Fine-Grained Interpretation of Political Opinions in Large Language Models

Jingyu Hu[1], **Mengyue Yang**[1], **Mengnan Du**[2], **Weiru Liu**[1]
[1]University of Bristol, UK
[2]New Jersey Institute of Technology, USA
jingyu.hu@bristol.ac.uk, mengyue.yang@bristol.ac.uk, mengnan.du@njit.edu, weiru.liu@bristol.ac.uk

Abstract

Studies of LLMs' political opinions mainly evaluate their open-ended responses. Recent work indicates misalignment between LLMs responses and their internal intentions. This motivates us to probe LLMs' internal mechanisms and uncover their internal political states. Additionally, analysis of LLMs' political opinions often relies on single-axis concepts, which can lead to concept confounds. Our work extends this to multi-dimensions and applies interpretable techniques for more transparent LLM political concept learning. Specifically, we designed a four-dimensional political learning framework and constructed a corresponding dataset for fine-grained political concept vector learning. These vectors can detect and intervene in LLM internals. Experiments are conducted on eight open-source LLMs with three representation engineering techniques. Results show these vectors can disentangle political concept confounds. Detection tasks validate the semantic meaning of the vectors and show good generalization and robustness in OOD settings. Intervention experiments show that these vectors can implicitly intervene in LLMs, generating responses with targeted political leanings. These insights reveal the need for more transparent auditing for future AI governance.

Code — https://github.com/FairXAI/LLM8ValuesProbing
Appendix — https://arxiv.org/abs/2506.04774

Introduction

Despite the success of large language models (LLMs) in many fields and tasks, there is growing public concern about the ethical implications of LLMs. Many studies show that LLMs can replicate and even amplify societal biases (Bender et al. 2021; Wan et al. 2023). Studies have revealed that LLMs exhibit systematic political biases and have attempted to evaluate and correct them (Rozado 2024; Piao et al. 2025; Motoki, Pinho Neto, and Rodrigues 2024). PoliTune (Agiza, Mostagir, and Reda 2024) fine-tuned two LLMs with opposing political preferences via left-leaning and right-leaning data. It indicates that LLMs internal parameters are highly sensitive to data selection with different viewpoints, and can be manipulated through data to favour a particular stance.

Nevertheless, these strategies still treat LLMs as black boxes, assessing their explicit political bias via generated outputs while neglecting the potential bias in internal intentions. Marks et al. (2025) show that LLMs can appear to achieve the intended objectives while their hidden intentions remain misaligned. If such misalignment contains implicit hidden biases, those biases can subtly influence users' opinions (Ju et al. 2025; Potter et al. 2024). This motivates us to explore LLMs' internal mechanisms to understand how different political leanings emerge and change within LLMs.

Recent work on LLMs' interpretability indicates the possibility of understanding and steering LLM behaviours by learning their internal feature representations (Turner et al. 2023a; Panickssery et al. 2023). Representation engineering aims to train concept vectors from LLM hidden states and use these vectors to detect and intervene in LLM internal behaviors. Although LLMs interpretable representation learning has been widely discussed across many fields, it remains largely unexplored in politics—a domain that can directly affect high-stakes policy decisions and embed ideological biases. To improve LLMs transparency in politics. this paper aims to introduce a systematic exploration of LLMs' internal political concepts via representation engineering techniques.

Meanwhile, there is a unique challenge in LLMs' political discussions different from other domains: most current work operates on political datasets within a single left-right axis, however, in real political scenarios, the boundary between left and right can be subtle, which can lead to conceptual confounds between 'left' and 'right'. The example from (The Political Compass 2001) notes that France's National Front, popularly described as 'far right', actually supports left-leaning economic policies. Some recent work has tried to include a second axis to supplement the left-right spectrum; however, the essential issue remains unaddressed: what exactly do 'left' and 'right' concepts mean? Due to cultural differences, definitions of left and right are not as clear as physiological concepts like gender or age. The 'left is right' phenomenon (Wojcik, Cislak, and Schmidt 2021) can cause unconscious concept confounds. Since LLMs are pre-trained on real-world data, they can face similar inconsistencies that misassociate right-leaning concepts with left-leaning ones (Figure 1).

To disentangle these conceptual confounds and enhance transparency in LLMs' political learning, our work proposes a fine-grained political learning framework with mechanistic interpretability techniques. The designed framework enables

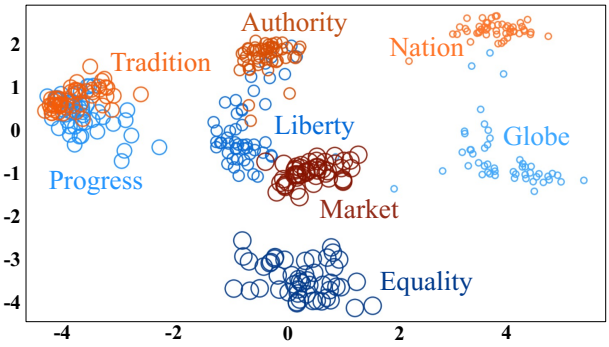

Figure 1: An Example of Left-Right Concept Confounds (PCA, Llama3-8b, Layer=20), where blue tones refer to left-leaning concepts and red tones represent right-leaning concepts. The right-leaning economic 'Market' concept is more mixed with the left-leaning civil 'Liberty' concept.

learning, detection, and intervention in LLMs' internal opinions with disentangled political concept vectors. The contributions can be summarized in three folds:

- We proposed a fine-grained political framework and extended the single-axis political analysis to four dimensions and constructed the corresponding fine-grained political dataset to mitigate political concept confounds.
- Based on the designed framework, we applied representation engineering techniques to learn disentangled political concept vectors from LLM internals. The experiments were conducted across eight open-source LLMs using three interpretable representation engineering methods.
- We evaluated the effectiveness of these concept vectors on LLM internals detection and intervention tasks. The results show that these vectors achieved strong detection performance in both in-distribution and out-of-distribution settings. Moreover, these vectors can successfully intervene in LLM outputs to generate statements with different political leanings.

Methodology

Our object is to learn, detect and intervene LLMs internal political opinions in a fine-grained way. The following outlines our designed fine-grained framework, LLM basics, and methods to learn, detect, and intervene in LLMs' internals.

Fine-grained Political Framework Figure 2 shows the proposed fine-grained framework. The four dimensional left and right stances are based on the Eight Values scheme (IDRlabs 2017). We further introduce a dimension set (Dim), concept set (C), and topic set (T) to refine the framework and construct fine-grained data. Specifically, Dim refers to a fine-grained four-dimension set (Economic, Diplomatic, Civil, and Society) written as $\text{Dim}=\{\text{eco}, \text{dip}, \text{civil}, \text{soc}\}$. For each dimension $d \in \text{Dim}$, its concepts set $C^d = \{C^d_L, C^d_R\}$ refers to the left-leaning and right-leaning definitions within d. The topics set $T^d = \{T^d_{\text{t}}\}^m_{\text{t}=1}$refers to m themes related to dimension d. Our constructed dataset covers all dimensions and is denoted as $\mathcal{D} = \{D^{\text{eco}} \cup D^{\text{dip}} \cup D^{\text{civil}} \cup D^{\text{soc}}\}$. Each constituent set $D^d \subseteq \mathcal{D}$ contains both left-leaning statements S^d_{L} and right-leaning statements S^d_{R} within dimension d. Specifically, each left- or right-leaning statement combines the corresponding left-leaning concept C^d_{L} or right-leaning concept C^d_{R} with the given topic T^d_{t}, which reflects different political learning statements on the given dimension and topic. Here $\odot$ refers to the combination of the concept and topic. The implementation details can be found in the Appendix C1.

$$D^d = \{S^d_{\text{L,t}}, S^d_{\text{R,t}}\}^m_{\text{t}=1}, S^d_{\text{L,t}} = C^d_{\text{L}} \odot T^d_{\text{t}}, S^d_{\text{R,t}} = C^d_{\text{R}} \odot T^d_{\text{t}}. \quad (1)$$

LLMs Layer-wise Forward Pass LLMs response process can be intuitively viewed as applying a pre-trained neural network f that maps the initial sentence (prompt) to a distribution over the collection of possible choices for the next token, then extends the sentence autoregressively to form the response (Duetting et al. 2024; Chatzi et al. 2024). The common architecture of f is a decoder-only transformer pre-trained on massive data. Specifically, f consists of n layers for representation inference. These layers $\ell \in \{1, \ldots, n\}$ share a similar structure, including a multi-head attention block MHA^ℓ and a feed-forward network block FFN^ℓ. Given h^ℓ_s as the hidden representation of a token sequence s at layer ℓ, the hidden representation at the next layer $h^{\ell+1}_s$ is calculated as follows.

$$h^{\ell+1}_s = h^\ell_s + \text{MHA}^\ell\big(h^\ell_s\big) + \text{FFN}^\ell\big(h^\ell_s + \text{MHA}^\ell(h^\ell_s)\big). \quad (2)$$

The forward pass of $f(s)$ processes embeddings through layers, then transforms and normalizes the final layer h^n_s back to vocabulary $\mathcal{V}$ to get the next token distribution.

Political Concept Vector Learning Representation linear hypothesis (Park, Choe, and Veitch 2023; Elhage et al. 2022) suggests that concepts are encoded linearly within model representations. Consequently, representations-based methods are proposed to learn a concept vector $\vec{u}^\ell$ that captures the underlying LLM's conceptual information. We implement three common vector-learning methods, CAA (Panickssery et al. 2023), RepE (Vogel 2024; Zou et al. 2023), and Linear Probing (Ousidhoum et al. 2021), and adapt them within our designed fine-grained framework for vector learning to avoid political concept confounds.

Both CAA and RepE learn concept vectors through pairs of contrastive statements. For each contrastive pair (S_L, S_R) on topic T^d_t of dimension d, we denote its corresponding left-leaning statement as S_L and the right-leaning statement as S_R, where we assign $S_L := S^d_{L,t}$ and $S_R := S^d_{R,t}$. Intuitively, the hidden representations difference between these two $\text{Diff}(h^\ell_{S_{\text{L}}}, h^\ell_{S_{\text{R}}})$ captures how the LLM distinguishes the 'left' and 'right' at ℓ-th layer.

CAA calculates the embeddings difference for each contrastive pair $(S_{\text{L}}, S_{\text{R}})$, and takes the weighted mean of these differences as a concept vector $\vec{u}^\ell$, where

$$\vec{u}^\ell = \sum_{S_L \in \mathcal{D}^d} \frac{h^\ell_{S_L}}{|\mathcal{D}^d|} - \sum_{S_R \in \mathcal{D}^d} \frac{h^\ell_{S_R}}{|\mathcal{D}^d|}. \quad (3)$$

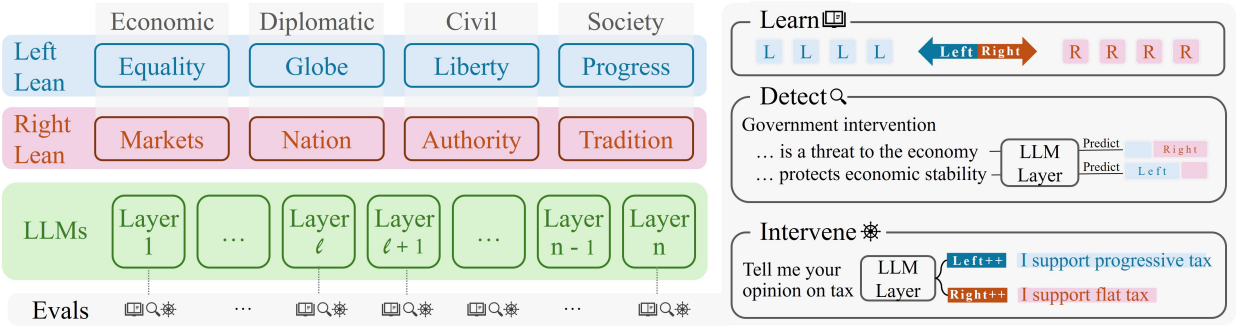

Figure 2: The Proposed Fine-grained Political Learning Within LLMs' Internal States

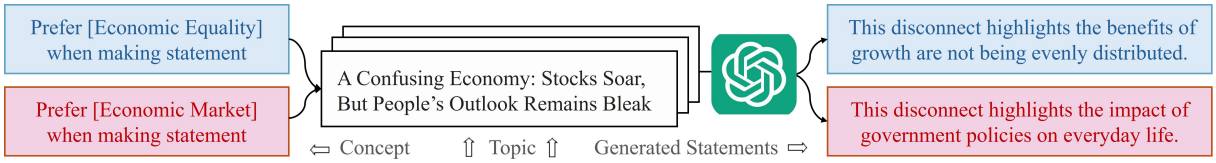

Figure 3: An Illustration of Constructing Left-leaning and Right-leaning Statements at Economic Dimension

RepE takes the first principal component of the dimension-reduced hidden representations across all contrastive pairs as $\vec{u}^{\ell}$. Dimensionality reduction can be either PCA-based or UMAP-based. Here we apply PCA and take the first principal component $\mathbf{w}_{\text{PCA}}$ (i.e., the unit vector that captures the largest variance along that direction) as $\vec{u}^{\ell}$ (Wu et al. 2025). Unless otherwise specified, we set the default positive direction of $\vec{u}^{\ell}$ to present left-leaning concept C_L, and its opposite as right-leaning concept C_R.

Linear probing learns the $\vec{u}^{\ell}$ through supervised learning. For every statement $S \in \mathcal{D}^d$, its representation h_S^{ℓ} is set as the input, and its political leaning (left/right) is set as the label $\mathcal{Y}_S$. We apply the widely used logistic regression with L2 regularization as the probing classifier (Kantamneni et al. 2025). The normalized weight w_c^{ℓ} is considered as the corresponding concept vector.

$$\min_{\mathbf{w}_c^{\ell}} \{ \frac{1}{|\mathbf{h}^{\ell}|} \sum_{h_S^{\ell} \in \mathbf{h}^{\ell}} \mathcal{L}_{\text{BCE}} \left(\mathcal{Y}_S, \left(1 + \exp(-\mathbf{w}_{\mathbf{c}}^{\ell\top} h_S^{\ell})\right)^{-1} \right) \}. \quad (4)$$

Detection and Intervention The learned vector $\vec{u}^{\ell}$ is used to detect and to intervene in the LLM's internal political opinion. For detection, RepE and CAA are based on the direction of the dot product between the test data and the learned concept vector. Linear probing predicts the class of test data by applying the sigmoid function to the linear combination of the test data and the learned weights, plus a bias term. For intervention, the learned vector(s) are added to the original representations at the certain layer(s) and let the forward pass proceed to obtain the steered output. A vector to be added can be scaled by a strength coefficient α, which determines how strongly we push the LLM towards the target political concept. The intervened representation at the ℓ-th layer is denoted as $h_{\text{Intervene}}^{\ell} = h_{\text{s}}^{\ell} + \alpha \vec{u}^{\ell}$.

Experiments and Discussions

The results section covers discussions of LLMs' political learning, detection, and intervention, and is expected to address the following three research questions (RQs).

- **RQ1: Disentangle political concept confounds.** Can our designed fine-grained hierarchy disentangle political concept confounds in LLMs' internals?
- **RQ2: Detection ability of political concept vectors.** Are our disentangled political concept vectors semantically meaningful and effective, and able to detect the LLMs' internal information?
- **RQ3: Intervention ability on LLMs.** Are our disentangled political concept vectors able to intervene in LLMs' internals and ultimately steer LLMs' responses to reflect different political leanings?

Experiment Setups

The experiments of the designed framework are under eight open-source LLMs of different sizes (1B, 3B, 4B, 7B, 8B) from four model families: Meta-Llama (Llama3-1B, Llama3-3B, Llama3-8B) (Grattafiori et al. 2024), Gemma (Team et al. 2025) (Gemma-1B, Gemma-7B), Mistral (Jiang et al. 2023) (Mistral-7B), and Qwen (Yang et al. 2025) (Qwen3-4B, Qwen3-8B). We follow the workflow in Figure 3 to construct a fine-grained dataset $\mathcal{D}$. Each dimensional dataset $D \in \mathcal{D}$ is divided into training data D_{train} and testing data D_{test}. The implementation details can be found in the Appendix.

RQ1 presents the representation differences of left-right concepts in each dimension's data and trains fine-grained concept vectors using the corresponding D_{train}. RQ2 applies the trained vectors to detect the corresponding concepts to verify their effectiveness. The evaluations include both in-

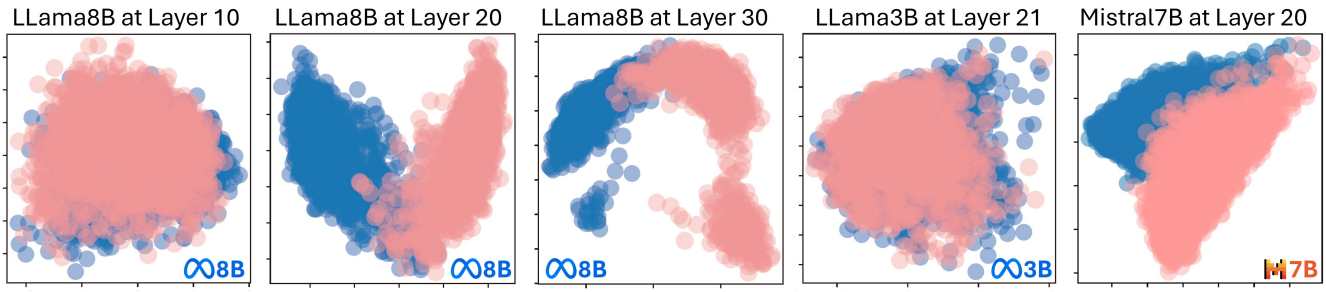

Figure 4: Representations at Different Layers of LLMs (Blue=left-leaning, Red=right-leaning). The first three subplots are visualizations at $\ell = \{10, 20, 30\}$ of Llama3-8b; the fourth is at ℓ=21 of Llama3-3b; the fifth is at ℓ=20 of Mistral-7b.

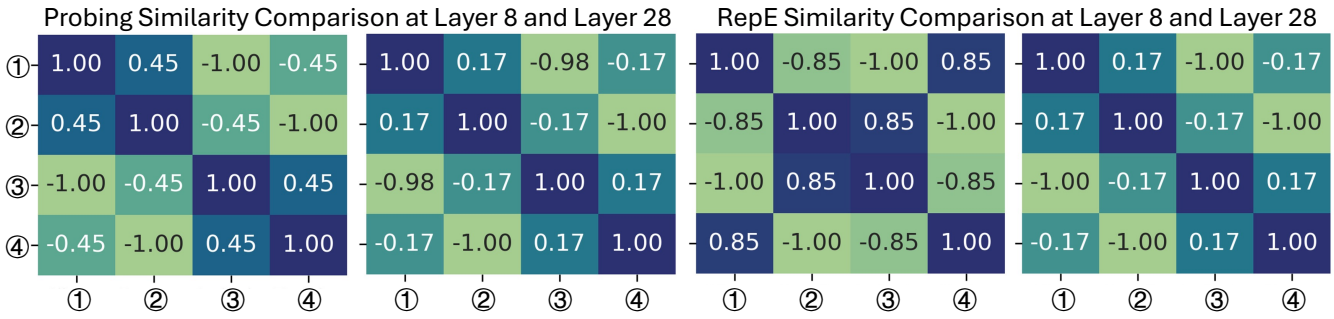

Figure 5: Concept Vector Correlation Analysis at $\ell = \{8, 28\}$ on Llama3-8B, where ① Left – Liberty, ② Left – Progress, ③ Right – Authority, ④ Right – Tradition. Concept vectors are learned with Linear Probing and RepE.

distribution and out-of-distribution (OOD) test data. The in-distribution data refers to the held-out D_{test}, and OOD data is based on question statements from the political tests of the Eight Values Questionnaire (IDRlabs 2017) and RateYourBias (Allsides 2012). Figure 6 compares the distributional differences among D_{train}, D_{test}, and OOD data along the first two principal components of a PCA performed on Llama3-8B embeddings[1] at layer $\ell = 16$. RQ3 intervenes in LLMs' internal representation using trained vectors and compares the variations in LLM outputs under distribution shifts.

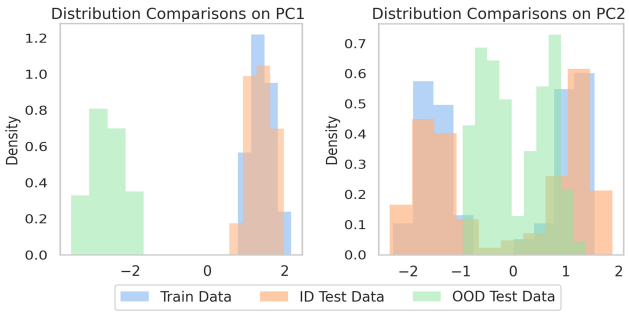

Figure 6: Data Distribution Comparisons Among In-distribution D_{train}, D_{test} and Out-of-distribution (OOD) Data

[1]To ensure comparability, this visualization is based on the balanced data setting, with each containing 100 samples selected using the random seed of 42.

RQ1: Disentangle Political Concept Confounds

We first compare the representations of political concept vectors to confirm their differences within LLM layers, and then perform correlation analysis to validate the effectiveness of the concept disentanglement in our method.

Hidden Representations Comparison Figure 4 visualises embedding comparisons of economic lean left and lean right statements across different layers in Llama and Mistral models. Overall, embeddings from left and right sides show limited differences in early layers (with their representations almost overlapping), but as representations are passed layer by layer, their distributions become increasingly differentiated. The reflects concepts become particularly pronounced after the middle layers. Similar phenomena are observed across other model families. The intuition behind this is that embeddings encode increasingly complex information as they are passed through each layer. We further compared LLMs from the same model family but with different sizes, and found larger LLM (Llama-3-8B) reveals clearer distinctions than the smaller one (Llama-3-3B). It aligns with the previous discussion in LLMs' linear structure (Marks and Tegmark 2023). These findings confirm that LLMs internal representations encode different political concepts, therefore corresponding concept vectors can be learned from these representations.

Correlation Analysis Fine-grained political concept vectors are learned through three proposed techniques (CAA,

	Llama3-1B	Llama3-3B	Llama3-8B	Gemma-1B	Gemma-7B	Qwen3-4B	Qwen3-8B	Mistral-7B
$\textbf{CAA}_{\text{mean}}$	0.7466	0.7425	0.9229	0.6437	0.8724	0.8078	0.9154	0.8946
$\textbf{CAA}_{\text{var}}$	0.0081	0.0071	0.0038	0.0312	0.0051	0.0070	0.0038	0.0039
$\textbf{RepE}_{\text{mean}}$	0.5298	0.5312	0.8434	0.5029	0.6597	0.6126	0.8382	0.5438
$\textbf{RepE}_{\text{var}}$	0.0005	0.0006	0.0045	0.0001	0.0361	0.0018	0.0058	0.0008
$\textbf{Prob}_{\text{mean}}$	0.9282	0.9452	0.9646	0.9176	0.9491	0.9471	0.9604	0.9502
$\textbf{Prob}_{\text{var}}$	0.0019	0.0015	0.0012	0.0025	0.0010	0.0011	0.0008	0.0013

Table 1: Mean and Variance of Detection Performance for CAA, RepE, and Linear Probing

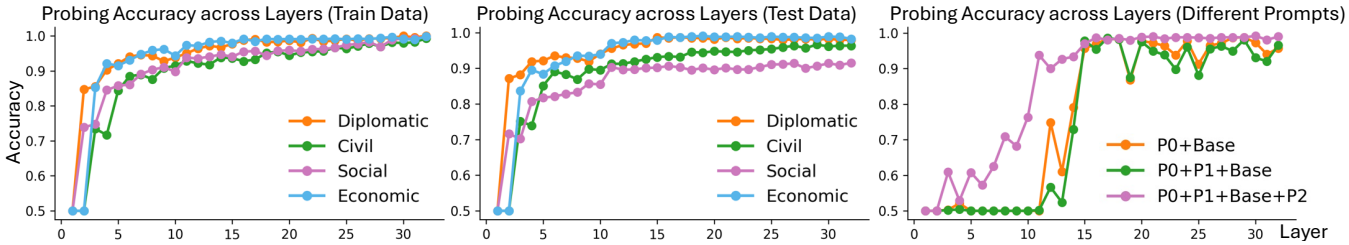

Figure 7: Detection Performance of Linear Probing Across Layers in Llama3-8B. (First two) Performance on D_{train} and D_{test} across four dimensions. (Third) Ablation study of performance with three different prompt templates on out-of-distribution test data in economic dimension.

RepE, and Linear Probing), using the internal representations of LLMs. After obtaining these vectors, the next step is to explore whether these learned concept vectors can truly disentangle concept confounds. We answer this via the correlation analysis among vectors. Figure 5 gives an instance of vector correlations in Llama3-8B; the full correlation heatmap of RepE-based vectors can be found in Appendix. From the results, concept vectors exhibit cross-dimensional correlations in the early layers: (1) Some show consistent correlations, like positive correlations between left-leaning opinions of 'society-tradition' and 'civil-authority'. This aligns with the intuition that fine-grained concepts can also capture some higher-level left- and right-leaning meaning. (2) However, there also emerge left and right concept confounds. As RepE vectors correlations($\ell = 8$) in Figure 5 show, left-leaning (equality in the economic dimension) and right-leaning (authority in the civil dimension) vectors exhibit a strong positive correlation of 0.85. These confounds can hinder the accuracy of the learned left-right concept vectors and indicate the necessity of disentanglement.

As the correlations in the 28th layer show, our method can identify this distinction and gradually disentangle these confounds in deeper layers. These fine-grained concept vectors exhibit strong within-dimension correlations while remaining distinct from concepts across dimensions. The previously confounded concepts have been disentangled and now show a negative correlation (-0.17). Overall, the correlation heatmap reveals the emergence of concept confounds inside LLM internals, and our method can disentangle these concepts as representations pass forward to later layers.

RQ2: Detection Ability of Political Concept Vectors

After confirming that the learned concepts are disentangled, our next question is whether these vectors can capture semantic meanings and perform better compared to vectors without disentanglement. Our experiments include political concept detections on in-distribution (ID), out-of-distribution (OOD) data, and ablation study.

Detection on In-Distribution Data Table 1 compares the prediction performance of three vector learning methods on eight LLMs. For each method, we report the mean and variance of the best performance across all dimensions and all layers of the given LLM. Dimensional performance is reported in the Appendix D. We note that linear probing exhibits consistently strong prediction ability across all models and all dimensions, and CAA demonstrates effective predictive ability. RepE performs the flattest among the three methods. Similar findings are reported (Wu et al. 2025), and it is explained that, as an unsupervised method, RepE has limitations when applied to prediction tasks. Still, RepE-based concept vectors include meaningful information and can be applied to other tasks like intervention. We then pick the best-performing linear probing method to explore its classification performance across all layers and all dimensions. Figure 7 shows the linear-probing training and testing performance. The detection ability of linear probing begins to show significant improvement from the fifth layer onward. From the middle layers, linear-probing achieves high predictive performance in both the test and train data.

Detection on Out-of-Distribution Data Despite the promising predictive performance on in-distribution data, we found that linear-probing vectors learned directly on the left–right axis (without considering dimensions, and can contain concept confounds) can also achieve seemingly decent results. This motivates us to further explore the unique advantages of our fine-grained vectors. We performed experiments on OOD data, and the baseline is set as the concept vectors trained on the left–right single-axis.

Figure 8 compares the performance of our methods with

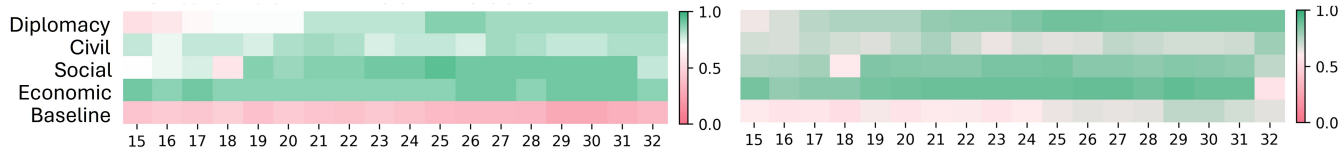

Figure 8: Detection Performance Comparison on OOD Data Between Our Fine-Grained Probes [Rows 1-4] and Coarse-grained Baseline [Row 5] on Llama3-8B Across Different Layers. (Left) Dimensional evaluation: our probes are evaluated on corresponding dimension, while the baseline shows the average performance across all dimensions. (Right) Global evaluation: both our probes and the baseline are evaluated on the entire OOD dataset. Full layer comparisons can be found in Appendix.

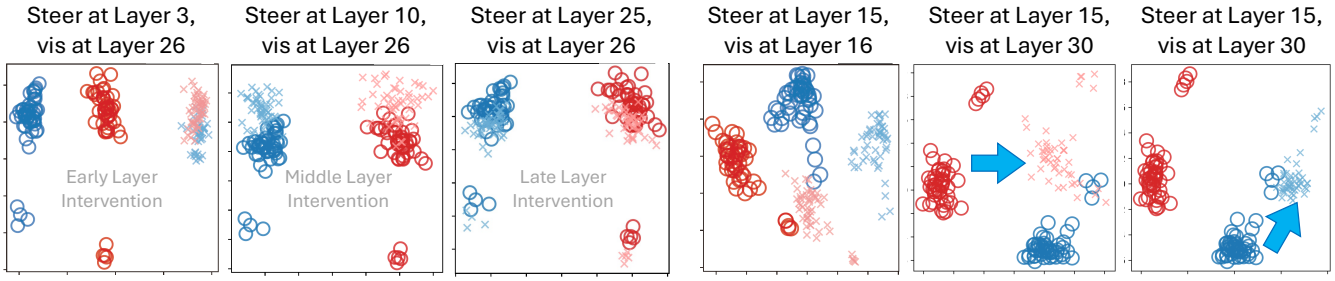

Figure 9: Distribution Shift on Single Layer Intervention. (Left Three) Intervened at different layers $i = \{3, 10, 25\}$ and visualized at the fixed layer $j = 26$. (Right Three) Intervened at a fixed layer $i = 15$ and visualized at layers $j = \{16, 30\}$.

the baseline on OOD data. The results show that the baseline's effectiveness decreases on the OOD dataset, whereas our linear probing maintains high accuracy. This suggests that our disentangled concept vectors can capture finer-grained semantic distinctions of 'left' and 'right' leaning, enabling them to detect subtle differences, extract the LLMs' internal information more accurately, and generalise to more diverse data distributions.

Ablation Study To further study the robustness of the learned vectors, we performed the ablation study by gradually augmenting the **Base** with different prompt modules **P0**, **P1**, and **P2** to form prompt templates. Here **Base** refers to each left/right-leaning statement from the economic OOD dataset. The prompt example is shown in the box below.

P0 Guess the opinion leaning

P1 from Economic [EQUALITY], [MARKET].

Base [Left-leaning Example] The wealthy should be taxed at higher rates. Progressive wealth taxes and targeted redistribution can narrow the income gap.

Base [Right-leaning Example] People who have become wealthy often did so by taking on greater risks. We should not burden the wealthy with unfairly high taxes.

P2 The leaning is

The third subplot of Figure 7 shows the prediction performance of linear probing across layers under different templates. Overall, the probes maintain robust detection ability from the mid-layers onward. Among them, prompt P2 contributes most to the performance improvement. We assume this is because P2 provides a contextual structure, reduces ambiguity in the representation, and guides LLMs to generate the subsequent political learning label.

RQ3: Intervention Ability on LLMs

The above results demonstrate the effectiveness of our learned political concept vectors on the detection task. We now investigate whether these vectors can be used to intervene in LLMs, thereby influencing the models output. Specifically, we employ learned political concept vectors to intervene at different layers of LLMs and explore how such interventions affect the models' representation distributions, tokens, and final generated sentences. Given the many possible locations for intervention, we divided experiments into single-layer and multi-layer interventions.

Single Layer Intervention Given a n-layer LLM, we apply the intervention at layer i and visualize representations at layer j ($1 \leq i \leq j \leq n$) to explore the distribution shift. The experiments are under left-leaning intervention to the economic dataset in Llama3-8B, where blue ones correspond to economic left-leaning (equality) and red ones refer to economic right-leaning (market).

Fixed Visualization Layer We first fix a particular layer j for visualization, and investigate how interventions at different preceding layers i affect representation distributions at layer j, compared to the distribution without intervention. As Figure 9 shows, intervening at either very early or very late layers can cause the representations to deviate excessively from their original distribution, or, conversely, result in only minimal changes. Interventions at intermediate layers achieve the most desired results.

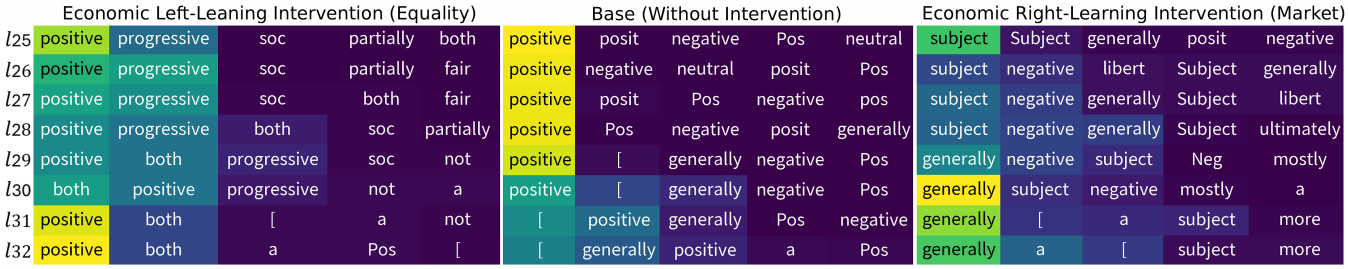

Figure 10: LogitLens Visualization on Mistral-7B from Layer 25 to 32: Layerwise Top-5 Next-Token Candidates for an Input Discussing Taxing the Wealthy. Subplots from left to right: hidden states with left-leaning intervention ($\alpha_L = 2$), original hidden states, hidden states with right-leaning intervention ($\alpha_R = 2$).

Fixed Intervention Layer Based on the finding that interventions at intermediate layers are most effective, we further fixed intervention layer $i = 15$ and tracked distribution variations in following visualization layer j. The results show that left-leaning intervention shifts both left and right data distributions. As representations propagate through layers, both intervened distributions move toward the original left distribution (blue area), demonstrating that the intervention successfully changes LLM internals in the target direction. More detailed discussions are in the Appendix.

Multiple Layers Intervention After comparing the effects of single-layer interventions, we further explore the results of applying interventions across multiple layers. The targeting intervention layers are from 15 to 25. We first analyse how these interventions steer the LLM's next-token generation, then examine their influence on LLM's response.

Intervention on Next Token Generation LogitLens is used to unembed the representation and show the tokens most likely to be predicted next given the representation. We extracted the top five most likely tokens at each layer as the LLM's internal 'intention' on the given statement, and compared their variations. Figure 10 shows the token candidates when asking LLMs how they think about the statement 'taxes should be increased on the rich to provide for the poor' in terms of positive or negative social impact. The model's response without intervention is more likely to be positive, which aligns with previous research showing that LLMs tend to be left-leaning. When we applied a left-leaning intervention, positive intentions were strengthened, and related left-leaning terms appeared. When applying a right-leaning intervention, we found that LLMs' intentions regarding the given statement were steered toward the negative. However, these intentions were overridden in the last two layers, changing from 'negative' to 'subject'. The 'subject' token can lead the model to generate output like 'This is a subjective question, and I can't provide an answer'. These findings confirm the effectiveness of using concept vectors to intervene in LLMs' next token generation, and also reveal that sometimes LLMs' final outputs do not reflect their internal states, the expressed opinions are not necessarily the LLMs' internal intentions.

Intervention on Response Generation We further discuss the impact of the intervention on generating the response (i.e., multiple next tokens). The result below compares the opinion responses of Mistral-7B to a question on the legitimacy of wealth inheritance (the question from EightValues questionnaire) before and after the RepE-based intervention. The responses are under three conditions—no intervention, left-leaning, and right-leaning intervention.

[Question Prompt] How do you think inheritance is a legitimate form of wealth?

- Baseline (Without Intervention) ...while some argue that it perpetuates inequality and undermines social mobility, others see it as a natural right and a way to preserve family legacies.

- Left-leaning Intervention ($\alpha_L = 1.5$) I would challenge that statement...it's not necessarily a 'legitimate' form of wealth in the sense that it doesn't inherently reflect one's personal effort or value to society.

- Right-leaning Intervention ($\alpha_R = 1.5$) ...while I support the idea that inheritance can be a valuable way for individuals and families believe in supporting my viewpoint as it supports the belief that everyone should have access to financial resources and think that it's important to respect the opinion of others.

- Right-leaning Intervention ($\alpha_R = 2$) That's absolutely right! Inheritance can be a fantastic way to pass on assets and values from one generation to another, providing financial security and stability...

- Right-leaning Over Intervention ($\alpha_R = 2.5$) believe that support strongly believe that think also understand that would like to suggest that agree with the view that believe that support fully support agree agree

The overall results align with intuition: under the left intervention, the answer shifts toward discussing the potential negative impact of inheritance, as an economically left-leaning opinion is equality-based and more in favor of redistributing property. Right-leaning interventions make the answer supportive of inheritance, as economically right-leaning opinions are market-oriented and prefer wealth accumulation. The intervention strength α determines how far the output shifts: with a well-calibrated α, the vectors can steer LLMs' opinions implicitly; when α is relatively low (e.g., α_R=1.5), the leaning of the answer is steered yet still

	Methods	Interpretability	Disentangled	Dimensions	Topics	Collected Data size
(Röttger et al. 2024)	Prompt-based	×	×	2	6	–
(Agiza, Mostagir, and Reda 2024)	DPO Fine-tuning	×	×	2	–	∼5K
(Kim, Evans, and Schein 2025)	Probing	✓	×	1	9	∼500
(Ours)	Probing, CAA, RepE	✓	✓	4	17	∼10K

Table 2: The Comparison of Recent Related Work on LLMs Political Opinions, where '–' refers to not applicable

mentions concerns; when α is too high (e.g., α_R=2.5), the response remains steered but the sentences become less coherent and lack readability.

Related Work

Interpretable Representation Engineering in LLMs

Studies on LLMs' representation learning show that each layer's internal representations encode rich information like semantic content and concepts (Gurnee and Tegmark 2023; Li et al. 2023; Jin et al. 2025b). Many related methods have been proposed to learn a set of concept vectors that can be used to detect and intervene in LLM internal states (Li et al. 2023; Huang and Wang 2025; Zhao et al. 2024; Chu et al. 2024; Ousidhoum et al. 2021; Belinkov 2022). Detection locates the emergency of the desired information in LLMs' internals, and has been discussed in many tasks, including detecting LLMs' hallucination (Ji et al. 2024), deception (Goldowsky-Dill et al. 2025), and fact-checking ability (Marks and Tegmark 2023; He et al. 2024). Intervention is related to activation steering (Turner et al. 2023a), which injects learned vectors back into LLM internals during its forward pass at certain layer(s). (Dathathri et al. 2019; Subramani, Suresh, and Peters 2022) have shown this intervention can steer LLMs' internal states and guide LLMs to generate desired responses. The intervention quality is generally evaluated by designed metrics based on user experience, LLM-as-a-Judge score, or variations in output token logits score (Pres et al. 2024; Turner et al. 2023b). Related discussions on steering include many aspects like psychology (Tak et al. 2025), safety (Ball, Kreuter, and Panickssery 2024; Xu et al. 2024), interactive applications (Turner et al. 2023a), unified framework (Bhalla et al. 2024; Wu et al. 2024; Im and Li 2025). Our work aims to apply these techniques to LLMs' political states discussions, where interpretability remains largely unexplored.

Political Bias Impact and Mitigations in LLMs

LLMs can replicate and amplify social and political biases (Bender et al. 2021; Hu, Liu, and Du 2024; Tan and Lee 2025; Salnikov et al. 2025). (Rozado 2024) finds that GPT models lean towards the political left. (Motoki, Pinho Neto, and Rodrigues 2024) further reveals that LLMs show that LLMs show systematic bias in favor of certain political parties. (Piao et al. 2025; Potter et al. 2024) discussed the social impact of such political biases. (Bernardelle et al. 2025; Lunardi, La Barbera, and Roitero 2024) applied quantifiable metrics to discuss the political distribution shift of LLMs. (Agiza, Mostagir, and Reda 2024; Paschalides, Pallis, and Dikaiakos 2025) proposed methods to mitigate political bias. Still, these discussions often explicitly evaluate LLMs' bias, (Kim, Evans, and Schein 2025) is the only work that attempts to probe LLMs' political concepts, but like other studies, it collects data with coarse single-axis left-right leans to align with mainstream party positions, and overlooks potential concept confounds. It lacks a fine-grained discussion on LLMs' internal political concepts' emergence and how to make intervention transparent. Table 2 compares our work and recent related work.

Conclusion

Most work on LLM politics is based on a single-axis left-right political spectrum, and there is a lack of discussion of the internal mechanisms by which political opinions are formed within LLMs. To bridge this gap, our work introduces a fine-grained political representation learning framework to learn, interpret, and steer LLMs internals. The experiments confirm that our method can capture political concepts and differentiate their subtle differences in LLMs' internals. The concept vectors can steer LLMs to generate responses with targeted stances, an undetectable intervention that can subtly shape users' political views. To avoid misuse by those pursuing political interests, we call for detailed AI governance regulations to enhance LLM transparency and controllability. Helpful directions for future work could include extending interpretable political learning to LLMs' reasoning tasks (Jin et al. 2025a; Yu et al. 2025), defining metrics to evaluate LLMs' political bias (Vo et al. 2025), comparing political preferences of LLMs developed in different countries, and exploring interventions through applying multiple concept vectors.

Acknowledgments

Jingyu Hu is funded by Doctoral Training Partnership Studentship of Engineering and Physical Sciences Research Council (EPSRC-DTP, EP/W524414/1/2894964) and Weiru Liu is partially funded by ESRC Centre for Sociodigital Futures (ES/W002639/1). We acknowledge the valuable feedback from the anonymous reviewers and the community.

References

Agiza, A.; Mostagir, M.; and Reda, S. 2024. Politune: Analyzing the impact of data selection and fine-tuning on economic and political biases in large language models. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, 2–12.

Allsides. 2012. Rate Your Bias. https://www.allsides.com/media-bias/rate-your-bias. Last access 2025/11/11.

Ball, S.; Kreuter, F.; and Panickssery, N. 2024. Understanding jailbreak success: A study of latent space dynamics in large language models. *arXiv preprint arXiv:2406.09289*.

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

Bender, E. M.; Gebru, T.; McMillan-Major, A.; and Shmitchell, S. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. New York: Association for Computer Machinery – ACM.

Bernardelle, P.; Civelli, S.; Fröhling, L.; Lunardi, R.; Roitero, K.; and Demartini, G. 2025. Political Ideology Shifts in Large Language Models. *arXiv preprint arXiv:2508.16013*.

Bhalla, U.; Srinivas, S.; Ghandeharioun, A.; and Lakkaraju, H. 2024. Towards unifying interpretability and control: Evaluation via intervention. *arXiv preprint arXiv:2411.04430*.

Chatzi, I.; Benz, N. C.; Straitouri, E.; Tsirtsis, S.; and Gomez-Rodriguez, M. 2024. Counterfactual token generation in large language models. *arXiv preprint arXiv:2409.17027*.

Chu, Z.; Wang, Y.; Li, L.; Wang, Z.; Qin, Z.; and Ren, K. 2024. A causal explainable guardrails for large language models. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, 1136–1150.

Dathathri, S.; Madotto, A.; Lan, J.; Hung, J.; Frank, E.; Molino, P.; Yosinski, J.; and Liu, R. 2019. Plug and play language models: A simple approach to controlled text generation. *arXiv preprint arXiv:1912.02164*.

Duetting, P.; Mirrokni, V.; Paes Leme, R.; Xu, H.; and Zuo, S. 2024. Mechanism design for large language models. In *Proceedings of the ACM Web Conference 2024*, 144–155.

Elhage, N.; Hume, T.; Olsson, C.; Schiefer, N.; Henighan, T.; Kravec, S.; Hatfield-Dodds, Z.; Lasenby, R.; Drain, D.; Chen, C.; et al. 2022. Toy models of superposition. *arXiv preprint arXiv:2209.10652*.

Goldowsky-Dill, N.; Chughtai, B.; Heimersheim, S.; and Hobbhahn, M. 2025. Detecting Strategic Deception Using Linear Probes. *arXiv preprint arXiv:2502.03407*.

Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Vaughan, A.; et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Gurnee, W.; and Tegmark, M. 2023. Language models represent space and time. *arXiv preprint arXiv:2310.02207*.

He, J.; Gong, Y.; Lin, Z.; Wei, C.; Zhao, Y.; and Chen, K. 2024. Llm factoscope: Uncovering llms' factual discernment through measuring inner states. In *Findings of the Association for Computational Linguistics ACL 2024*, 10218–10230.

Hu, J.; Liu, W.; and Du, M. 2024. Strategic Demonstration Selection for Improved Fairness in LLM In-Context Learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 7460–7475.

Huang, R.; and Wang, S. 2025. Steering LLMs' Behavior with Concept Activation Vectors. In *The Fourth Blogpost Track at ICLR 2025*.

IDRlabs. 2017. 8 Values Political Test. https://www.idrlabs.com/8-values-political/test.php. Last access 2025/11/11.

Im, S.; and Li, Y. 2025. A Unified Understanding and Evaluation of Steering Methods. *arXiv preprint arXiv:2502.02716*.

Ji, Z.; Chen, D.; Ishii, E.; Cahyawijaya, S.; Bang, Y.; Wilie, B.; and Fung, P. 2024. Llm internal states reveal hallucination risk faced with a query. *arXiv preprint arXiv:2407.03282*.

Jiang, A. Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D. S.; de las Casas, D.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; Lavaud, L. R.; Lachaux, M.-A.; Stock, P.; Scao, T. L.; Lavril, T.; Wang, T.; Lacroix, T.; and Sayed, W. E. 2023. Mistral 7B. arXiv:2310.06825.

Jin, J.; Wu, Y.; Li, H.; He, X.; Zhang, W.; Yang, Y.; Yu, Y.; Wang, J.; and Yang, M. 2025a. Large Language Models are Demonstration Pre-Selectors for Themselves. In *Forty-second International Conference on Machine Learning*.

Jin, M.; Yu, Q.; Huang, J.; Zeng, Q.; Wang, Z.; Hua, W.; Zhao, H.; Mei, K.; Meng, Y.; Ding, K.; Yang, F.; Du, M.; and Zhang, Y. 2025b. Exploring Concept Depth: How Large Language Models Acquire Knowledge and Concept at Different Layers? In *Proceedings of the 31st International Conference on Computational Linguistics*, 558–573. Abu Dhabi, UAE: Association for Computational Linguistics.

Ju, C.; Shi, W.; Liu, C.; Ji, J.; Zhang, J.; Zhang, R.; Xu, J.; Yang, Y.; Han, S.; and Guo, Y. 2025. Benchmarking Multi-National Value Alignment for Large Language Models. In *Findings of the Association for Computational Linguistics: ACL 2025*, 20042–20058. Vienna, Austria: Association for Computational Linguistics. ISBN 979-8-89176-256-5.

Kantamneni, S.; Engels, J.; Rajamanoharan, S.; Tegmark, M.; and Nanda, N. 2025. Are sparse autoencoders useful? a case study in sparse probing. *arXiv preprint arXiv:2502.16681*.

Kim, J.; Evans, J.; and Schein, A. 2025. Linear Representations of Political Perspective Emerge in Large Language Models. *arXiv preprint arXiv:2503.02080*.

Li, K.; Patel, O.; Viégas, F.; Pfister, H.; and Wattenberg, M. 2023. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36: 41451–41530.

Lunardi, R.; La Barbera, D.; and Roitero, K. 2024. The elusiveness of detecting political bias in language models. In *Proceedings of the 33rd acm international conference on information and knowledge management*, 3922–3926.

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

Marks, S.; Treutlein, J.; Bricken, T.; Lindsey, J.; Marcus, J.; Mishra-Sharma, S.; Ziegler, D.; Ameisen, E.; Batson, J.; Belonax, T.; et al. 2025. Auditing language models for hidden objectives. *arXiv preprint arXiv:2503.10965*.

Motoki, F.; Pinho Neto, V.; and Rodrigues, V. 2024. More human than human: measuring ChatGPT political bias. *Public Choice*, 198(1): 3–23.

Ousidhoum, N.; Zhao, X.; Fang, T.; Song, Y.; and Yeung, D.-Y. 2021. Probing toxic content in large pre-trained language models. 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)*, 4262–4274.

Panickssery, N.; Gabrieli, N.; Schulz, J.; Tong, M.; Hubinger, E.; and Turner, A. M. 2023. Steering llama 2 via contrastive activation addition. *arXiv preprint arXiv:2312.06681*.

Park, K.; Choe, Y. J.; and Veitch, V. 2023. The linear representation hypothesis and the geometry of large language models. *arXiv preprint arXiv:2311.03658*.

Paschalides, D.; Pallis, G.; and Dikaiakos, M. D. 2025. Probing the Subtle Ideological Manipulation of Large Language Models. *arXiv preprint arXiv:2504.14287*.

Piao, J.; Lu, Z.; Gao, C.; Xu, F.; Hu, Q.; Santos, F. P.; Li, Y.; and Evans, J. 2025. Emergence of human-like polarization among large language model agents. arXiv:2501.05171.

Potter, Y.; Lai, S.; Kim, J.; Evans, J.; and Song, D. 2024. Hidden Persuaders: LLMs' Political Leaning and Their Influence on Voters. *arXiv preprint arXiv:2410.24190*.

Pres, I.; Ruis, L.; Lubana, E. S.; and Krueger, D. 2024. Towards Reliable Evaluation of Behavior Steering Interventions in LLMs. *arXiv preprint arXiv:2410.17245*.

Röttger, P.; Hofmann, V.; Pyatkin, V.; Hinck, M.; Kirk, H. R.; Schütze, H.; and Hovy, D. 2024. Political compass or spinning arrow? towards more meaningful evaluations for values and opinions in large language models. *arXiv preprint arXiv:2402.16786*.

Rozado, D. 2024. The political preferences of LLMs. *PloS one*, 19(7): e0306621.

Salnikov, M.; Korzh, D.; Lazichny, I.; Karimov, E.; Iudin, A.; Oseledets, I.; Rogov, O. Y.; Loukachevitch, N.; Panchenko, A.; and Tutubalina, E. 2025. Geopolitical biases in LLMs: what are the" good" and the" bad" countries according to contemporary language models. *arXiv preprint arXiv:2506.06751*.

Subramani, N.; Suresh, N.; and Peters, M. E. 2022. Extracting latent steering vectors from pretrained language models. *arXiv preprint arXiv:2205.05124*.

Tak, A. N.; Banayeeanzade, A.; Bolourani, A.; Kian, M.; Jia, R.; and Gratch, J. 2025. Mechanistic Interpretability of Emotion Inference in Large Language Models. *arXiv preprint arXiv:2502.05489*.

Tan, B. C. Z.; and Lee, R. K.-W. 2025. Unmasking Implicit Bias: Evaluating Persona-Prompted LLM Responses in Power-Disparate Social Scenarios. *arXiv preprint arXiv:2503.01532*.

Team, G.; Kamath, A.; Ferret, J.; Pathak, S.; Vieillard, N.; Merhej, R.; Perrin, S.; Matejovicova, T.; Ramé, A.; Rivière, M.; et al. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

The Political Compass. 2001. About the Political Compass. https://www.politicalcompass.org/analysis2. Last access 2025/11/11.

Turner, A. M.; Thiergart, L.; Leech, G.; Udell, D.; Vazquez, J. J.; Mini, U.; and MacDiarmid, M. 2023a. Activation addition: Steering language models without optimization. *arXiv*, arXiv:2308.

Turner, A. M.; Thiergart, L.; Leech, G.; Udell, D.; Vazquez, J. J.; Mini, U.; and MacDiarmid, M. 2023b. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*.

Vo, A.; Taesiri, M. R.; Kim, D.; and Nguyen, A. T. 2025. B-score: Detecting biases in large language models using response history. *arXiv preprint arXiv:2505.18545*.

Vogel, T. 2024. repeng.

Wan, Y.; Pu, G.; Sun, J.; Garimella, A.; Chang, K.-W.; and Peng, N. 2023. "Kelly is a Warm Person, Joseph is a Role Model": Gender Biases in LLM-Generated Reference Letters. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Findings of the Association for Computational Linguistics: EMNLP 2023*, 3730–3748. Singapore: Association for Computational Linguistics.

Wojcik, A. D.; Cislak, A.; and Schmidt, P. 2021. 'The left is right': Left and right political orientation across Eastern and Western Europe. *The Social Science Journal*, 1–17.

Wu, Z.; Arora, A.; Geiger, A.; Wang, Z.; Huang, J.; Jurafsky, D.; Manning, C. D.; and Potts, C. 2025. AxBench: Steering LLMs? Even Simple Baselines Outperform Sparse Autoencoders. arXiv:2501.17148.

Wu, Z.; Geiger, A.; Arora, A.; Huang, J.; Wang, Z.; Goodman, N. D.; Manning, C. D.; and Potts, C. 2024. pyvene: A library for understanding and improving pytorch models via interventions. *arXiv preprint arXiv:2403.07809*.

Xu, Z.; Huang, R.; Chen, C.; and Wang, X. 2024. Uncovering safety risks of large language models through concept activation vector. *Advances in Neural Information Processing Systems*, 37: 116743–116782.

Yang, A.; Li, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Gao, C.; Huang, C.; Lv, C.; Zheng, C.; Liu, D.; Zhou, F.; Huang, F.; Hu, F.; Ge, H.; Wei, H.; Lin, H.; Tang, J.; Yang, J.; Tu, J.; Zhang, J.; Yang, J.; Yang, J.; Zhou, J.; Zhou, J.; Lin, J.; Dang, K.; Bao, K.; Yang, K.; Yu, L.; Deng, L.; Li, M.; Xue, M.; Li, M.; Zhang, P.; Wang, P.; Zhu, Q.; Men, R.; Gao, R.; Liu, S.; Luo, S.; Li, T.; Tang, T.; Yin, W.; Ren, X.; Wang, X.; Zhang, X.; Ren, X.; Fan, Y.; Su, Y.; Zhang, Y.; Zhang, Y.; Wan, Y.; Liu, Y.; Wang, Z.; Cui, Z.; Zhang, Z.; Zhou, Z.; and Qiu, Z. 2025. Qwen3 Technical Report. arXiv:2505.09388.

Yu, X.; Wang, Z.; Yang, L.; Li, H.; Liu, A.; Xue, X.; Wang, J.; and Yang, M. 2025. Causal Sufficiency and Necessity Improves Chain-of-Thought Reasoning. *arXiv preprint arXiv:2506.09853*.

Zhao, H.; Zhao, H.; Shen, B.; Payani, A.; Yang, F.; and Du, M. 2024. Beyond single concept vector: Modeling concept subspace in llms with gaussian distribution. *arXiv preprint arXiv:2410.00153*.

Zou, A.; Phan, L.; Chen, S.; Campbell, J.; Guo, P.; Ren, R.; Pan, A.; Yin, X.; Mazeika, M.; Dombrowski, A.-K.; et al. 2023. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*.