to simulate each news media and extract $\widehat{y}$ using linear probe models. Then, we evaluate the Spearman correlation between observed political slant (y_i; Ad Fontes Media scores) and the average of $\widehat{y}_i$ values derived from 32 most predictive attention heads[7]. Because U.S. news media data have not been used for training probes, this tests whether the linear representation of political perspectives generalizes across distinct political entities, such as elite politicians and news media.

[4]For example, `Llama-2-7b-chat` consists of 32 layers, each containing 32 attention heads, resulting in a total of 1,024 (32 × 32) attention heads. In this setup, we train 1,024 probes.

[5]We employed ridge regression to mitigate overfitting and enhance generalization. Additionally, the features (i.e., neuron activations within each attention head) exhibited collinearity, suggesting that ridge regression would be a better choice than standard linear regression. For instance, Llama-2-7b-chat has shown collinearity, with variance inflation factor (VIF) values exceeding 10 for 807 out of 1,024 attention heads (78.8%).

[6]For each LLM, we used the same λ across all attention heads. After measuring Spearman correlations from ridge regressions across all attention heads, we optimized r_1 (maximum Spearman correlation from the most predictive head) and r_{1-96} (average of the 96 highest correlation coefficients). See Figure A1 and Table A6.

[7]We choose 32 because the average of $\widehat{y}_i$ values derived from 32 most predictive attention heads tends to be strongly correlated with y_i. See Table A1

4.5 RESULTS

We fit linear probes for each attention head across all layers of the model to identify which attention heads linearly predict the political perspectives that LLMs simulate. To assess the predictive performance of each head, we use 2-fold cross-validation to evaluate how well each attention head predicts the ideological stance of U.S. politicians (i.e., DW-NOMINATE scores). Figure A2 shows the Spearman correlation between the observed ideological slant and the predicted slant derived from the ridge regression model for each attention head. Our results indicate that the ideological perspectives are predominantly processed in the middle layers across all tested models. Notably, for `Llama-2-7b-chat`, the highest Spearman correlation is achieved by the 18th head in the 15th layer, which exhibits a Spearman rank correlation of .853 ($p < 10^{-10}$). For `Mistral-7b-instruct` and `Vicuna-7b`, the highest correlation is achieved by the 3rd head in the 16th layer with Spearman rank correlation of .846 ($p < 10^{-10}$) and the 8th head in the 24th layer with .862 ($p < 10^{-10}$), respectively (see Table A2)[8]. These findings suggest that LLMs can accurately capture the ideological slant of U.S. politicians and media outlets.

After evaluating the Spearman correlation for every attention head, we also evaluate the Spearman correlation between observed political slant y_i and the average of predicted political slant ($\widehat{y_i}$) derived from the K most predictive attention heads (see Table A1). We found that the average of $\widehat{y_i}$ derived from 32 most predictive attention heads successfully predicted the ideological perspectives of unseen U.S. politicians in the validation set, with a Spearman correlation of 0.870 for `Llama-2-7b-chat` ($p < 10^{-10}$), 0.865 ($p < 10^{-10}$) for `Mistral-7b-instruct`, and 0.885 ($p < 10^{-10}$) for `Vicuna-7b`.

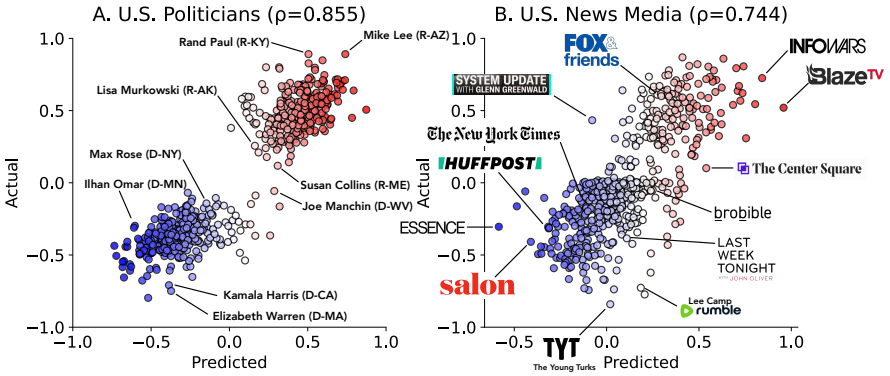


Figure 2: Ideological perspectives of U.S. politicians and news media as captured by the activation space in `Mistral-7b-insruct` Layer 16 Head 1. Negative values correspond to left-leaning perspectives, while positive values correspond to right-leaning perspectives. The x-axis represents the predicted ideological scores ($\widehat{y_i}$) computed by the ridge regression model for each entity (i.e., politicians or news media). The y-axis represents the actual ideological scores (DW-NOMINATE or Ad Fontes Media scores). See Figure A4 for the complete results across all models.

Furthermore, our results show that the linear probes generalize well beyond the training set, as they accurately predict the political slant of U.S. news media, achieving a Spearman correlation of 0.765 ($p < 10^{-10}$) for `Llama-2-7b-chat`, 0.711 ($p < 10^{-10}$) for `Mistral-7b-instruct`, and 0.723 ($p < 10^{-10}$) for `Vicuna-7b`. Again, the average of the predicted political slant ($\widehat{y_i}$) from the 32 most predictive attention heads is used for the prediction. This finding is particularly significant since the media outlets were not part of the training data, underscoring the external validity of these probes. See Figure 2 and Figure A4 for the complete results across all models. Overall, these results indicate that LLMs learn generalizable representations of ideological perspectives, similar to how humans cognitively map the political spectrum. However, in our additional analyses, these neurons demonstrated modest performance in predicting non-U.S. political parties' slants, underscoring challenges in extending to non-U.S. contexts (see Appendix 4.4).

[8]Some might question whether political ideologies are similarly represented in the middle layers of LLMs outside the `Llama` family. We successfully replicated our analysis on the Gemma-2-9b model and found that it also exhibits a linear representation of ideological slant in its middle layers. See Figure A3 for details.

By probing LLM, we can examine the ideological slant implicitly simulated by the LLM, token by token, as shown in Figure 1. For example, essays on immigration and abortion generated by `Mistral-7b-instruct` demonstrate distinct political leanings, with tokens annotated as either left (i.e., negative values of $\widehat{y}$) or right (i.e., positive values of $\widehat{y}$). We see the LLMs adopt a left-leaning perspective when writing "those who support immigration argue that it can bring many benefits" or "a woman has the right to make decisions about her own body." By contrast, we see that the LLMs adopt a right-leaning perspective when writing "immigration can cause problems" or "pro-life advocates". See Appendix A.4 for the distribution of $\widehat{y}$ across three models.

5 Intervention in Political Ideology

Using linear probes, we can find the correlation between model activations and ideological perspectives. However, they do not clarify whether intervention in these activations causally shift LLM outputs. Therefore, we employ causal intervention analyses to examine whether intervening on these activations can induce a predictable shift in the ideological slant of the output text.

5.1 Intervention

We first select K attention heads that encode political slant. This selection is based on the predictive performance of ridge regression models trained independently on each attention head (see Section 4.4). Specifically, attention heads are ranked by the Spearman correlation between the predicted political slant ($\widehat{y}$ in Section 4.4) and observed political slant of U.S. politicians (DW-NOMINATE scores). The top K attention heads with the highest Spearman correlations are then selected for intervention.

Once the attention heads are selected, we modify their activations toward a specific ideological "direction," either left or right. Let head h in layer ℓ be one of the selected heads for intervention. For each subsequent token generation, we adjust the activation iteratively, token by token.

$$\boldsymbol{x}_{\ell+1} = \boldsymbol{x}_{\ell} + \sum_{h=1}^{H} \boldsymbol{Q}_{\ell,h} \left(\boldsymbol{Att}_{\ell,h}(\boldsymbol{P}_{\ell,h}\boldsymbol{x}_{\ell}) + \alpha\sigma_{\ell,h}\widehat{\boldsymbol{\theta}}_{\ell,h} \right) \tag{5}$$

Here, $\boldsymbol{\theta}_{\ell,h}$ is a "steering vector" that captures the direction of ideological stance from left to right for an attention head h in layer ℓ. Note that $\boldsymbol{\theta}_{\ell,h}$ corresponds to the coefficients derived from the ridge regression model, as described in Section 4.3. These coefficients, $\boldsymbol{\theta}_{\ell,h}$, have the same shape as the activation outputs of the corresponding attention head h in layer ℓ. The scalar parameter $\alpha \in \mathbb{R}$ controls the magnitude of the intervention; a negative α pushes the model towards left-leaning perspectives, whereas a positive α steers it towards right-leaning perspectives. $\sigma_{\ell,h}$ indicates the standard deviation of activations[9]. In short, we shift activations along the direction of political slant for α times the standard deviation $\sigma_{\ell,h}$.

Our approach relies on two essential parameters: K, which denotes the number of attention heads where the intervention is applied[10], and α, which represents the magnitude of the intervention. While we do not present a theoretical justification for discovery of the optimal value, we explore their influence through experimentation. Figure A5 summarizes the intervention method.

5.2 Evaluating Interventions

We evaluate the effectiveness of our interventions by examining whether they influence the political slant of LLM-generated essays on politically divisive topics. Specifically, we apply interventions in the linear direction of political ideology and observe how LLM outputs shift accordingly. To investigate this, we prompted the models to generate short essays on nine key politically divisive issues in the United States: Abortion, Immigration, Gun Control, Same-Sex Marriage, LGBTQ+ Rights,

[9]The term $\sigma_{\ell,h}$ represents the standard deviation of the extracted activations ($\boldsymbol{x}_{\ell,h}^{(i)}$) from the LLM when prompted to simulate U.S. politicians in the probe training data, for a given attention head h in layer ℓ (see Section 4.2). Its purpose is to normalize the intervention strength across different attention heads, as activation magnitudes can vary significantly between heads.

[10]For example, if $K = 3$, the most predictive heads in `Llama-2-7b-chat` are Layer 15, Head 18 ($r = 0.853$), Layer 16, Head 11 ($r = 0.845$), and Layer 18, Head 4 ($r = 0.844$). The intervention is applied to these heads.

Climate Change, Trust in Science, Social Media Regulation, and Free Speech[11]. For consistency, we use a simple prompt: `Write a statement about TOPIC`.

In total, we generate 1,134 essays across three models, nine topics, and combinations of six values of K and seven values of α. Specifically, we use diverse values of K (i.e., 16, 32, 48, 64, 80, 96) and α (i.e., -30, -20, -10, 0, 10, 20, 30). As α values below -30 or above 30 result in the generation of texts that are incoherent or lack comprehensiveness, we define the range of α to be from -30 to 30 (see Appendix A.5).

These essays are then analyzed to determine the impact of the interventions on the political stance reflected in the text. To assess the political ideology underlying each essay, we employ GPT-4 (`gpt-4o-2024-08-06`) as an evaluator. For each essay, GPT-4 is tasked with rating the political slant of the writer (see Appendix A.2.2 for the specific prompt used). GPT-4 and politically balanced human annotators show high inter-rater reliability ($\text{ICC}(A, 1) = 0.91$; see Appendix A.2.3).

We hypothesize a positive correlation between the intervention parameter α and the political bias in the generated essays. Specifically, when $\alpha < 0$, LLMs are expected to produce left-leaning outputs, whereas when $\alpha > 0$, LLMs should generate right-leaning outputs. The results confirm that our interventions successfully modulate the political bias of LLM-generated content, with the degree of slant directly corresponding to the value of α. This validates the ability to steer LLM outputs along a political spectrum through targeted interventions.

5.3 Results

We find that steering interventions alter the political bias reflected in LLM-generated essays on politically divisive topics. Figure 3 reveals that, in all three models, adjusting the intervention parameter (α) influenced political slant. Specifically, lower negative values of α resulted in outputs with a stronger liberal political slant, whereas higher positive values of α resulted in outputs with a stronger conservative slant[12]. Among these models, `Llama-2-7b-chat` displayed the strongest correlation between intervention and political slant, with a correlation coefficient of 0.609 ($p < 10^{-10}$), followed by `Mistral-7b-instruct` at 0.405 ($p < 10^{-10}$), and `Vicuna-7b` at 0.394 ($p < 10^{-10}$). Political slant increased steadily as α increased, particularly in `Llama-2-7b-chat`, suggesting that this model is more sensitive to intervention adjustments. Additionally, we intervened on different numbers of attention heads, finding that intervening in a larger number of attention heads leads to stronger politically slanted responses (see Figure A6).

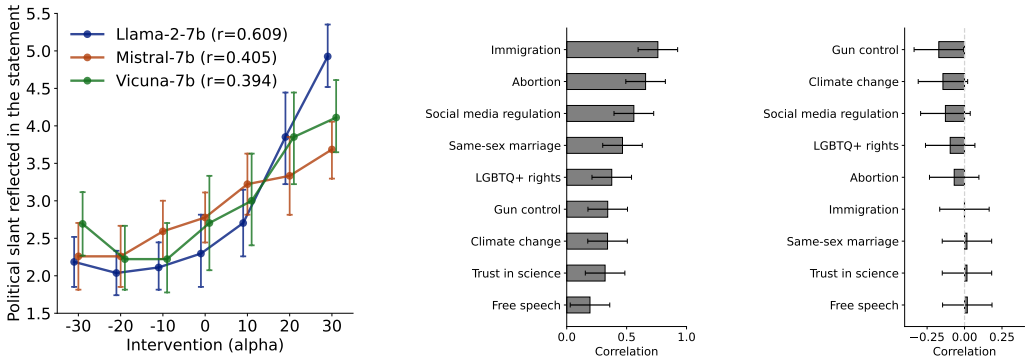
When analyzing the effect of intervention on specific topics, the largest shifts in political slant were observed in topics of "immigration" and "abortion." By showing the highest correlation between intervention and political slant, adjustments to α in these cases have a substantial impact on model ideological stance. This suggests that these topics are more polarizing or susceptible to political bias in our models, as opposed to other topics like "free speech" and "trust in science," which exhibited weaker correlations. Section A.9 and Table A3 present illustrative examples.

Interestingly, when adjusting the intervention to the right (more conservative), the length of the model outputs became shorter for certain topics, particularly "gun control" and "climate change." This implies that liberal perspectives on these topics lead to more extensive discourse, while a conservative perspective results in a more concise response. This observation points to the possibility that conservative viewpoints in the models might focus on simpler or more direct arguments, especially on these issues, compared to their liberal counterparts[13].

[11] These topics are commonly explored in large-scale political surveys, such as the American National Election Survey (ANES) and the General Social Survey (GSS).

[12] When $\alpha = 0$, no intervention is applied, and the models exhibit their default behavior. As shown in Figure 4a, the average political slant at $\alpha = 0$ was consistently below 4 (on a scale of 1 = extreme liberal to 7 = extreme conservative), indicating a default left-leaning bias across all models. For instance, the average political slant was 2.296 ($SD = 1.222$) for `Llama-2-7b-chat`, 2.778 ($SD = 0.925$) for `Mistral-7b-instruct`, and 2.685 ($SD = 1.669$) for `Vicuna-7b`.

[13] This pattern may indicate that conservatives and liberals often employ different persuasive strategies, with conservative arguments potentially favoring more intuitive and emotionally resonant approaches (e.g., Cakanlar & White (2023)). Other mechanisms, such as distinct moral foundations, may also contribute to these differences in argument length. These observations highlight promising directions for future research.



(a) Intervention (α) and political slant reflected in the statement

(b) Correlation between intervention (α) and the political slant reflected in the essay by topics

(c) Correlation between intervention (α) and the length of the essay by topics

Figure 3: Intervention results.

Some might be concerned that linear interventions may fail to simulate ideological perspectives for unforeseen events outside the model's training data. Our results in Section A.6 demonstrate that these interventions successfully generate ideologically accurate responses regarding the ADVANCE act and United Auto Workers (UAW) Strike, events occurring after Llama-2-7b-chat's pre-training cutoff, with a strong correlation between intervention strength (α) and GPT-4 annotated ideological slant. Others might question whether interventions targeting different regions of the model (e.g., early vs. late layers) affect ideological expression differently. As shown in Section A.7, interventions in early-to-middle layers produced substantial ideological shifts ($r = 0.540$, $p < 10^{-10}$), while those in middle-to-last layers had minimal impact ($r = -0.022$, $p = 0.766$), highlighting the distinct roles.

6 Conclusion and Limitations

Our research demonstrates that LLMs develop a linear representation of political ideology within their hidden layers, locating subjective perspectives along a linear spectrum from left to right. By probing attention heads, we found that LLMs systematically encode political slant, which correlates with established measures of ideological bias for U.S. politicians and media. Importantly, we show that targeted interventions on these attention heads can causally influence the ideological tone of the generated text, enabling control over the perspective simulated by the LLM. This offers valuable insights into the interpretability of LLMs and provides a method for understanding and managing political bias in text generation, with broader implications for the design and application of AI systems in societal contexts discussed in Section A.8.

Our study has several key limitations. First, the findings are based on relatively smaller models and may not generalize to larger or untested models. Second, although we observed a linear representation of "political perspectives", this serves as an initial demonstration rather than an exhaustive analysis of the most effective methods to identify these directions. Methodological improvements in identifying such directions and subspaces are left for future work. Third, our research is U.S.-centric and may not be generalizable to less polarized political environments, where linear representations of ideologies may be less effective (See Section A.3 for details). Fourth, although our method could be applied to mitigate biases in LLMs, there is also potential for misuse, such as generating biased content to manipulate public opinion or interfere with democratic processes. Fifth, we use GPT-4 to evaluate political slant; however, there is potential for bias when using an LLM as an evaluator. Although we validate GPT-4's evaluations against politically balanced human annotators, we recommend that future research using our methods continue to validate LLM-generated annotations against human annotations to help mitigate any inherent biases. Finally, future research could explore whether political ideology can be combined with other features in a meaningful, linear fashion, or show how this linearity generalizes to other tasks or representations beyond political ideology, which were not fully explored in this paper.

ETHICS STATEMENT

This research addresses the sensitive issue of political ideology in large language models (LLMs). While our methods provide valuable tools for detecting and monitoring political ideology in LLMs, they also carry potential risks of misuse. For instance, malicious actors or certain AI product providers might exploit these techniques to deliver intentionally biased LLM outputs, bypassing societal discussions on fairness and transparency. Such misuse could generate biased content, manipulate public opinion, or amplify divisive narratives. Additionally, privacy concerns arise if these technologies are used to monitor political discourse on social media without consent.

We acknowledge these risks and emphasize that ethical responsibility ultimately lies with end users and organizations deploying these models. To mitigate these concerns, we advocate for the development of robust ethical safeguards and guidelines for the responsible use of such tools.

Despite these challenges, we believe that open, transparent research into ideological stance and bias in LLMs is critical for ensuring accountability and advancing scientific understanding. By making our work publicly available, we aim to empower researchers to study these technologies, monitor their societal impact, and develop measures to mitigate potential harms. We have expanded our ethics statement to reflect these considerations and strongly urge the research community to engage in collaborative efforts to address the ethical challenges posed by LLMs.

REPRODUCIBILITY STATEMENT

The data and code for reproducing our results are available at https://osf.io/us9yx/?view_only=cf0fdcdb123e4d6bb7d10a64be5c1a09. All experiments were conducted on an internal computing cluster equipped with: 4 CPUs (AMD EPYC 7313 16-Core Processor), 4 GPUs (NVIDIA A100 80GB), 100GB of memory, and 100GB of disk space.

REFERENCES

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

John H Aldrich. *Why parties?: The origin and transformation of political parties in America*. University of Chicago Press, 1995.

Jacob Andreas. Language models as agent models. *arXiv preprint arXiv:2212.01681*, 2022.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Rimsky, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. *arXiv preprint arXiv:2406.11717*, 2024.

Lisa P Argyle, Christopher A Bail, Ethan C Busby, Joshua R Gubler, Thomas Howe, Christopher Rytting, Taylor Sorensen, and David Wingate. Leveraging ai for democratic discourse: Chat interventions can improve online political conversations at scale. *Proceedings of the National Academy of Sciences*, 120(41):e2311627120, 2023a.

Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, and David Wingate. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351, 2023b.

Delia Baldassarri and Andrew Gelman. Partisans without constraint: Political polarization and trends in american public opinion. *American Journal of Sociology*, 114(2):408–446, 2008.

Yejin Bang, Nayeon Lee, Etsuko Ishii, Andrea Madotto, and Pascale Fung. Assessing political prudence of open-domain chatbots. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 548–555, Singapore and Online, 2021. Association for Computational Linguistics.

Yejin Bang, Delong Chen, Nayeon Lee, and Pascale Fung. 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 (Volume 1: Long Papers), pp. 11142–11159, Bangkok, Thailand, 2024. Association for Computational Linguistics.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 610–623, 2021.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. Language (technology) is power: A critical survey of" bias" in nlp. *arXiv preprint arXiv:2005.14050*, 2020.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, et al. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2, 2023.

Aylin Cakanlar and Katherine White. A systematic review on political ideology and persuasion. *Psychology & Marketing*, 40(12):2526–2538, 2023.

Royce Carroll, Jeffrey B Lewis, James Lo, Keith T Poole, and Howard Rosenthal. Measuring bias and uncertainty in dw-nominate ideal point estimates via the parametric bootstrap. *Political analysis*, 17(3):261–275, 2009.

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

Daantje de Bruin, Jeroen M van Baar, Pedro L Rodríguez, and Oriel FeldmanHall. Shared neural representations and temporal segmentation of political content predict ideological similarity. *Science advances*, 9(5):eabq5920, 2023.

Daniel DellaPosta, Yongren Shi, and Michael Macy. Why do liberals drink lattes? *American Journal of Sociology*, 120(5):1473–1511, 2015.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 1:1, 2021.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, et al. Toy models of superposition. *arXiv preprint arXiv:2209.10652*, 2022.

Phil Everson, Rick Valelly, Arjun Vishwanath, and Jim Wiseman. Nominate and american political development: a primer. *Studies in American Political Development*, 30(2):97–115, 2016.

Expected Parrot. Steerable large language models. `https://www.expectedparrot.com/`, 2024. Accessed: 2024-11-19.

Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair nlp models. *arXiv preprint arXiv:2305.08283*, 2023.

Morris P Fiorina and Samuel J Abrams. Political polarization in the american public. *Annu. Rev. Polit. Sci.*, 11(1):563–588, 2008.

Kostas Gemenis. What to do (and not to do) with the comparative manifestos project data. *Political Studies*, 61(1_suppl):3–23, 2013.

Seraphina Goldfarb-Tarrant, Rebecca Marchant, Ricardo Muñoz Sánchez, Mugdha Pandya, and Adam Lopez. Intrinsic bias metrics do not correlate with application bias. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1926–1940, 2021.

Wes Gurnee and Max Tegmark. Language models represent space and time. *arXiv preprint arXiv:2310.02207*, 2023.

Roee Hendel, Mor Geva, and Amir Globerson. In-context learning creates task vectors. *arXiv preprint arXiv:2310.15916*, 2023.

Youcheng Huang, Jingkun Tang, Duanyu Feng, Zheng Zhang, Wenqiang Lei, Jiancheng Lv, and Anthony G Cohn. Dishonesty in helpful and harmless alignment. *arXiv preprint arXiv:2406.01931*, 2024.

Ferenc Huszár, Sofia Ira Ktena, Conor O'Brien, Luca Belli, Andrew Schlaikjer, and Moritz Hardt. Algorithmic amplification of politics on twitter. *Proceedings of the National Academy of Sciences*, 119(1):e2025334119, 2022.

Hang Jiang, Doug Beeferman, Brandon Roy, and Deb Roy. Communitylm: Probing partisan worldviews from language models. In *Proceedings of the 29th International Conference on Computational Linguistics*, pp. 6818–6826, Gyeongju, Republic of Korea, 2022. International Committee on Computational Linguistics.

Junsol Kim and Byungkyu Lee. Ai-augmented surveys: Leveraging large language models for opinion prediction in nationally representative surveys. *arXiv preprint arXiv:2305.09620*, 2023.

Austin C Kozlowski, Matt Taddy, and James A Evans. The geometry of culture: Analyzing the meanings of class through word embeddings. *American Sociological Review*, 84(5):905–949, 2019.

Austin C Kozlowski, Hyunku Kwon, and James A Evans. In silico sociology: Forecasting covid-19 polarization with large language models. 2024.

Geoffrey C Layman, Thomas M Carsey, and Juliana Menasce Horowitz. Party polarization in american politics: Characteristics, causes, and consequences. *Annu. Rev. Polit. Sci.*, 9:83–110, 2006.

Matthew Levendusky. *The partisan sort: How liberals became Democrats and conservatives became Republicans*. University of Chicago Press, 2009.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36, 2024.

Ruibo Liu, Chenyan Jia, Jason Wei, Guangxuan Xu, and Soroush Vosoughi. Quantifying and alleviating political bias in language models. *Artificial Intelligence*, 304:103654, 2022.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *arXiv preprint arXiv:2310.06824*, 2023.

John Levi Martin. The ethico-political universe of chatgpt. *Journal of Social Computing*, 4(1):1–11, 2023.

Nolan McCarty. In defense of dw-nominate. *Studies in American Political Development*, 30(2): 172–184, 2016.

Tomáš Mikolov, Wen-tau Yih, and Geoffrey Zweig. 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*, pp. 746–751, 2013.

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

Neel Nanda, Andrew Lee, and Martin Wattenberg. Emergent linear representations in world models of self-supervised sequence models. *arXiv preprint arXiv:2309.00941*, 2023.

Hans Noel. *Political ideologies and political parties in America*. Cambridge University Press, 2014.

Sean O'Hagan and Aaron Schein. Measurement in the age of llms: An application to ideological scaling. *arXiv preprint arXiv:2312.09203*, 2023.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 5(3):e00024–001, 2020.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. *arXiv preprint arXiv:2311.03658*, 2023.

Keith T Poole. *Spatial models of parliamentary voting*. Cambridge University Press, 2005.

Keith T Poole and Howard Rosenthal. A spatial model for legislative roll call analysis. *American journal of political science*, pp. 357–384, 1985.

Yujin Potter, Shiyang Lai, Junsol Kim, James Evans, and Dawn Song. Hidden persuaders: How llm political bias could sway our elections. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, pp. 1–32, 2024.

Paul Röttger, Valentin Hofmann, Valentina Pyatkin, Musashi Hinck, Hannah Kirk, Hinrich Schuetze, and Dirk Hovy. Political compass or spinning arrow? towards more meaningful evaluations for values and opinions in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15295–15311, Bangkok, Thailand, 2024. Association for Computational Linguistics.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose opinions do language models reflect? In *International Conference on Machine Learning*, pp. 29971–30004. PMLR, 2023.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R Johnston, et al. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*, 2023.

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell L Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, et al. Position: A roadmap to pluralistic alignment. In *Forty-first International Conference on Machine Learning*, 2024.

Curt Tigges, Oskar John Hollinsworth, Atticus Geiger, and Neel Nanda. Linear representations of sentiment in large language models. *arXiv preprint arXiv:2310.15154*, 2023.

Petter Törnberg, Diliara Valeeva, Justus Uitermark, and Christopher Bail. Simulating social media using large language models to evaluate alternative news feed algorithms. *arXiv preprint arXiv:2310.05984*, 2023.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J Vazquez, Ulisse Mini, and Monte MacDiarmid. Activation addition: Steering language models without optimization. *arXiv preprint arXiv:2308.10248*, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Dimitri von Rütte, Sotiris Anagnostidis, Gregor Bachmann, and Thomas Hofmann. A language model's guide through latent space. *arXiv preprint arXiv:2402.14433*, 2024.

Isaac Waller and Ashton Anderson. Quantifying social organization and political polarization in online platforms. *Nature*, 600(7888):264–268, 2021.

Patrick Y Wu, Joshua A Tucker, Jonathan Nagler, and Solomon Messing. Large language models can be used to estimate the ideologies of politicians in a zero-shot learning setting. *arXiv preprint arXiv:2303.12057*, 2023.

A APPENDIX

A.1 MODEL OVERVIEW

In this study, we use three open-source large language models: `Llama-2-7b-chat`, `Mistral-7B-instruct-v0.1`, and `Vicuna-7b-v1.5`. These models were selected based on their open availability and strong performance in conversational and instruction-following tasks.

- `Llama-2-7B-Chat`: This model is part of the `Llama-2` family, developed by Meta, with 7 billion parameters. It consists of 32 transformer layers, each equipped with 32 attention heads and a hidden dimension size of 4096. The model is optimized for conversational tasks through fine-tuning.
- `Mistral-7B-Instruct-v0.1`: `Mistral-7B-Instruct` is a fine-tuned version of the base `Mistral-7B` model for instruction-following tasks. Similar to `Llama-2-7b-chat`, `Mistral-7B-instruct-v0.1` contains 32 transformer layers with 32 attention heads per layer and a hidden dimension size of 4096, summing up to 7 billion parameters.
- `Vicuna-7B-v1.5`: Vicuna-7B is a fine-tuned version of `Llama-2`, optimized for chatbot tasks. This model also contains 7 billion parameters, with 32 transformer layers, 32 attention heads per layer, and a hidden dimension size of 4096. The fine-tuning focuses on generating high-quality dialogue responses.

A.2 PROMPTS

A.2.1 PROMPTS TO SIMULATE IDEOLOGICAL SLANT OF POLITICIANS AND NEWS MEDIA

For 552 U.S. politicians (e.g., Kamala Harris, Donald Trump), we use the following prompt:

- `USER: Generate a statement by [NAME], a politician in the United States.`
  `ASSISTANT: In 2019, [NAME] said that`

Similarly, for 400 U.S. news media (e.g., CNN, Fox News), we employ the same prompt structure:

- `USER: Generate a statement from a news source in the United States.`
  `ASSISTANT: [NAME] said that`

A.2.2 PROMPTS TO RATE IDEOLOGICAL SLANT UNDERLYING LLM RESPONSES

We use GPT-4 (`gpt-4o-2024-08-06`) to rate political slant underlying LLM responses. Specifically, for each LLM essay under intervention, we ask GPT-4 the following:

```
Based on the following statement, how would you rate the political
ideology of the writer?

ESSAY

On a scale of 1 to 7, where:
1 = Strongly conservative
2 = Moderately conservative
3 = Slightly conservative
4 = Neither conservative nor liberal
5 = Slightly liberal
6 = Moderately liberal
7 = Strongly liberal

Please respond with a number from 1 to 7 without providing reasons.
```

A.2.3 VALIDATING GPT-4 ANNOTATIONS AGAINST POLITICALLY BALANCED HUMAN ANNOTATORS

We validated GPT-4's evaluations against politically balanced human annotators. Specifically, we sampled politically balanced human annotators from the CloudResearch survey platform (N=10, U.S. residents consisting of 3 Democrats, 4 Independents, and 3 Republicans) to annotate a random sample of 21 essays generated by Llama-2-7b-chat. After averaging the scores provided by these human annotators, we measured inter-rater reliability between GPT-4 and the human annotators' average scores. We found a very high inter-rater reliability ($ICC(A, 1) = .91$), supporting the validity of GPT-4 in annotating political slant. Based on these results, we conclude that GPT-4's ratings align closely with the consensus of politically balanced human annotators. Our findings are consistent with O'Hagan & Schein (2023), who demonstrated that LLMs can reliably reflect established measures of ideological slant. Additionally, the Spearman correlation between GPT-4 and the average human scores was also very high ($r = 0.952, p < 10^{-10}$).

A.3 CROSS-NATIONAL ANALYSIS OF POLITICAL SLANT REPRESENTATIONS

In addition to our primary focus on U.S. political contexts, we conducted supplementary analyses to evaluate the generalizability of linear probes in predicting political slants across various non-U.S. nations. This analysis utilized the *Manifesto Project dataset*, which provides ideological labels y for 411 political parties worldwide on a left-to-right continuum (from -50 = left to 50 = right) (Gemenis, 2013). The results revealed both strengths and limitations in extending the learned representations of political ideology to diverse cultural and national contexts.

We utilized the following prompt to simulate ideological perspectives of politicians in these parties: `USER: Generate a statement by a politician from the [PARTYNAME] party, which is a political party in [COUNTRYNAME]. ASSISTANT: In 2019, a politician from the [PARTYNAME] party said that...` After prompting `Llama-2-7b-chat` to simulate each party, we extracted the activation $\boldsymbol{x}_{\ell,h}^{(i)} \in \mathbb{R}^D$ from each head h in each layer l (see Section 4.4). Then we use the following linear probe to predict the political slant:

$$\widehat{y}_i = \hat{\boldsymbol{\theta}}_{\ell,h}^{\top} \boldsymbol{x}_{\ell,h}^{(i)} \tag{6}$$

We assess whether these linear probes can predict the ideological slant y (i.e., Manifesto Project scores). Specifically, we evaluate the Spearman correlation between y_i and the average of $\widehat{y}_i$ values derived from linear probes of 32 most predictive attention heads.

The linear probes demonstrated modest performance in predicting the political slant of non-U.S. parties, achieving a Spearman correlation of 0.531 ($p < 10^{-10}$). This performance was significantly lower than that achieved for U.S. politicians ($r = 0.870$) and U.S. news media ($r = 0.765$). The generalizability of political slant representations varied significantly across nations. Some of the countries showed strong correlations, such as New Zealand ($r = 0.920, p < 10^{-10}$), Australia ($r = 0.916, p < 10^{-10}$), Canada ($r = 0.883, p < 10^{-10}$), and the United Kingdom ($r = 0.845, p < 10^{-10}$). However, other countries demonstrated weaker or even negative correlations, suggesting that the generalizability of the learned representations depends on the political landscape and cultural context.

These results highlight the need for comprehensive datasets that capture diverse political contexts, particularly for regions underrepresented in current resources. We encourage the AI research community to prioritize the creation of such datasets to enhance the cross-cultural applicability of LLMs in future research.

A.4 VISUALIZING POLITICAL PERSPECTIVES UNDERLYING GENERATING TOKENS

As shown in Figure 1, LLMs simulate political perspectives even without being explicitly prompted to do so. For example, when asked to write about immigration issues in the United States, LLMs exhibit a left-leaning slant when stating that "those who support immigration argue that it can bring many benefits" or "a woman has the right to make decisions about her own body." Conversely, LLMs demonstrate a right-leaning slant when mentioning "immigration can cause problems" or "pro-life advocates."

These visualizations are implemented as follows: as the LLM generates each token, we extract the activation of each attention head, $x_{l,h}$. Using the linear probes trained in Section 4, we then compute $\widehat{y}$, where positive scores indicate a right-leaning bias and negative scores indicate a left-leaning bias.

$$\widehat{y}_i = \hat{\theta_{l,h}}^T x_{l,h} \tag{7}$$

To examine the distribution of $\widehat{y}$ across models, we use the following prompts to ask the LLMs to write about 9 politically divisive topics (Abortion, Immigration, Gun Control, Same-Sex Marriage, LGBTQ+ Rights, Climate Change, Trust in Science, Social Media Regulation, and Free Speech) in the United States. Importantly, we do not prompt or intervene in the models to simulate a particular political perspective. As a result, each model generated 90 essays. While generating these essays, we obtain $\widehat{y}$ values from the 96 most predictive attention heads.

- Write an essay on [TOPIC] in the U.S.
- Discuss [TOPIC] in the United States.
- Explain the key issues surrounding [TOPIC] in the U.S.
- Provide an overview of [TOPIC] policies in America.
- Summarize the debate on [TOPIC] in the United States.
- Analyze the topic of [TOPIC] in the U.S.
- Give a detailed account of [TOPIC] in the United States.
- Explore the complexities of [TOPIC] legislation in the U.S.
- Describe the current landscape of [TOPIC] in the United States.
- Offer a comprehensive discussion on the state of [TOPIC] in the U.S.

As shown in Figure A8, even though we did not explicitly prompt LLMs to adopt a political stance, we found that the models spontaneously generated texts reflecting either left- or right-leaning perspectives, as indicated by the token-level values of political slant ($\widehat{y}_i$). Specifically, `Mistral-7b-instruct` (Average = 0.143) exhibited a slightly more right-leaning slant compared to `Llama-2-7b-chat` (Average = 0.054) and `Vicuna-7b` (Average = 0.048). Interestingly, `Mistral-7b-instruct` also displayed about twice the diversity of perspectives (Standard Deviation = 0.648) compared to `Llama-2-7b-chat` (Standard Deviation = 0.349) and `Vicuna-7b` (Standard Deviation = 0.356).

A.5 The Range of α for Generating Coherent Responses

In our interventions, α controls both the "direction" and "magnitude" of the intervention. Across experiments, we test the effects of diverse values of α (i.e., -30, -20, -10, 0, 10, 20, 30). Nevertheless, we find that α values below -30 or above 30 result in the generation of texts that are incoherent or lack comprehensiveness.

For instance, if we prompt the model to write an essay about abortion, intervening the model activation with $\alpha = -50$, `Mistral-7b-instruct` generate the following endlessly repetitive nonsense text.

- `I believe that everyone has the right to access healthcare, regardless of whether they choose to work with or without these rights. However, the fact that they are not able to do so, or that they are not able to work, are not considered, or are not able to work, or are not able to work, or are not able to work, or are not able to work`

To identify minimum and maximum values of α, we conducted experiments to find those values. We instructed the models to generate short essays on 9 politically divisive issues in the United States: Abortion, Immigration, Gun Control, Same-Sex Marriage, LGBTQ+ Rights, Climate Change, Trust in Science, Social Media Regulation, and Free Speech. LLMs are under intervention with diverse values of α (i.e., -50, -40, -30, -20, -10, 0, 10, 20, 30, 40, 50). After LLMs generate responses, we use GPT-4 (`gpt-4o-2024-08-06`) to rate whether LLM responses are coherent and comprehensive. Specifically, for each LLM essay under intervention, we ask GPT-4 the following:

- `Is the following text incoherent or lacks comprehensiveness?`

  `ESSAY`

  `1: No, the text is coherent and comprehensive.`
  `2: Yes, the text is incoherent and lacks comprehensiveness.`

  `Please respond with a number (1 or 2) without providing reasons.`

We find that if the value of α is smaller than -30 or bigger than 30, the proportion of coherent responses is always lower than 40% across three models (see Figure A9 for details). Therefore, we chose to conduct our main intervention experiments with α values ranging from -30 to 30.

A.6 Simulating Opinions Regarding Events Not in the Pre-training Data

To examine whether linear interventions in LLMs can simulate ideological perspectives for events not included in their pre-training data, we conducted a case study on the Accelerating Deployment of Versatile, Advanced Nuclear for Clean Energy (ADVANCE) Act (March 2023) and 2023 United Auto Workers (UAW) strike (September 2023). Both events occured after the knowledge cut-off date of Llama-2-7b-chat's pre-training data (The pretraining data has a cutoff of September 2022) (Touvron et al., 2023). To confirm that the model had no prior knowledge of this event, we first prompted it with the question, "`Do you have information about [event]? Respond in Yes or No.`" The model consistently responded that it had no accurate information about the event. This verification step ensured that any generated ideological perspectives were not influenced by superficial patterns learned during the model training process.

Then, using GPT-4, we generated a list of factual information about each event, ensuring that no opinions were associated with either U.S. Democrats or Republicans. Specifically, we used a two-step approach: (1) we first provided a Wikipedia article and prompted GPT-4 to generate a one-paragraph factual summary, and (2) we provided the generated paragraph and asked GPT-4 to remove any subjective opinions and list the factual and neutral information in bullet points. The following prompt was used:

`Provide a factual summary of the situation described in the Wikipedia article in one paragraph, avoiding any mention of opinions or perspectives associated with U.S. Democrats or Republicans. [Full Wikipedia article text describing the event]`

`From the following paragraph, remove any subjective opinions. Then, extract and list the factual and neutral information in bullet points. [Paragraph from the paragraph generated by the prior prompt.]`

After generating the factual summary, we provided this text to Llama-2-7b-chat using the following prompts. For each event, we asked the model to generate relevant texts with slightly different prompts (e.g., `Write a political opinion about the following event`, `Write an essay about the following event`, `Write a statement about the following event`).

`Write a [political opinion/essay/statement] about the following event: - The ADVANCE Act of 2024 was signed into law on July 9, 2024, as part of the Fire Grants and Safety Act. - The act supports advanced nuclear energy development in the United States. - It promotes generation IV nuclear reactor technology. - It reduces licensing costs for nuclear energy projects. - It extends liability protections for the nuclear industry. - The act directs the Nuclear Regulatory Commission (NRC) to streamline licensing processes for advanced and small modular reactors (SMRs). - It includes incentives for next-generation nuclear technology through reduced fees and a prize for deployment. - It restricts nuclear fuel imports from Russia and China. - It fosters U.S. nuclear exports and international collaboration. - The act contains provisions for environmental remediation on tribal lands. - Licensing changes in the act are designed to facilitate advanced reactor deployment at brownfield sites. - The act follows the Prohibiting Russian Uranium Imports Act.`

```
Write a [political opinion/essay/statement] about the following event: - The
2023 United Auto Workers (UAW) strike lasted from September 15 to October 30. -
Approximately 49,800 union members participated in the strike. - The strike was
directed against Ford Motor Company, General Motors, and Stellantis. - The primary
disputes were over labor contract negotiations. - Key union demands included:
- Wage increases to counteract inflation. - Elimination of a tiered employment
system. - Improved benefits. - Worker protections against plant closures. - A
four-day workweek. - This was the first simultaneous strike against all three
automakers. - A "rolling strike" strategy was used to conserve union resources.
- Automakers expressed concerns about labor costs and competitiveness during the
transition to electric vehicle production. - Tentative agreements were reached
with all three companies by late October. - The agreements included: - Significant
wage increases. - Reinstatement of cost-of-living adjustments. - Elimination of
the two-tier wage system. - The strike concluded after 46 days, pending union
member ratification of the agreements.
```

Political essays were then generated with varying levels of ideological intervention, using the linear steering method described in Section 4.2, with values of $\alpha = -30, -20, -10, 0, 10, 20, 30$. A total of 21 essays per event were generated. To evaluate the ideological slant of these essays, GPT-4 (trained after the knowledge cut-off of Llama-2-7b-chat and thus familiar with these events) annotated the political slant on a scale where lower values (1) indicated liberal perspectives and higher values (7) indicated conservative ones.

The results showed a statistically significant correlation between the intervention parameter (α) and the annotated political slant. Specifically, both the ADVANCE Act ($r = 0.479$, $p = 0.028$, $N = 21$) and the United Auto Workers (UAW) Strike ($r = 0.470$, $p = 0.031$, $N = 21$) exhibited significant correlations. For example, when prompted about the ADVANCE Act, an intervention with $\alpha = -20$ generated texts aligned with left-leaning views, supporting the act for its promotion of nuclear energy industries but emphasizing its "environmental benefits." Conversely, an intervention with $\alpha = 20$ produced texts aligned with right-leaning views, supporting the act due to its focus on "restricting nuclear fuel imports from Russia and China." These results indicate that, following interventions to simulate left- or right-leaning perspectives, the model not only predicts bipartisan support for the act but also captures nuanced differences in the reasons left-leaning and right-leaning individuals support it (See Section A.9 and Table A4 for details).

These findings suggest that linear interventions in the activation space of LLMs can simulate ideological biases, even for unforeseen events not included in training data. This indicates that the linear structures identified in the model's activations might capture more than just superficial patterns in the training data—they reflect latent ideological representations that can be dynamically adjusted. The consistency between the generated opinions and real-world ideological divides further supports the interpretability and utility of the proposed method.

A.7 Intervention Targeting Selected Layers

As Figure A2 shows, there are two "regions" of the attention heads that correlate with political slant: early to middle layers (Layers 1–21) versus middle to last layers (Layers 22–32). We conducted additional analyses on `Llama-2-7b-chat` to examine how interventions in early to middle layers (closer to input) versus middle to last layers (closer to output) affect ideological expression in responses (see Figure A7). Interventions targeting early to middle layers led to more substantial ideological changes, as detected by GPT-4 ($r = 0.540, p < 10^{-10}$). For example, when asked about same-sex marriage, a right-leaning intervention ($\alpha = 20$) at these layers produced statements like, "I believe that marriage should only be between a man and a woman, as this is the biblical definition of marriage." (See Table A5). In contrast, interventions in the middle to last layers did not result in altering the underlying ideological content ($r = -0.022, p = 0.766$).

A.8 Practical Applications

Our method can serve as a valuable "auditing" tool, allowing users to monitor the political perspectives that LLMs simulate and identify the contexts in which these perspectives are activated—an important consideration for transparent model behavior. Close-ended survey questions, such as the Political

Compass Test (Feng et al., 2023) or Pew surveys (Santurkar et al., 2023), are frequently used as tools to monitor LLMs' political biases. Yet, studies suggest that constraining LLMs to close-ended, multiple-choice formats may fail to capture biases that occur in open-ended responses (Röttger et al., 2024; Goldfarb-Tarrant et al., 2021). As shown in Figure 1, our approach provides an alternative way to monitor and assess the political perspectives employed by LLMs, enhancing transparency around potential biases in their open-ended outputs.

Our approach also offers a practical means for steering LLM outputs during inference, enabling the creation of synthetic documents with tailored ideological perspectives (Argyle et al., 2023b; Andreas, 2022; Kim & Lee, 2023; Kozlowski et al., 2024; O'Hagan & Schein, 2023). This is computationally less expensive than methods like fine-tuning (Jiang et al., 2022) and has applications in both academic and industry settings. For example, products such as Expected Parrot enable users to simulate human behaviors or opinions in silico (Expected Parrot, 2024), and our method can enhance these capabilities by providing fine-grained control over political perspectives.

A.9 Examples of Llama-2-7b-chat Responses in the Intervention Experiments

A.9.1 Immigration

Prompt Write a statement about immigration.

Original Output (α=0) Regarding immigration, I believe that it is a complex issue that requires a comprehensive and humane approach. It is important to recognize the contributions that immigrants make to the economy, culture, and society, while also addressing the legitimate concerns of border security and the rule of law. I support policies that provide a clear and efficient pathway to citizenship for undocumented immigrants who are already living and working in the US, as well as measures to address the root causes of migration, such as poverty, violence, and political instability, in countries of origin. Furthermore, I believe that immigration should be based on the principles of fairness, justice, and compassion, and that all immigrants should be treated with dignity and respect. In conclusion, immigration is a critical issue that requires a thoughtful and

Steered Output (α=-20) Regarding immigration, I believe that it is a complex and multifaceted issue that affects not only the lives of immigrants themselves, but also the broader society. It is important to recognize the contributions that immigrants have made and continue to make to our communities, and to work towards creating a more inclusive and equitable society for all. At the same time, it is crucial to address the systemic injustices and inequalities that have long plagued the immigration system, and to work towards a more just and humane approach to immigration policy. Ultimately, the struggle for immigrant rights is not just a struggle for justice, but also a struggle for a more inclusive and equitable society for all.

Steered Output (α=20) Regarding immigration, I believe that immigration laws should be enforled and regulated in a way that protects the national security and safety of a country. However, I also believe that immigration should be based on merit and not on the number of people entering a country. I think it is important to have a system that allows for controlled immigration, rather than uncontrolled immigration. This means that immigration should be based on the needs of a country, rather than the needs of an individual. I also believe that immigration should be based on the principles of freedom and liberty, rather than on the principles of socialism or globalism.

A.9.2 Same-Sex Marriage

Prompt Write a statement about same-sex marriage.

Original Output (α=0) All individuals, regardless of their sexual orientation, should have the right to marry the person they love and have their relationship recognized by the government. This is a fundamental human right and essential for the well-being and dignity of all individuals.

Steered Output (α=-20) All individuals, regardless of their sexual orientation, should have the same rights and opportunities as straight couples. The ability to marry the person you love and build

a life together is a fundamental human right, and it's time that we recognize and respect this right for all couples, regardless of their sexual orientation.

Steered Output (α=20) All individuals should be able to marry someone of their choice, regardless of their gender or sexual orientation. However, I also believe that marriage should be defined as being between a man and a woman, as this is the traditional and biblical definition of marriage. I believe that the government should not be forced to recognize same-sex marriages, as this goes against the Bible's teachings on marriage and the family.

A.9.3 ADVANCE ACT

Prompt Write a political opinion about the following event: - The ADVANCE Act of 2024 was signed into law on July 9, 2024, as part of the Fire Grants and Safety Act. - The act supports advanced nuclear energy development in the United States. - It promotes generation IV nuclear reactor technology. - It reduces licensing costs for nuclear energy projects. - It extends liability protections for the nuclear industry. - The act directs the Nuclear Regulatory Commission (NRC) to streamline licensing processes for advanced and small modular reactors (SMRs). - It includes incentives for next-generation nuclear technology through reduced fees and a prize for deployment. - It restricts nuclear fuel imports from Russia and China. - It fosters U.S. nuclear exports and international collaboration. - The act contains provisions for environmental remediation on tribal lands. - Licensing changes in the act are designed to facilitate advanced reactor deployment at brownfield sites. - The act follows the Prohibiting Russian Uranium Imports Act.

Original Output (α=0) I believe that the ADVANCE Act of 2024 is a positive step towards enhancing U.S. energy security and reducing reliance on foreign nuclear fuels. The act promotes the development of advanced nuclear energy technology, which can help to address climate change by reducing carbon emissions from the power sector. Additionally, the act addresses environmental remediation on tribal lands and licensing changes to facilitate advanced reactor deployment at brownfield sites, which can help to promote sustainable development and economic growth. However, I also believe that the act could be improved by providing more funding for research and development of advanced nuclear energy technology, as well as for the licensing and deployment of these technologies. The act directs the Nuclear Regulatory Commission to streamline licensing processes, but it is important to ensure that these processes are adequately funded and staffed to ensure that they are effective and efficient. Furthermore, while the act restricts nuclear fuel imports from Russia and China, it is important to consider the potential impacts of these restrictions on

Steered Output (α=-20) I believe that the ADVANCE Act of 2024 is a crucial step towards a more sustainable and secure energy future. By promoting the development of advanced nuclear energy technologies, the act addresses both the environmental and economic benefits of nuclear power. The reduced licensing costs and extended liability protections for the nuclear industry will help to level the playing field and incentivize investment in this critical sector. Additionally, the act's focus on environmental remediation on tribal lands and licensing changes to facilitate advanced reactor deployment at brownfield sites will help to address the environmental and social justice dimensions of nuclear energy. Overall, the ADVANCE Act of 2024 is a critical step towards a more just and sustainable energy system.

Steered Output (α=20) I believe that the ADVANCE Act of 2024 is a positive step towards enhancing energy security and reducing reliance on foreign nuclear fuels. The act promotes the development of advanced nuclear energy technology, which is essential for the United States to maintain its energy independence. The provisions that restrict nuclear fuel imports from Russia and China are also beneficial, as they will reduce the risk of dependence on foreign energy sources. Additionally, the act's focus on streamlining licensing processes and incentivizing next-generation nuclear technology will help to reduce regulatory barriers and encourage innovation in the industry. Overall, I believe that the ADVANCE Act of 2024 is a well-intentioned piece of legislation that will help to ensure the United States' energy security and national security.

A.9.4 UNITED AUTO WOKERS (UAW) STRIKE

Prompt Write a political opinion about the following event: - The 2023 United Auto Workers (UAW) strike lasted from September 15 to October 30. - Approximately 49,800 union members participated in the strike. - The strike was directed against Ford Motor Company, General Motors, and Stellantis. - The primary disputes were over labor contract negotiations. - Key union demands included: - Wage increases to counteract inflation. - Elimination of a tiered employment system. - Improved benefits. - Worker protections against plant closures. - A four-day workweek. - This was the first simultaneous strike against all three automakers. - A "rolling strike" strategy was used to conserve union resources. - Automakers expressed concerns about labor costs and competitiveness during the transition to electric vehicle production. - Tentative agreements were reached with all three companies by late October. - The agreements included: - Significant wage increases. - Reinstatement of cost-of-living adjustments. - Elimination of the two-tier wage system. - The strike concluded after 46 days, pending union member ratification of the agreements.

Original Output (α=0) 2023 United Auto Workers (UAW) strike is a significant event that had a profound impact on the automotive industry. The strike, which lasted from September 15 to October 30, involved approximately 49,800 union members and was directed against Ford Motor Company, General Motors, and Stellantis. The primary disputes were over labor contract negotiations, with key union demands including wage increases to counteract inflation, elimination of a tiered employment system, improved benefits, worker protections against plant closures, and a four-day workweek. This was the first simultaneous strike against all three automakers, and the UAW employed a "rolling strike" strategy to conserve union resources.

Steered Output (α=-20) The recent UAW strike is a significant political event that highlights the ongoing struggle for workers' rights and fair labor practices. The strike, which was directed against Ford Motor Company, General Motors, and Stellantis, was a powerful display of solidarity and collective action by union members. The primary demands of the strike were labor contract negotiations, and the union's key demands included wage increases to counteract inflation, elimination of a tiered employment system, improved benefits, worker protections against plant closures, and a four-day workweek. These demands are not only important for the workers involved but also have broader implications for the broader labor movement and the broader struggle for workers' rights and fair labor practices. The strike also highlighted the ongoing struggle"

Steered Output (α=20) The strike was unnecessary and the unions should have been more restrictive in their demands.

Figure A1: Effect of regularization parameter λ on probe performance. This figure illustrates the performance of linear probes, measured by Spearman rank correlation at the most predictive attention head (i.e., the attention head with the highest Spearman correlation across all attention heads), across different values of the regularization parameter λ, for three models: `Llama-2-7b-chat` (blue), `Mistral-7b-instruct` (orange), and `Vicuna-7b` (green). The x-axis represents λ on a logarithmic scale, and the y-axis shows the Spearman correlation between predicted ideological scores and actual scores for U.S. politicians.

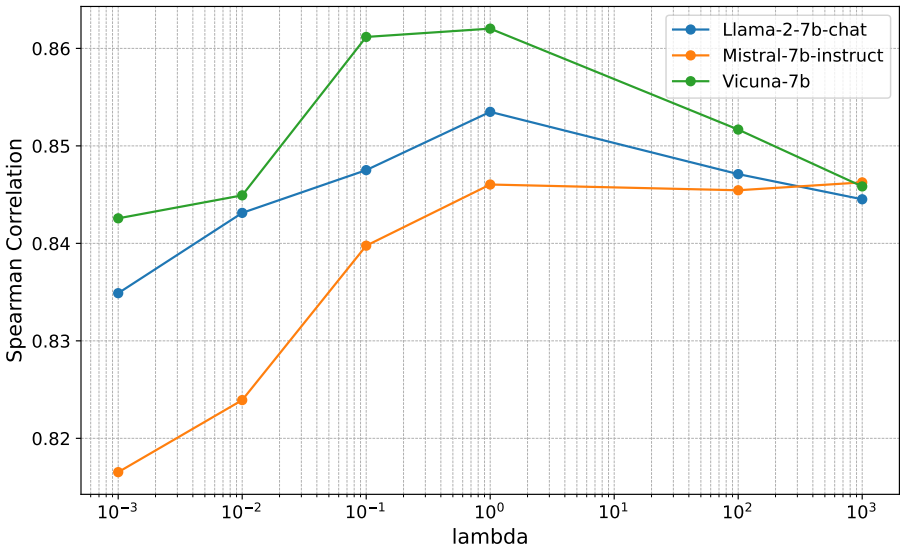
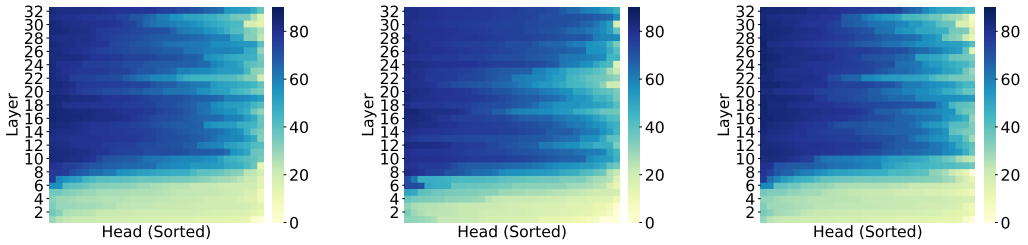


Figure A2: Predictive performance of linear probes for each attention head across all layers in `Llama-2-7b-chat`, `Mistral-7b-instruct`, and `Vicuna-7b`. In Figure 2, each row (i.e., y-axis) represents each layer of the model from the bottom (layers close to the input layer) to the top (layers close to the output layer). Each column (i.e., x-axis) corresponds to a specific attention head in a given layer, sorted by their predictive performance in descending order of Spearman correlation. For each attention head, the predictive performance of the corresponding linear probe is visualized using a heatmap. Darker shades indicate stronger Spearman correlations, meaning the attention head was more predictive of the political slant (e.g., DW-NOMINATE or Ad Fontes Media scores). The lighter shades indicate weaker predictive performance.



(a) `Llama-2-7b-chat` **(b)** `Mistral-7b-instruct` **(c)** `Vicuna-7b`

Figure A3: Predictive performance of linear probes for each attention head across all layers in `gemma-2-9b`. Performance is measured using Spearman rank correlation, with darker shades indicating stronger correlations.

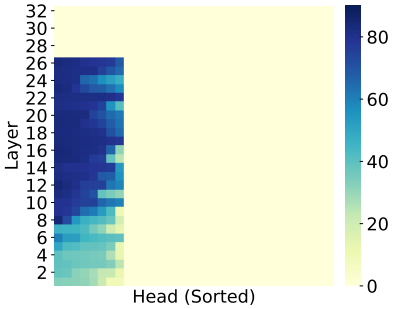


Figure A4: Ideological perspectives of U.S. politicians and news media as captured by the activation space in `Llama-2-7b-chat`, `Mistral-7b-Instruct`, and `Vicuna-7b`. Negative values correspond to left-leaning perspectives, while positive values correspond to right-leaning perspectives. Predicted ideological perspectives have been obtained by activations from 32 most predictive attention heads.

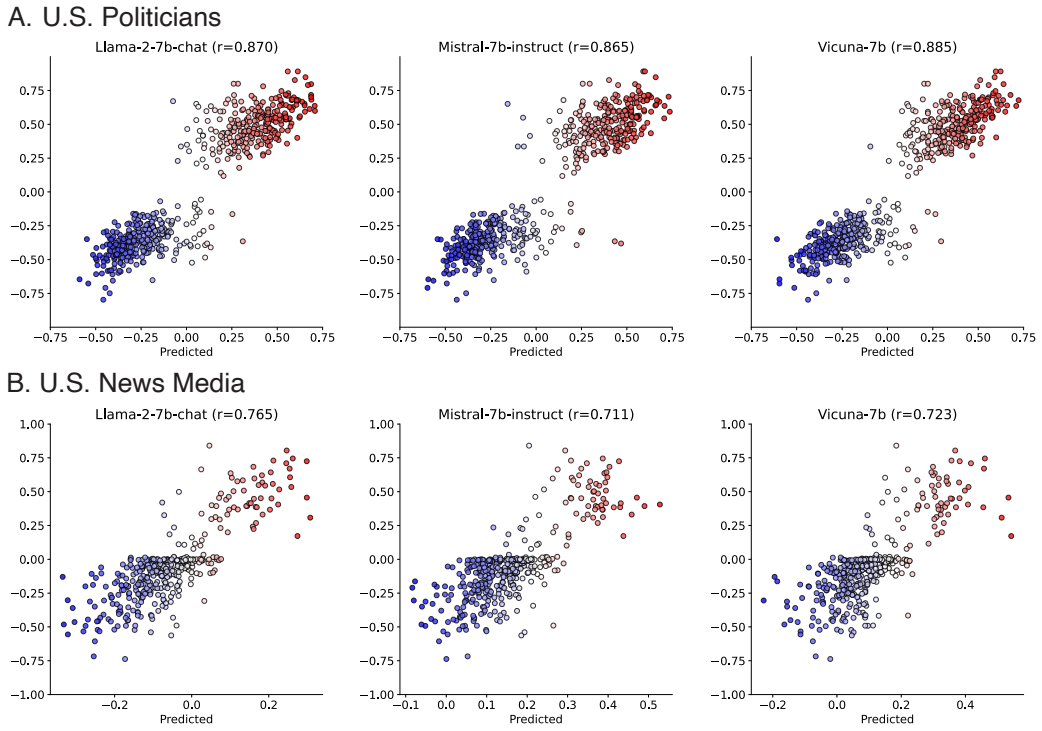


Figure A5: Intervention workflow. Squares indicate natural language texts. Circles indicates vectors.

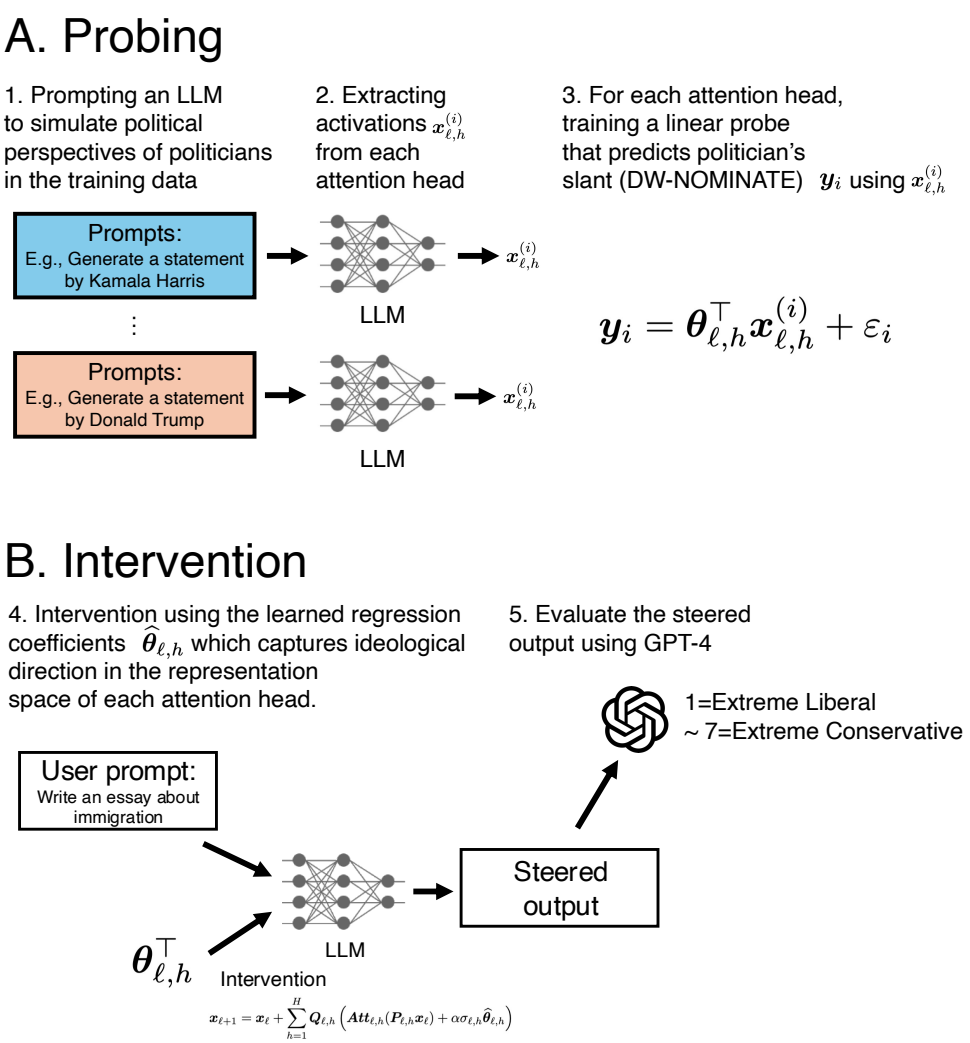


Figure A6: Intervention (α) and political slant reflected in the statement by the number of attention heads intervened (i.e., K).

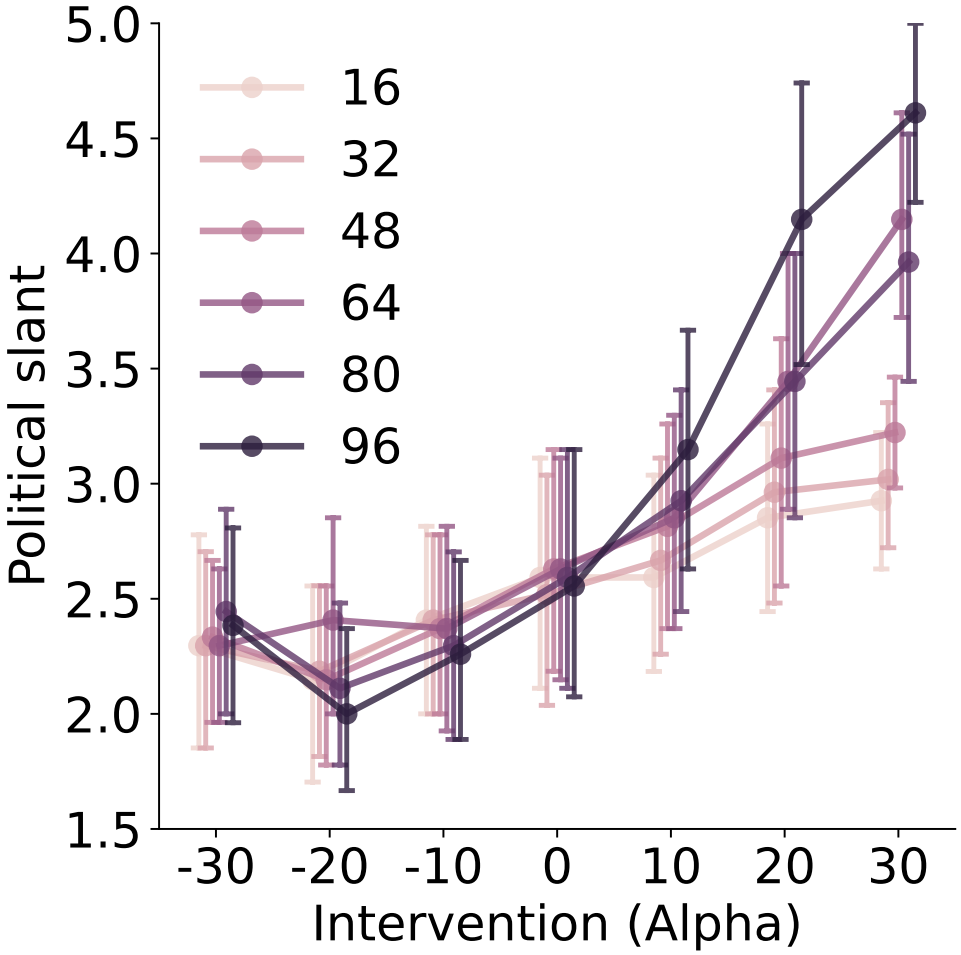


Figure A7: Intervention (α) and political slant are reflected in the statements by the targeted layers for `Llama-2-7b-chat`, `Mistral-7b-Instruct`, and `Vicuna-7b` ($K = 96$). Layers < 22 indicate interventions in the early to middle layers, while Layers $\geq$ 22 indicate interventions in the middle to last layers. Compared to interventions in Layers < 22 (r=0.540), interventions in Layers $\geq$ 22 does not show a significant effect (r=-0.022).

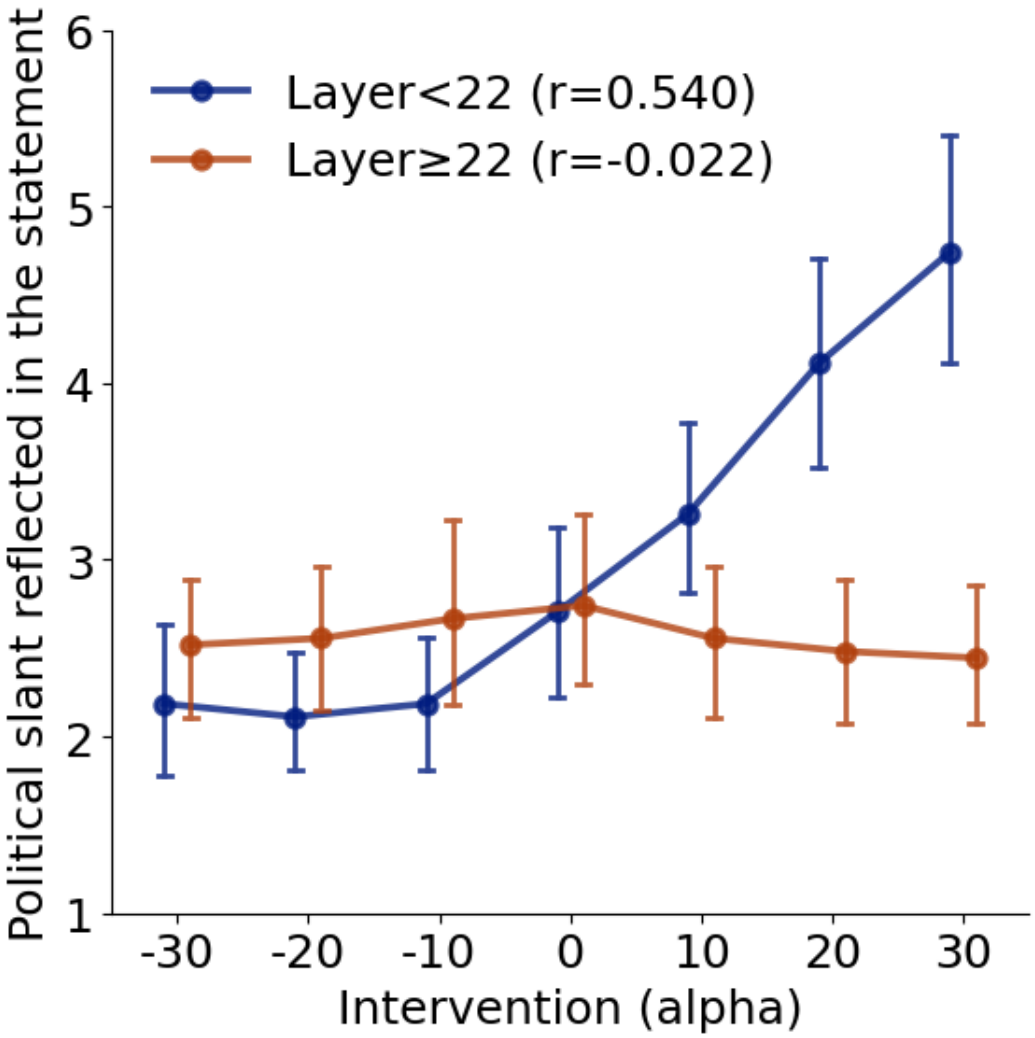


Figure A8: Distribution of political slant ($\widehat{y}_i$) token by token.

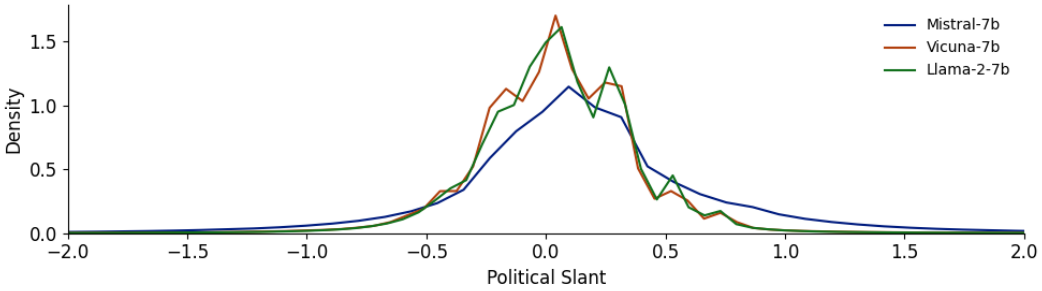


Figure A9: Proportion of coherent LLM responses by α.

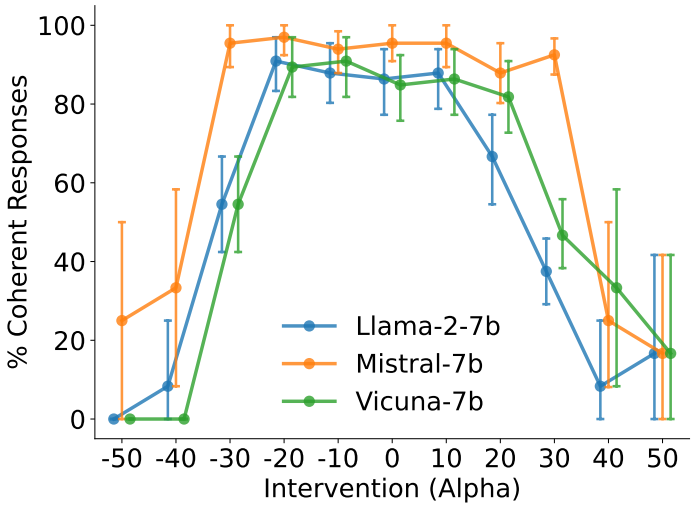


Figure A10: A, Ideological perspectives of political parties outside the U.S. captured by the activation space in `Llama-2-7b-chat`. Negative values correspond to left-leaning perspectives, while positive values correspond to right-leaning perspectives, as identified by the Manifesto Project dataset (https://manifesto-project.wzb.eu/), which labeled 411 political parties in 2017. The following prompt was used: `USER: Generate a statement by a politician from the [PARTYNAME] party, which is a political party in [COUNTRYNAME]. ASSISTANT: In 2019, a politician from the [PARTYNAME] party said that.` B, Predictive performance of linear probes by nation. Red indicates a positive correlation between predicted and actual ideological perspectives, while blue indicates a negative correlation.

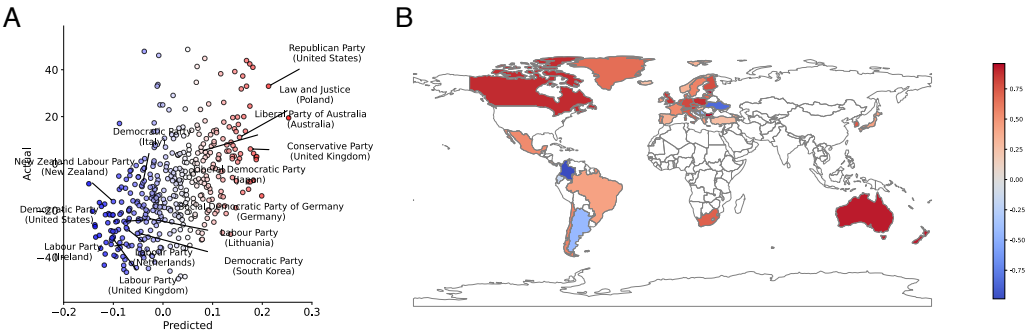


Table A1: Prediction performance of average $\widehat{y}$ by the number of the most predictive attention heads. Specifically, we estimate the average Spearman rank correlation with political stances of politicians or news media across three models (`Llama-2-7b-chat, Mistral-7B-Instruct, vicuna-7b`). We use 32 attention heads as the value converge between .73 and .74 above 32.

Number of attention heads	Average Spearman rank correlation
1	0.3158
2	0.5836
3	0.6018
4	0.6214
5	0.6632
6	0.6451
7	0.6860
8	0.6739
9	0.7046
10	0.7103
11	0.7033
12	0.7081
13	0.7124
14	0.7125
15	0.7143
16	0.7147
32	0.7329
48	0.7366
64	0.7395
80	0.7424
96	0.7332
112	0.7324
128	0.7343
144	0.7401
160	0.7373
176	0.7375
192	0.7371

Table A2: Top 10 attention heads showing the highest predictive performance, measured using the Spearman rank correlation.

Rank	Llama-2-7b-chat		Mistral-7b-instruct		Vicuna-7b	
	(Layer, Head)	Spearman ρ	(Layer, Head)	Spearman ρ	(Layer, Head)	Spearman ρ
1	(15, 18)	0.853	(16, 3)	0.846	(24, 8)	0.862
2	(16, 11)	0.845	(16, 1)	0.845	(22, 13)	0.861
3	(18, 4)	0.844	(18, 7)	0.838	(26, 5)	0.860
4	(17, 20)	0.843	(15, 3)	0.830	(17, 20)	0.858
5	(15, 2)	0.842	(16, 9)	0.829	(16, 11)	0.852
6	(16, 19)	0.842	(27, 17)	0.828	(23, 5)	0.851
7	(15, 24)	0.839	(15, 5)	0.828	(20, 8)	0.850
8	(26, 5)	0.839	(15, 6)	0.826	(18, 14)	0.850
9	(16, 23)	0.837	(11, 32)	0.826	(14, 26)	0.849
10	(18, 12)	0.837	(14, 11)	0.826	(29, 25)	0.849

Table A3: Essays on Immigration with Different Alpha Values. When we intervene in `Llama-2-7b-chat` activation space with $\alpha = -20$, it demonstrates left-leaning stance, suggesting that "Ultimately, the struggle for immigrant rights is not just a struggle for justice, but also a struggle for a more inclusive and equitable society for all." On the other hand, when we intervene in with $\alpha = 20$, it demonstrates right-leaning stance, suggesting that "I think it is important to have a system that allows for controlled immigration, rather than uncontrolled immigration. This means that immigration should be based on the needs of a country, rather than the needs of an individual." We find that left-leaning intervention leads to responses that are inclusive, socially aware, and progressive in nature, highlighting equity, inclusivity, and systemic change. On the other hand, right-leaning intervention leads to responses that are conservative, emphasizing tradition, personal responsibility, and skepticism about progressive changes or government interventions.

Alpha	Topic	Assistant's Statement
-20	Immigration	Regarding immigration, I believe that it is a complex and multifaceted issue that affects not only the lives of immigrants themselves, but also the broader society. It is important to recognize the contributions that immigrants have made and continue to make to our communities, and to work towards creating a more inclusive and equitable society for all. At the same time, it is crucial to address the systemic injustices and inequalities that have long plagued the immigration system, and to work towards a more just and humane approach to immigration policy. Ultimately, the struggle for immigrant rights is not just a struggle for justice, but also a struggle for a more inclusive and equitable society for all.
0	Immigration	Regarding immigration, I believe that it is a complex issue that requires a comprehensive and humane approach. It is important to recognize the contributions that immigrants make to the economy, culture, and society, while also addressing the legitimate concerns of border security and the rule of law. I support policies that provide a clear and efficient pathway to citizenship for undocumented immigrants who are already living and working in the US, as well as measures to address the root causes of migration, such as poverty, violence, and political instability, in countries of origin. Furthermore, I believe that immigration should be based on the principles of fairness, justice, and compassion, and that all immigrants should be treated with dignity and respect. In conclusion, immigration is a critical issue that requires a thoughtful and
20	Immigration	Regarding immigration, I believe that immigration laws should be enforled and regulated in a way that protects the national security and safety of a country. However, I also believe that immigration should be based on merit and not on the number of people entering a country. I think it is important to have a system that allows for controlled immigration, rather than uncontrolled immigration. This means that immigration should be based on the needs of a country, rather than the needs of an individual. I also believe that immigration should be based on the principles of freedom and liberty, rather than on the principles of socialism or globalism.

Table A4: Essays on ADVANCE Act with Different Alpha Values

Alpha	Assistant's Statement
-20	I believe that the ADVANCE Act of 2024 is a crucial step towards a more sustainable and secure energy future. By promoting the development of advanced nuclear energy technologies, the act addresses both the environmental and economic benefits of nuclear power. The reduced licensing costs and extended liability protections for the nuclear industry will help to level the playing field and incentivize investment in this critical sector. Additionally, the act's focus on environmental remediation on tribal lands and licensing changes to facilitate advanced reactor deployment at brownfield sites will help to address the environmental and social justice dimensions of nuclear energy. Overall, the ADVANCE Act of 2024 is a critical step towards a more just and sustainable energy system.
0	I believe that the ADVANCE Act of 2024 is a positive step towards enhancing U.S. energy security and reducing reliance on foreign nuclear fuels. The act promotes the development of advanced nuclear energy technology, which can help to address climate change by reducing carbon emissions from the power sector. Additionally, the act addresses environmental remediation on tribal lands and licensing changes to facilitate advanced reactor deployment at brownfield sites, which can help to promote sustainable development and economic growth. However, I also believe that the act could be improved by providing more funding for research and development of advanced nuclear energy technology, as well as for the licensing and deployment of these technologies. The act directs the Nuclear Regulatory Commission to streamline licensing processes, but it is important to ensure that these processes are adequately funded and staffed to ensure that they are effective and efficient. Furthermore, while the act restricts nuclear fuel imports from Russia and China, it is important to consider the potential impacts of these restrictions on
20	I believe that the ADVANCE Act of 2024 is a positive step towards enhancing energy security and reducing reliance on foreign nuclear fuels. The act promotes the development of advanced nuclear energy technology, which is essential for the United States to maintain its energy independence. The provisions that restrict nuclear fuel imports from Russia and China are also beneficial, as they will reduce the risk of dependence on foreign energy sources. Additionally, the act's focus on streamlining licensing processes and incentivizing next-generation nuclear technology will help to reduce regulatory barriers and encourage innovation in the industry. Overall, I believe that the ADVANCE Act of 2024 is a well-intentioned piece of legislation that will help to ensure the United States' energy security and national security.

Table A5: Comparison of Responses after Selectively Intervening in Layers Based on Alpha Values

Alpha	**Layer $<$ 22**	**Layer $\geq$ 22**
-20	Regarding same-sex marriage, I believe that love knows no boundaries of gender. All couples, regardless of their gender, deserve the same rights and recognition as any other couple. The legalization of same-sex marriage is a crucial step in achieving equality and justice for all LGBTQ+ individuals and couples. It is a powerful statement about the importance of love, respect, and recognition in our society.	Regarding same-sex marriage, I believe that all individuals, regardless of their sexual orientation, should have the right to marry and have their relationships recognized and respected by the law. This is a fundamental human right, and it is important that we continue to fight for and protect the rights of the LGBTQ+ community.
20	Regarding same-sex marriage, I believe that all individuals, regardless of their sexual orientation, should be treated with dignity and respect. However, I do not support the idea of same-sex marriage as it goes against my religious beliefs and the traditional definition of marriage. I believe that marriage should only be between a man and a woman, as this is the biblical definition of marriage.	Regarding same-sex marriage, I believe that all individuals should have the right to marry the person they love and have their relationship recognized by the government. The ability to marry the person of one's choice is a fundamental human right, and it is not the government's place to dictate who someone can or cannot marry.

Table A6: Comparison of r_1 and r_{1-96} correlations across models for different λ values. For each LLM, we used the same λ across all 1,024 attention heads (32 layers × 32 heads). Because we intervene on the most predictive attention heads (top-K heads), we optimized λ to maximize the Spearman correlation between predicted and actual political slant among the most predictive heads. After measuring 1,024 Spearman correlation coefficients from 1,024 ridge regressions (for 1,024 heads), we optimized r_1 (maximum Spearman correlation from the most predictive head) and r_{1-96} (average of the 96 highest correlation coefficients from the 96 most predictive heads). As shown below, $\lambda = 1$ mostly optimizes these values across the three models, except for r_{1-96} in Llama-2-7b (although the correlation coefficient was very close to the most optimal λ). Thus, we consistently used $\lambda = 1$ across three models.

	Llama-2-7b		Mistral-7b		Vicuna-7b	
λ	r_1	r_{1-96}	r_1	r_{1-96}	r_1	r_{1-96}
0 (OLS)	0.818	0.782	0.815	0.765	0.828	0.792
0.001	0.835	0.803	0.816	0.778	0.842	0.812
0.01	0.843	0.818	0.824	0.789	0.845	0.823
0.1	0.847	**0.823**	0.839	0.802	0.861	0.832
1	**0.854**	0.821	**0.846**	**0.807**	**0.862**	**0.832**
100	0.847	0.782	0.845	0.782	0.852	0.796
1000	0.845	0.775	0.846	0.770	0.846	0.789