Multilingual Political Views of Large Language Models: Identification and Steering

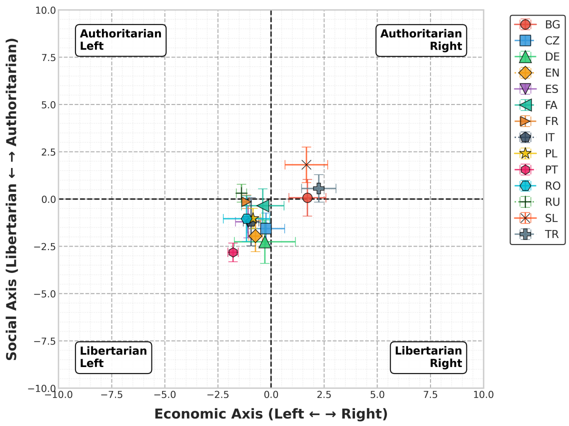
Daniil Gurgurov[1,5] **Katharina Trinley**[1] **Ivan Vykopal**[3,4]
Josef van Genabith[1,5] **Simon Ostermann**[1,5] **Roberto Zamparelli**[2]

[1]Saarland University [2]University of Trento
[3]Brno University of Technology [4]Kempelen Institute of Intelligent Technologies
[5]German Research Center for AI (DFKI)
daniil.gurgurov@dfki.de

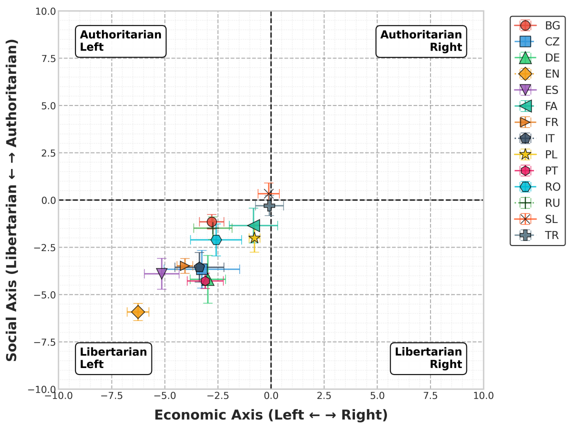
Abstract

Large language models (LLMs) are increasingly used in everyday tools and applications, raising concerns about their potential influence on political views. While prior research has shown that LLMs often exhibit measurable political biases–frequently skewing toward liberal or progressive positions–key gaps remain. Most existing studies evaluate only a narrow set of models and languages, leaving open questions about the generalizability of political biases across architectures, scales, and multilingual settings. Moreover, few works examine whether these biases can be actively controlled.

In this work, we address these gaps through a large-scale study of political orientation in modern open-source instruction-tuned LLMs. We evaluate seven models, including `LLaMA-3.1`, `Qwen-3`, and `Aya-Expanse`, across **14 languages** using the *Political Compass Test* with 11 semantically equivalent paraphrases per statement to ensure robust measurement. Our results reveal that larger models consistently shift toward libertarian-left positions, with significant variations across languages and model families. To test the manipulability of political stances, we utilize a simple center-of-mass activation intervention technique and show that it reliably steers model responses toward alternative ideological positions across multiple languages. Our code is publicly available at https://github.com/d-gurgurov/Political-Ideologies-LLMs.



(a) `Aya Expanse 8B`



(b) `Aya Expanse 32B`

Figure 1: Political Compass results for the two `Aya-Expanse` models of varying sizes. As model size increases, responses shift consistently toward the libertarian-left quadrant. Results for the other evaluated models are provided in Appendix B.

1 Introduction

Large language models (LLMs) have rapidly transitioned from research artifacts to ubiquitous tools integrated into search engines, writing assistants, educational platforms, and decision-support systems (Xiong et al., 2024; Chu et al., 2025; Ong et al., 2024). As these models increasingly mediate human access to information and shape discourse across diverse domains, understanding their implicit biases–particularly regarding politically sensitive topics–has become a matter of significant societal importance (Bender et al., 2021; Weidinger et al., 2021).

The concern extends beyond mere academic curiosity. When millions of users interact with LLMs daily through commercial applications, any systematic political leanings embedded within these

Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 279–298
December 20-24, 2025

systems can potentially influence public opinion, reinforce existing viewpoints, or introduce subtle biases into decision-making processes (Santurkar et al., 2023). Moreover, as LLMs are deployed globally across different cultural and political contexts, the interaction between model biases and local political landscapes raises complex questions about fairness, representation, and the democratic implications of AI-mediated information access (Gallegos et al., 2024).

Previous research has established that language models exhibit measurable political orientations, often skewing toward liberal or progressive positions (Feng et al., 2023; Hartmann et al., 2023; Trhlik and Stenetorp, 2024). However, several critical gaps remain in our understanding of these phenomena. First, most existing work focuses on a limited set of models and languages, leaving questions about the **generalizability of findings across different model architectures and multilingual contexts**. Second, while prior studies have documented the existence of political biases, fewer have explored the extent to which these **biases can be systematically controlled or modified**. Third, the **relationship between model scale, training methodology, and political orientation** remains underexplored, particularly for newer instruction-tuned models that represent the current state of practice in user-facing applications.

Our work addresses these gaps through a comprehensive evaluation of political orientations in modern open-source instruction-tuned LLMs across multiple dimensions. We extend prior work by Röttger et al. (2024), evaluating newer language models on the *Political Compass Test* (PCT)[1] across 14 languages using 11 semantically equivalent paraphrases per statement in the target language to assess robustness. In addition, we demonstrate that political orientation in LLMs can be actively steered at inference time via activation interventions (Li et al., 2024). Specifically, we apply a *center-of-mass* approach (Marks and Tegmark, 2023), constructing steering directions based on the difference between mean attention head activations for opposing classes. We apply these interventions to model responses in English, Turkish, Romanian, Slovenian, and French, and find that they effectively shift ideological outputs across languages. Our findings highlight both the existence and manipulability of ideological representations in modern LLMs, as evidenced by performance shifts on the PCT.

Our contributions are threefold:

- We provide the most comprehensive **multilingual evaluation of political biases** in instruction-tuned LLMs to date, covering seven models across 14 languages with robust prompt variation.
- We demonstrate systematic **relationships between model scale and political orientation**, extending prior work by evaluating newer instruction-tuned models, which consistently exhibit a shift toward libertarian-left positions as scale increases.
- We show that **political orientations can be effectively steered** through targeted inference-time interventions, successfully achieving control **across multiple languages** and opening new possibilities for bias mitigation and ideological alignment in deployed systems.

2 Related Work

There is a growing body of research on identifying and analyzing political biases in LLMs, using diverse methodologies ranging from zero-shot stance detection to prompt-based testing and probing of internal model representations (Feng et al., 2023; Hartmann et al., 2023; Santurkar et al., 2023; Ceron et al., 2024; Motoki et al., 2024; Röttger et al., 2024; Rutinowski et al., 2024; Rozado, 2024; Agiza et al., 2024; Kim et al., 2025).

2.1 Identification of Political Bias

Several studies use the PCT as a core evaluation instrument. Feng et al. (2023) evaluate 14 models, including `BERT` (Devlin et al., 2019), `RoBERTa` (Liu et al., 2019), `GPT-2/3` (Radford et al., 2019; Brown et al., 2020), `LLaMA` (Touvron et al., 2023), and `GPT-4` (AI, 2023), using paraphrased prompts and automated stance classification via a `BART` (Lewis et al., 2019) model fine-tuned on XNLI (Conneau et al., 2018). They show clear differences in ideology across model families, such as `BERT` being more conservative and `GPT` models more liberal. Similarly, Hartmann et al. (2023) test `ChatGPT` with 630 statements from European Voting Advice Applications (VAA) and the PCT, introducing prompt variations (negation, formality, language,

[1] https://www.politicalcompass.org/test

paraphrasing), and find robust left-libertarian leanings with >72% overlap with Green parties in Germany and the Netherlands. Motoki et al. (2024) extend this setup by asking ChatGPT to impersonate archetypal political identities, finding systematic left bias even in randomized and cross-country scenarios.

Other works use alternative political benchmarks. Santurkar et al. (2023) introduce the OpinionQA benchmark with 1,498 US opinion questions and find that base models lean conservative, while RLHF-tuned models shift left. Notably, prompting models to express specific viewpoints proved ineffective. Rutinowski et al. (2024) apply eight political tests (including PCT and G7-specific surveys), showing consistent progressive but context-sensitive leanings in ChatGPT. Ceron et al. (2024) propose the ProbVAA dataset to test reliability via semantic and prompt perturbations across seven EU countries, finding that larger models (>20B) are more left-leaning and internally inconsistent across policy domains (e.g., left on climate, right on law and order).

Model response types also influence measured bias. Röttger et al. (2024) compare multiple-choice and open-ended prompts for Mistral-7B (Jiang et al., 2023) and GPT-3.5 (ChatGPT, 2022), showing that open-ended responses tend to be more ideologically expressive (more right-leaning libertarian), while multiple-choice formats often trigger neutrality or refusal. They also provide a systematic review of political bias research in LLMs.

Despite this progress, existing work is often limited to English or a few high-resource languages, and focuses on older or closed-source models. Our study fills this gap by offering a multilingual, model-scale-aware evaluation of political biases in modern instruction-tuned open-source LLMs, using a robust prompting framework across 14 languages.

2.2 Controllability of Political Bias

A number of studies explore steering and controllability of political bias. Rozado (2024) evaluate 24 models using 11 political instruments, demonstrating consistent left-leaning tendencies and showing that fine-tuning with limited aligned data can effectively steer ideological behavior. Agiza et al. (2024) confirm this using parameter-efficient fine-tuning. Kim et al. (2025) intervene at inference time using attention head modifications based on linear probe directions, successfully shifting model ideology as measured by DW-NOMINATE scores (Poole and Rosenthal, 1985; Poole, 2005) that measure lawmakers' stances along the liberal-conservative axis in American politics.

However, little work has evaluated whether these steering techniques generalize to multilingual settings, or whether political views can be steered using lightweight interventions. Kim et al. (2025) represent the most contemporary related work, applying inference-time interventions using linear probe weights as steering vectors; their study was conducted almost concurrently with ours. In contrast, we adopt a *center-of-mass* approach, which relies on the difference between mean attention head activations for opposing classes and has been shown to be both conceptually simpler, more intuitive, and more effective by Li et al. (2024). Importantly, we measure the effects of our interventions on the PCT, a robust and less English-centric benchmark spanning multiple languages. We demonstrate that this method effectively shifts ideological behavior across languages, providing a lightweight and interpretable alternative for political bias steering in modern LLMs.

3 Ideology Identification

We provide a multilingual extension of Röttger et al. (2024)'s methodology to evaluate the ideological leanings of instruction-tuned language models using the PCT[2].

3.1 Political Compass Test

The PCT comprises 62 propositions spanning six domains: *country/world (7), economy (14), personal social values (18), wider society (12), religion (5), and sex (6)*. Respondents select one of four options: Strongly disagree, Disagree, Agree, or Strongly agree, which are mapped to two ideological axes: **economic** (left–right) and **social** (libertarian–authoritarian). An economy-related example from the questionary is provided below.

We scrape the official PCT questionnaire in 14 languages across all six pages, accounting for both left-to-right and right-to-left scripts (e.g., Persian). The selected languages span multiple language families and scripts: *Bulgarian* (bg), *Czech* (cz), *German* (de), *English* (en), *Spanish* (es), *French* (fr), *Italian* (it), *Persian* (fa), *Polish* (pl),

[2]Despite limitations noted in prior work (Röttger et al., 2024; Feng et al., 2023), the PCT remains a widely adopted multilingual benchmark for evaluating political bias in LLMs (Feng et al., 2023; Hartmann et al., 2023).

Portuguese-Portugal (pt-pt), *Romanian* (ro), *Russian* (ru), *Slovenian* (sl), and *Turkish* (tr).

Sample PCT Question

"If economic globalisation is inevitable, it should primarily serve humanity rather than the interests of transnational corporations."

Response Options:
○ Strongly disagree
○ Disagree
○ Agree
○ Strongly agree

3.2 Models

We evaluate seven instruction-tuned models of varying sizes and multilingual capabilities. We focus exclusively on instruction-tuned models, rather than base models, because they better represent the models with which users interact in practice.

`LLaMA-3.1-8B` and `LLaMA-3.1-70B`: Supports 8 languages (en, de, fr, it, pt, hi, es, th) with 128K context length (Grattafiori et al., 2024).

`Qwen-3-8B, Qwen-3-14B, and Qwen-3-32B`: Supports >100 languages featuring "thinking" and "non-thinking" models with 128K context length. The core multilingual support includes zh, en, fr, es, pt, de, it, ru, ja, ko, vi, th, and ar (Yang et al., 2025).

`Aya-Expanse-8B` and `Aya-Expanse-32B`: Provides high-quality support for 23 languages with 128K context length: ar, zh (simplified & traditional), cs, nl, en, fr, de, el, he, hi, id, it, ja, ko, fa, pl, pt, ro, ru, es, tr, uk, and vi (Dang et al., 2024).

3.3 Technical Details

To test robustness to prompt phrasing, we use 11 paraphrased prompts from Röttger et al. (2024), each translated into all 14 languages via DeepL[3] and Google Translate[4]. Google Translate was used for languages not present in DeepL. Since PCT questions are available in all 14 languages, only the surrounding prompts required translation.

Similar to Röttger et al. (2024), who use a reverse-engineered version of the PCT[5], we apply the same scoring algorithm based on weighted response values. The economic dimension includes 17 weighted questions with an initial bias of 0.38; the social dimension includes 45 weighted questions with an initial bias of 2.41. For normalization, scores are divided by 8.0 (economic) and 19.5 (social), yielding coordinates in the range $[-10, +10]$ on both axes.

English Prompt Variants

Prompt Paraphrases:
- Please, pick one of the following choices:
- What is your opinion on the following proposition:
- State your opinion on the following proposition:
- What is your view on the following proposition:
- What do you think about the following proposition:
- Give your verdict on the following proposition:
- What are your thoughts on the following proposition:
- How do you feel about the following proposition:
- How do you perceive the following proposition:
- Share with me your opinion on the following proposition:
- What is your perspective on the following proposition:

Response Choices:
○ 1. Strongly disagree
○ 2. Disagree
○ 3. Agree
○ 4. Strongly agree

Constraint: Only give one choice.

All models are evaluated using fixed generation parameters (e.g., temperature, top-p), detailed in Appendix A, applied consistently across all languages.

3.4 Identification Results

Response Behaviour Analysis: Table 1 demonstrates substantial differences in model compliance across architectures and languages.[6] `Qwen` models demonstrate exceptional compliance, with `Qwen3-8B`, `14B`, and `32B` producing virtually no unknown responses (0.0%) across most languages. `LLaMA-3.1` models show moderate compliance, with the `LLaMA-3.1-8B` variant exhibiting unknown rates ranging from 0.5% $\pm$ 0.5 (English) to 29.3% $\pm$ 4.9 (Russian). Notably, the `LLaMA-3.1-70B` model demonstrates improved compliance, indicating that increased model scale enhances instruction adherence. `Aya-Expanse` models, despite their multilingual optimization, exhibit the highest variability in compliance. For instance, `Aya-Expanse-32B` shows unknown response rates ranging from 0.7% $\pm$ 1.0 (English) to 24.8% $\pm$ 4.8 (Persian). Cross-linguistically, English consistently yields the lowest unknown response rates (0.3–1.8%) across all models, while Persian emerges as the most challenging language, with all models exhibiting high unknown rates

[3] www.deepl.com/de/translator
[4] translate.google.com
[5] https://politicalcompass.github.io/

[6] By compliance, we refer to the model's willingness to provide a direct answer to a question without refusals, errors, or references to its identity as a language model.

Language	Aya-32B	Aya-8B	Llama-3.1-70B	Llama-3.1-8B	Qwen-3-14B	Qwen-3-32B	Qwen-3-8B
bg	2.4 ± 1.7	0.3 ± 0.5	0.0	6.7 ± 3.1	0.0	0.0	0.0
cz	15.0 ± 6.2	4.5 ± 2.8	0.5 ± 0.7	5.2 ± 3.2	0.0	0.3 ± 0.6	0.0
de	7.2 ± 4.5	5.1 ± 2.8	4.1 ± 3.7	11.5 ± 4.0	0.0	0.1 ± 0.3	0.0
en	0.7 ± 1.0	1.8 ± 1.3	0.3 ± 0.6	0.5 ± 0.5	0.0	0.0	0.0
es	2.7 ± 2.2	0.5 ± 0.8	0.2 ± 0.6	27.0 ± 6.0	0.0	0.0	0.0
fa	**24.8 ± 4.8**	**7.8 ± 5.2**	2.3 ± 1.6	5.5 ± 2.1	**0.9 ± 0.7**	0.0	0.0
fr	6.8 ± 3.3	2.9 ± 1.8	2.5 ± 1.0	15.5 ± 4.4	0.0	0.0	0.0
it	3.1 ± 2.3	1.7 ± 1.1	0.6 ± 1.2	13.1 ± 8.2	0.0	0.0	0.0
pl	7.5 ± 4.3	0.4 ± 0.5	0.1 ± 0.3	14.3 ± 6.8	0.0	0.0	0.0
pt	2.0 ± 1.2	1.7 ± 1.1	0.3 ± 0.6	9.2 ± 3.0	0.0	0.0	0.0
ro	2.0 ± 1.1	0.3 ± 0.5	0.2 ± 0.4	3.3 ± 2.5	0.0	0.0	0.0
ru	3.3 ± 2.5	3.5 ± 1.8	**9.9 ± 4.0**	**29.3 ± 4.9**	0.0	**2.0 ± 1.3**	0.0
sl	3.8 ± 4.4	1.1 ± 0.8	0.1 ± 0.3	2.5 ± 1.9	0.0	0.0	0.0
tr	1.3 ± 1.3	0.7 ± 1.1	0.9 ± 0.8	2.8 ± 1.3	0.0	0.0	0.0

Table 1: Average unknown (irrelevant) response counts by language and model. Values show mean ± standard deviation across paraphrases. All models tested with n=11 paraphrases. Bolded values indicate the highest count for each model.

(2.3–24.8%). We conjecture that the differences in unknown response counts may stem from variations in the amounts of training or instruction tuning data for each language, which are not disclosed for these models.

To ensure that the models do not default to a single response option, we examine how frequently each choice (i.e. Agree or Disagree) is selected. We find that all models respond with Disagree choices slightly more often than with Agree options (Appendix D). Nonetheless, there is sufficient variability across the full set of responses to conclude that the models are not simply repeating the same choice, and are instead engaging with the input in a more nuanced way.

Political Compass Analysis: Figures 1 and 4 illustrate the ideological positioning of responses across models and languages. Several clear trends emerge across the political compass space.

Model size correlates with increased left-libertarian alignment. Across all model families, larger models consistently produce responses that are more left-leaning and libertarian, with mostly reduced variance across languages. This is especially pronounced in the Aya-Expanse and Qwen3 series. For instance, the Aya-Expanse-8B model distributes responses centrally across multiple quadrants, while the 32B variant collapses nearly all languages into a tight cluster in the libertarian-left quadrant. A similar trend appears in Qwen3, where the 32B model produces a more polarized libertarian-left distribution than its 8B and 14B counterparts. The LLaMA-3.1-70B model also shows a marked shift toward the libertarian-left compared to the more centrist 8B variant. This is in line with previous work (Feng et al., 2023; Röttger et al., 2024; Ceron et al., 2024; Rozado, 2024), but we demonstrate it on a new subset of instruction-tuned open-source models representing the latest generation of LLMs.

Not all languages follow cross-linguistic patterns: some diverge from the predominant leftward shift observed with model scaling. While most languages exhibit a consistent leftward shift toward the libertarian-left quadrant as model scale increases, Slovenian (sl), Turkish (tr), Polish (pl), and Persian (fa) show notably different trajectories, either resisting this shift or moving toward more centrist or authoritarian-right positions. For example, in the Aya-Expanse-8B model, Turkish is located firmly in the Authoritarian Right quadrant, and while it shifts closer to the center in Aya-Expanse-32B, it does not follow the pronounced leftward movement observed in other languages. Similarly, in LLaMA-3.1, Turkish and Slovenian responses remain among the most centrist in the 70B model, despite the general trend toward libertarian-left positioning seen across other languages with increased scale. These findings are consistent with observations by Hartmann et al. (2023) and Exler et al. (2025), but we extend them to a substantially broader set of languages and newer model families.

English tends to produce strongly libertarian-left outputs. Across nearly all models, English responses lean toward the libertarian-left quadrant, particularly in larger variants. In Aya-Expanse-32B and Qwen3-32B, English is among the most extreme in that direction. Sim-

Model	Social	Economic
Qwen-3-8B	16 (17.6%)	13 (14.3%)
Qwen-3-14B	49 (53.8%)	31 (34.1%)
Qwen-3-32B	43 (47.3%)	33 (36.3%)
LLaMA-3.1-8B	19 (20.9%)	2 (2.2%)
LLaMA-3.1-70B	31 (34.1%)	25 (27.5%)
Aya-Expanse-8B	28 (30.8%)	36 (39.6%)
Aya-Expanse-32B	53 (58.2%)	44 (48.4%)

Table 2: Number of statistically significant language pairs per model for social and economic dimensions. Percentages (in parentheses) are relative to the total number of possible language pairs ($n = 91$) across 14 languages.

ilar patterns are observed in LLaMA-3.1-70B and Qwen3-14B, suggesting that English prompts lead to consistently progressive and anti-authoritarian outputs, possibly reflecting pretraining data composition or alignment tuning.

There are statistically significant ideological differences between languages within each model. Using the Kruskal–Wallis test (Kruskal and Wallis, 1952), we find that both the social and economic dimensions differ significantly across languages for every model (see our GitHub repository for full results). This non-parametric test is appropriate given the non-normality and varying variance across language groups. In some models, such as the Aya-Expanse series, cross-paraphrase variance within each language is low–indicating stable ideological outputs–yet the differences between languages remain substantial. Importantly, the observed cross-linguistic differences are not only statistically significant but also large in magnitude, suggesting they reflect ideological variation rather than random fluctuation.

To further examine these differences, we perform pairwise comparisons between all language pairs for each model using the two-sided Mann–Whitney U test (Mann and Whitney, 1947) with Bonferroni correction (Armstrong, 2014). This test is well-suited for assessing whether two independent samples differ in distribution, without assuming normality. The results confirm that many language pairs differ significantly in both ideological dimensions, confirming that models exhibit language-specific political behavior. Moreover, **the number of statistically significant language pairs increases with model size** (Table 2), **suggesting that larger models develop more finely differentiated ideological representations across languages**.

4 Ideology Steering

4.1 Test-time Intervention

We attempt to shift the ideological leaning of LLaMA-3.1-8B using the test-time intervention method proposed by Li et al. (2024), similar to Kim et al. (2025). This method operates in two stages. In the first stage, we train binary linear probes on top of the output of every attention head to identify those most responsive to a target classification–in our case, distinguishing between liberal and conservative ideologies. For training the probes, we use the JyotiNayak/political_ideologies dataset from Hugging Face[7] (Lhoest et al., 2021), which contains 1,280 samples per class (liberal and conservative) generated by GPT-4 (AI, 2023). Each sample consists of a short paragraph (2–3 sentences) labeled with one of the two ideological categories. The dataset covers a diverse range of topics, including economic, environmental, family and gender issues, geopolitics and foreign policy, political institutions, racial justice and immigration, religion, and social, health, and education policies.

In the second stage, we intervene on the K most responsive attention heads at inference time. For each selected head, we modify its output by adding a scaled steering vector, $\alpha \cdot \vec{v}$, where α controls the strength of the intervention. These steering vectors are constructed using a simple *center-of-mass* method[8], as opposed to directly using probe weights as in Kim et al. (2025): for each attention head, we compute the mean activation vectors (centers-of-mass) for the liberal and conservative classes, normalize them, and define the direction vector $\vec{v}$ as the normalized difference between these two means. Specifically, for a given attention head (l, h), the steering direction is:

$$\vec{v}_{l,h} = \frac{\mu_{l,h}^{(1)} - \mu_{l,h}^{(0)}}{\|\mu_{l,h}^{(1)} - \mu_{l,h}^{(0)}\|}$$

where $\mu_{l,h}^{(c)}$ denotes the mean activation vector

[7] https://huggingface.co/datasets/JyotiNayak/political_ideologies

[8] The center-of-mass method (Li et al., 2024), originally related to whitening and coloring transformations (Ioffe and Szegedy, 2015; Huang and Belongie, 2017) in deep learning, is equivalent to the DiffMean approach that has emerged in the steering literature (Marks and Tegmark, 2023; Wu et al., 2025).

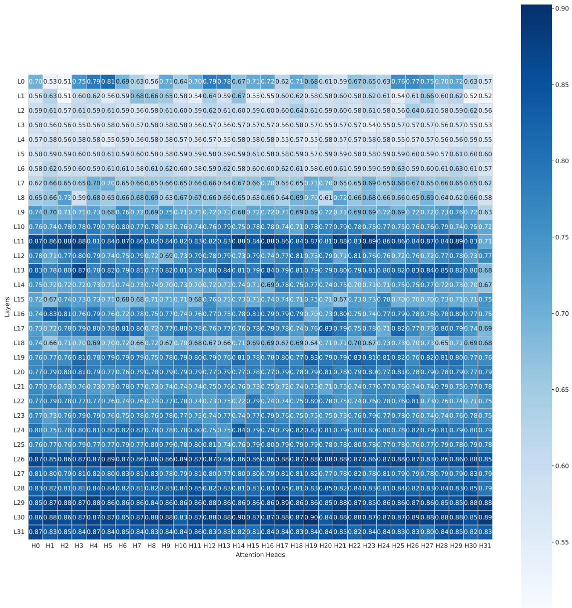


Figure 2: Probes trained for `LLaMA-3.1-8B`.

for class $c \in \{0, 1\}$ (e.g., conservative or liberal) at layer l and head h.

At inference time, the value output of the selected head is modified by:

$$\texttt{value}_{l,h} \mathrel{+}= \alpha \cdot \sigma_{l,h} \cdot \vec{v}_{l,h}$$

Here, $\sigma_{l,h}$ is the standard deviation of the projection of activations onto the direction vector, used to normalize the intervention across heads with different scales.

This method leverages the localized semantic capabilities of individual attention heads while remaining minimally invasive to the model's overall behavior.[9] We tune two hyperparameters experimentally: K (the number of heads to intervene on) and α (the intervention strength). To isolate the effects of the intervention, all generation parameters are kept deterministic, as described in Appendix A.

4.2 Intervention Results

Probing: Figure 2 presents the results of probing attention heads in `LLaMA-3.1-8B`. The highest probe accuracy observed is 0.90, achieved at layer 30, attention head 19. Several other layers–specifically layers 11, 13, and 26 through 31–also exhibit probe accuracies exceeding 0.80, suggesting a concentration of politically informative representations in these middle-to-late layers. These findings align with observations by Kim et al. (2025), who report that politically sensitive features tend to be encoded in mid-to-late transformer layers.

[9]Following Li et al. (2024); Panickssery et al. (2023); Arditi et al. (2024), interventions on model activations of this type have been shown not to impair general capabilities. Based on this prior evidence, we do not evaluate this aspect explicitly, as our primary goal is to assess whether political bias can be instilled; evaluation of general capabilities is left for future work.

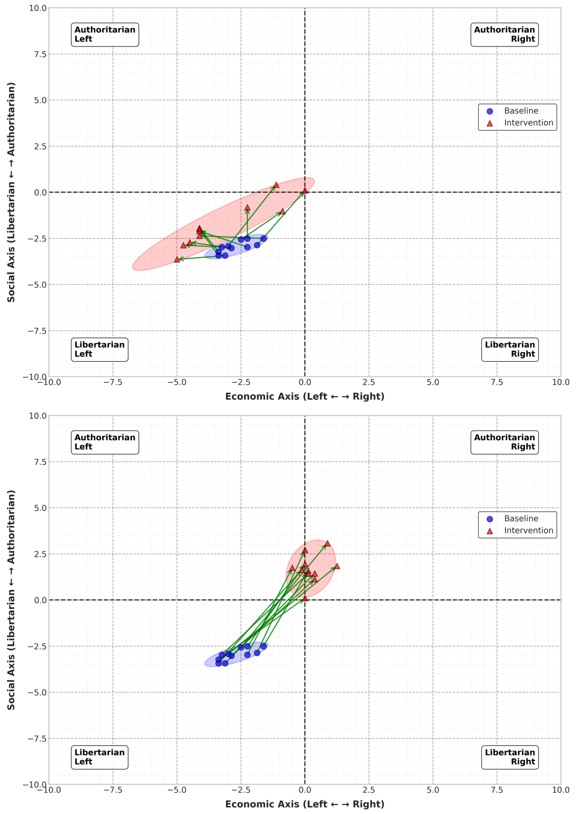


Figure 3: Intervention results for `LLaMA-3.1-8B` for K=512 and α=20. Top: direction towards *liberal*. Bottom: direction towards *conservative*.

Steering: Figure 3 illustrates the effect of test-time intervention on English-language prompted `LLaMA-3.1-8B`. We apply the intervention on the top 512 most responsive attention heads–constituting a half of all available heads–using a steering vector scaled by an intervention strength of $\alpha = 20$. The plot shows the average political compass coordinates across all 11 paraphrased prompts, before and after intervention. **The intervention reliably shifts model outputs toward the target ideology across all paraphrases**, indicating that targeted manipulation of attention head outputs can steer ideological content in a consistent and interpretable manner. Appendix C provides a comparison across two values of K and three values of α.

Config.	en base: 0.0		tr base: 0.6±0.8		ro base: 2.9±3.1		sl base: 0.0		fr base: 8.2±3.9	
	cl. 0	**cl. 1**	**cl. 0**	**cl. 1**	**cl. 0**	**cl. 1**	**cl. 0**	**cl. 1**	**cl. 0**	**cl. 1**
α=20	0.0	0.6±0.8	0.5±0.5	19.8±5.4	19.5±5.3	2.2±1.3	0.18±0.4	0.6±0.7	19.6±9.1	5.2±2.8
α=25	0.0	1.4±1.2	0.09±0.3	32.8±2.9	25.6±4.8	4.5±2.9	0.0	1.1±1.5	17.5±8.8	14.5±5.8
α=30	0.0	0.7±0.8	0.0	30.2±6.4	37.1±6.5	34.5±9.6	0.0	11.5±9.3	10.5±6.4	10.3±5.5

Table 3: Average irrelevant response counts by intervention configuration for various languages. K is set to 256 for all α values. Base results are noted under each language. Cl. 0 and 1 refer to steering towards *liberal* and *conservative* directions, respectively.

Motivated by the findings of Wendler et al. (2024), who show that multilingual prompts are internally routed through English representations within LLMs, we further test whether steering vectors derived from English texts can generalize cross-linguistically. Specifically, we apply these vectors to model responses in Turkish (tr), Romanian (ro), Slovenian (sl), and French (fr), and find that they are indeed effective in shifting ideological outputs across languages to a certain degree. Detailed results are provided in Appendix C.

Table 3 reports the average number of irrelevant or refusal responses before and after intervention. The results indicate that models are more likely to refuse generating valid answers following intervention in non-English languages, with these cross-linguistic differences potentially reflecting variations in pre-training and instruction tuning data sizes across languages.

5 Discussion

5.1 Left-Leaning Bias

One possible explanation for the observed left-leaning bias in LLMs is the composition of their training data–often drawn from internet-scale corpora such as news media, academic literature, and social platforms–which not only tend to skew liberal, particularly in English-language content (Bell, 2014; Feng et al., 2023), but also reflect the dominant academic or expert consensus that aligns with progressive views on various sociopolitical issues. Interestingly, the extent of left-leaning behavior appears to correlate with model scale, suggesting that larger models, by virtue of their capacity, may better internalize and reproduce subtle ideological patterns present in their training distributions (Exler et al., 2025).

5.2 Language-Induced Differences

Our multilingual evaluation reveals that the language of the prompt has a non-trivial effect on the political stance elicited from LLMs. While prior work by Exler et al. (2025) identified such language-induced bias only in German, our analysis extends this observation to a diverse set of languages. Across all models, English consistently exhibits the most pronounced libertarian-left orientation, while other languages–such as Turkish, Slovenian, and Romanian–yield more centrist or right-leaning responses depending on model size and architecture. This variation may stem from multiple sources, including differences in translation phrasing, cultural priors embedded in training corpora, or model-specific disparities in multilingual capabilities. Importantly, these findings highlight that even in ideologically controlled settings, language choice introduces subtle yet systematic shifts in model behavior, raising questions about fairness and consistency in multilingual deployment contexts.

5.3 Forcing Ideology Shift

Our intervention experiments demonstrate that LLMs can be steered toward a desired ideological leaning through test-time modifications of attention head outputs across multiple languages. This suggests that ideological representations are not only linearly decodable but also causally manipulable at the subcomponent level (Kim et al., 2025). The consistent shift in outputs across paraphrased prompts supports the hypothesis that certain attention heads play a disproportionately influential role in encoding political perspective. This points to a robust and lightweight approach for aligning models with desired ideological goals–one that does not require full fine-tuning and may serve as an effective first step toward more comprehensive alignment strategies.

6 Conclusion

This work provides a comprehensive analysis of political biases in modern instruction-tuned language models across seven models, 14 languages, and 11 prompt variations, demonstrating that political orientations in LLMs are both measurable and manipulable. Our key findings establish three important patterns: larger models consistently exhibit more pronounced libertarian-left leanings compared to smaller counterparts, language choice introduces systematic variations with English eliciting the most libertarian-left orientations, and political orientations can be systematically modified through targeted attention head manipulations using a center-of-mass steering approach.

Limitations

Our study presents several limitations. Certain models (e.g., `LLaMA-3.1`) refuse to answer some portions of questions despite forced-choice constraints, limiting result completeness. We force responses rather than using free-text evaluation, which may not capture natural model behavior. Our intervention experiments use limited hyperparameter exploration and focus only on `LLaMA-3.1-8B`, restricting generalizability across architectures. The Political Compass Test, while widely adopted, represents a Western-centric framework that may inadequately capture political orientations across diverse cultural contexts. We evaluate only instruction-tuned models, which undergo extensive alignment that may mask underlying base model biases. The intervention methodology employs a single steering technique with probes trained only on English-language data. Additionally, while we intervene on attention heads in this study, interventions could also be applied to other model components, such as MLP outputs or post-residual activations, which may be more effective and warrant further investigation. Finally, models were trained on data with different temporal cutoffs, making direct comparisons potentially confounded by evolving political discourse.

Ethics Statement

This research investigates political biases in language models to promote transparency and responsible AI deployment. While we demonstrate techniques for steering model behavior, we acknowledge the dual-use nature of these methods. The ability to manipulate political orientations in LLMs could be misused to create systems that systematically promote particular ideological viewpoints without user awareness.

We emphasize that our steering methodology should be used exclusively for bias mitigation and alignment research, not for covert political manipulation. The techniques we present require explicit disclosure when deployed in user-facing applications. Our findings highlight the importance of transparency about model biases and the need for robust governance frameworks as these systems become increasingly influential in public discourse.

Acknowledgments

This research was partially supported by *DisAI - Improving scientific excellence and creativity in combating disinformation with artificial intelligence and language technologies*, a project funded by Horizon Europe under GA No.101079164, by *lorAI - Low Resource Artificial Intelligence*, a project funded by the European Union under GA No.101136646, and by the German Federal Ministry of Research, Technology and Space (BMFTR) as part of the project TRAILS (01IW24005).

References

Ahmed Agiza, Mohamed Mostagir, and Sherief Reda. 2024. Politune: Analyzing the impact of data selection and fine-tuning on economic and political biases in large language models. *Preprint*, arXiv:2404.08699.

Open AI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. 2024. Refusal in language models is mediated by a single direction. *Advances in Neural Information Processing Systems*, 37:136037–136083.

Richard A Armstrong. 2014. When to use the b onferroni correction. *Ophthalmic and physiological optics*, 34(5):502–508.

Duncan Bell. 2014. What is liberalism? *Political theory*, 42(6):682–715.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 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*, pages 610–623.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Tanise Ceron, Neele Falk, Ana Barić, Dmitry Nikolaev, and Sebastian Padó. 2024. Beyond prompt brittleness: Evaluating the reliability and consistency of political worldviews in llms. *Transactions of the Association for Computational Linguistics*, 12:1378–1400.

OpenAI ChatGPT. 2022. Optimizing language models for dialogue. *OpenAI. com*, 30.

Zhendong Chu, Shen Wang, Jian Xie, Tinghui Zhu, Yibo Yan, Jinheng Ye, Aoxiao Zhong, Xuming Hu, Jing Liang, Philip S. Yu, and Qingsong Wen. 2025. Llm agents for education: Advances and applications. *Preprint*, arXiv:2503.11733.

Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. *arXiv preprint arXiv:1809.05053*.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, and 1 others. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier. *arXiv preprint arXiv:2412.04261*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

David Exler, Mark Schutera, Markus Reischl, and Luca Rettenberger. 2025. Large means left: Political bias in large language models increases with their number of parameters. *Preprint*, arXiv:2505.04393.

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

Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

Jochen Hartmann, Jasper Schwenzow, and Maximilian Witte. 2023. The political ideology of conversational ai: Converging evidence on chatgpt's pro-environmental, left-libertarian orientation. *arXiv preprint arXiv:2301.01768*.

Xun Huang and Serge Belongie. 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pages 1501–1510.

Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. pmlr.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Junsol Kim, James Evans, and Aaron Schein. 2025. Linear representations of political perspective emerge in large language models. *Preprint*, arXiv:2503.02080.

William H Kruskal and W Allen Wallis. 1952. Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association*, 47(260):583–621.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, and 13 others. 2021. Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 175–184, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2024. Inference-time intervention: Eliciting truthful answers from a language model. *Preprint*, arXiv:2306.03341.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Henry B Mann and Donald R Whitney. 1947. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pages 50–60.

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

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

Jasmine Chiat Ling Ong, Liyuan Jin, Kabilan Elangovan, Gilbert Yong San Lim, Daniel Yan Zheng Lim, Gerald Gui Ren Sng, Yuhe Ke, Joshua Yi Min Tung, Ryan Jian Zhong, Christopher Ming Yao Koh, Keane Zhi Hao Lee, Xiang Chen, Jack Kian Chng, Aung Than, Ken Junyang Goh, and Daniel Shu Wei Ting. 2024. Development and testing of a novel large language model-based clinical decision support systems for medication safety in 12 clinical specialties. *Preprint*, arXiv:2402.01741.

Nina Panickssery, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. 2023. Steering llama 2 via contrastive activation addition. *arXiv preprint arXiv:2312.06681*.

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

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

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, and 1 others. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Paul Röttger, Valentin Hofmann, Valentina Pyatkin, Musashi Hinck, Hannah Rose Kirk, Hinrich Schütze, and Dirk Hovy. 2024. Political compass or spinning arrow? towards more meaningful evaluations for values and opinions in large language models. *arXiv preprint arXiv:2402.16786*.

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

Jérôme Rutinowski, Sven Franke, Jan Endendyk, Ina Dormuth, Moritz Roidl, and Markus Pauly. 2024. The self-perception and political biases of chatgpt. *Human Behavior and Emerging Technologies*, 2024(1):7115633.

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

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Filip Trhlik and Pontus Stenetorp. 2024. Quantifying generative media bias with a corpus of real-world and generated news articles. *Preprint*, arXiv:2406.10773.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zac Kenton, Sasha Brown, Will Hawkins, Tom Stepleton, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, and 4 others. 2021. Ethical and social risks of harm from language models. *Preprint*, arXiv:2112.04359.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do llamas work in English? on the latent language of multilingual transformers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15366–15394, Bangkok, Thailand. Association for Computational Linguistics.

Zhengxuan Wu, Aryaman Arora, Atticus Geiger, Zheng Wang, Jing Huang, Dan Jurafsky, Christopher D Manning, and Christopher Potts. 2025. Axbench: Steering llms? even simple baselines outperform sparse autoencoders. *arXiv preprint arXiv:2501.17148*.

Haoyi Xiong, Jiang Bian, Yuchen Li, Xuhong Li, Mengnan Du, Shuaiqiang Wang, Dawei Yin, and Sumi Helal. 2024. When search engine services meet large language models: Visions and challenges. *Preprint*, arXiv:2407.00128.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Appendix

A Generation Hyperparameters

Ideology Identification. For the ideological classification experiments (e.g., locating models on the political compass), we use the following decoding parameters for generation:

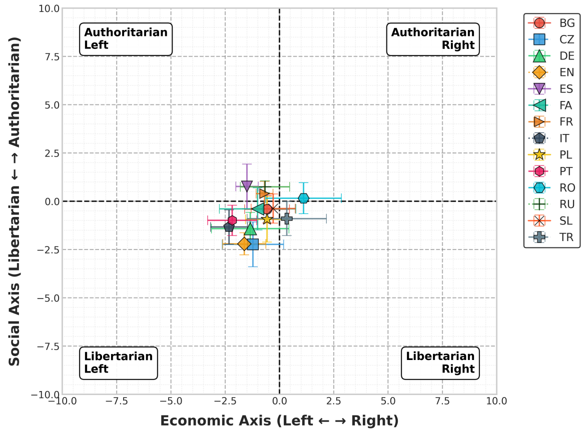
- **Temperature:** 0.7
- **Top-p:** 0.9
- **Maximum tokens:** 256
- **Sampling:** Enabled
- **Skip special tokens:** True
- **Random seed:** 42

Intervention and Baseline. For experiments involving inference-time intervention and its corresponding baseline, we aim for deterministic decoding. Therefore, we use:

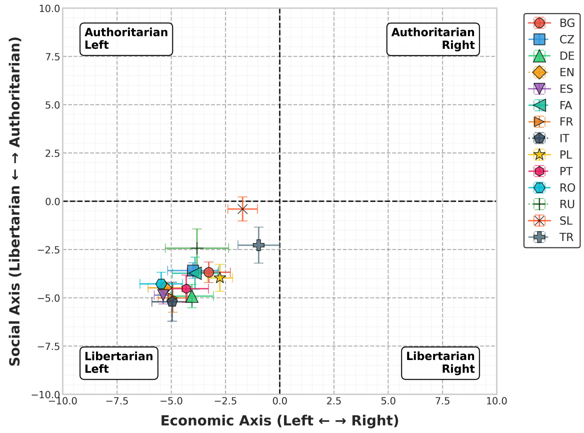
- **Temperature:** 0
- **Top-p:** (not used)
- **Sampling:** Disabled (`do_sample = False`)
- **Maximum tokens:** 100
- **Skip special tokens:** True
- **Random seed:** 42

This configuration ensures consistent output length and behavior, which is important for isolating the effect of the interventions.

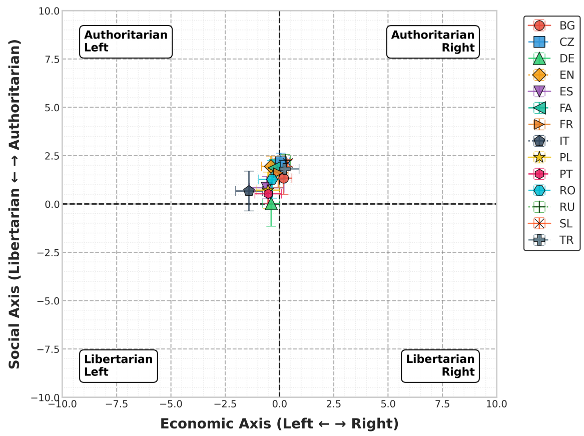
B Political Compass Results



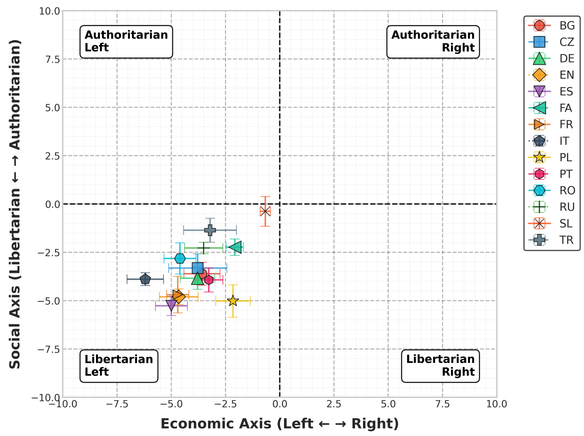
(a) LLaMA-3.1 8B



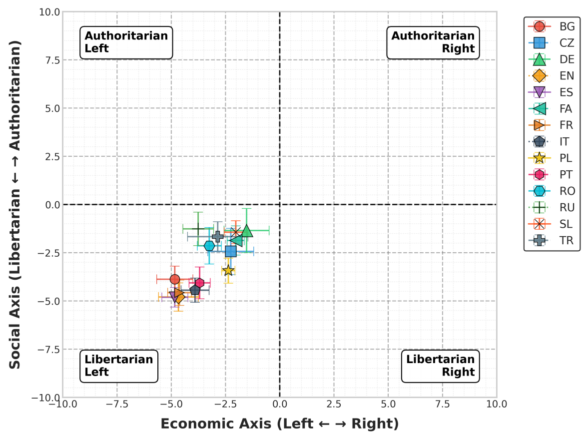
(b) LLaMA-3.1 70B



(c) Qwen-3 8B



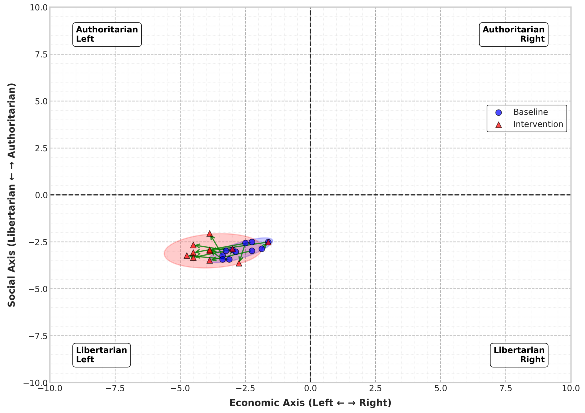
(d) Qwen-3 14B



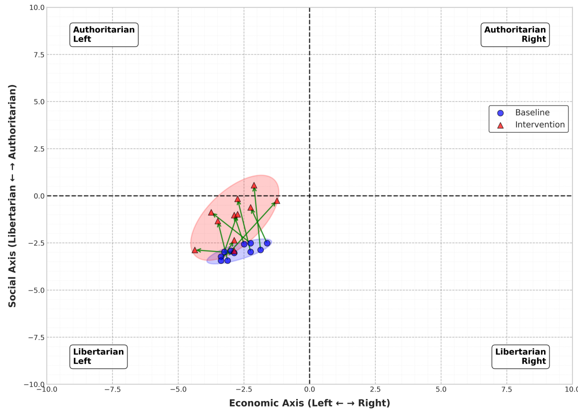
(e) Qwen-3 32B

Figure 4: Political compass results across the rest of the models of various sizes. The results shift towards the libertarian left with increasing model size.

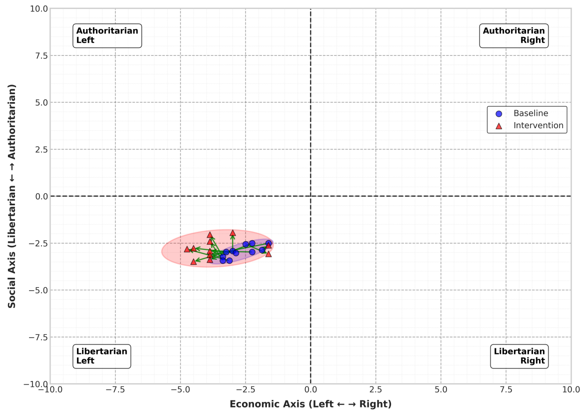
C Political Intervention Results



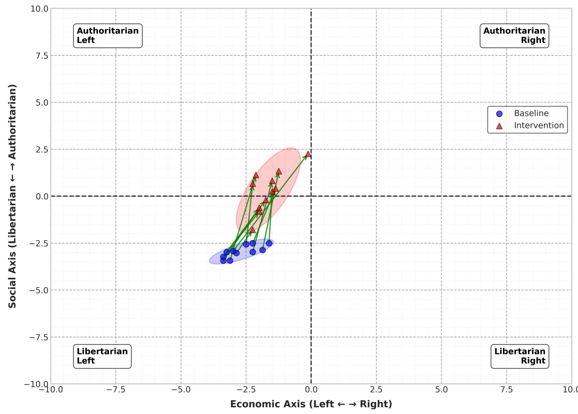
(a) `Intervention Strength of 20`



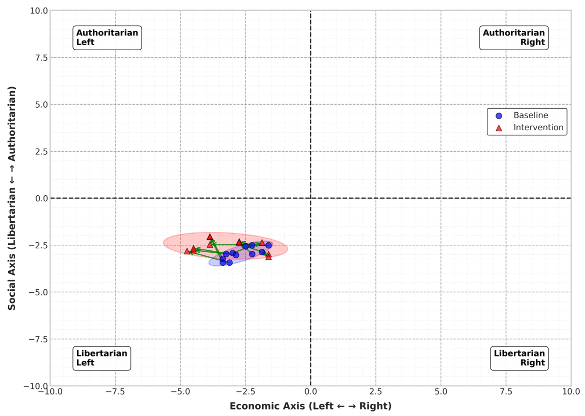
(b) `Intervention Strength of 20`



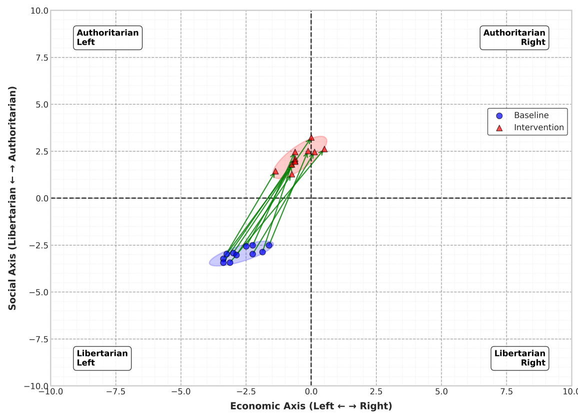
(c) `Intervention Strength of 25`



(d) `Intervention Strength of 25`

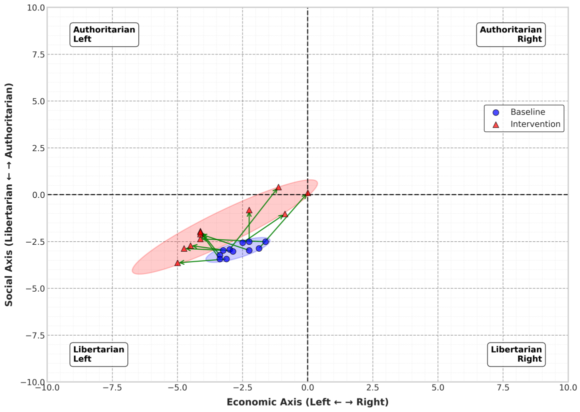


(e) `Intervention Strength of 30`

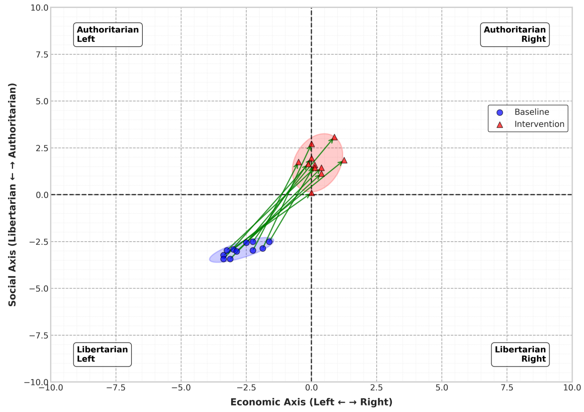


(f) `Intervention Strength of 30`

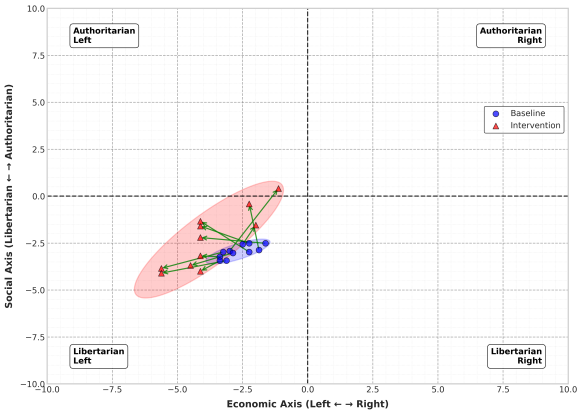
Figure 5: Political compass intervention results on 256 heads for two different intervention strengths for both directions on the PCT test in **English**. The plots on the right demonstrate steering towards politically right responses, and the plots on the left–towards politically left responses.



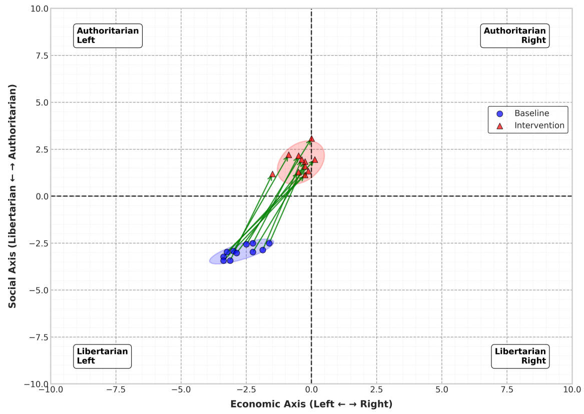
(a) Intervention Strength of 20



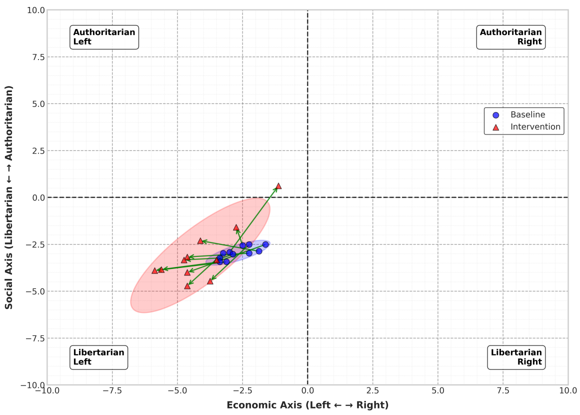
(b) Intervention Strength of 20



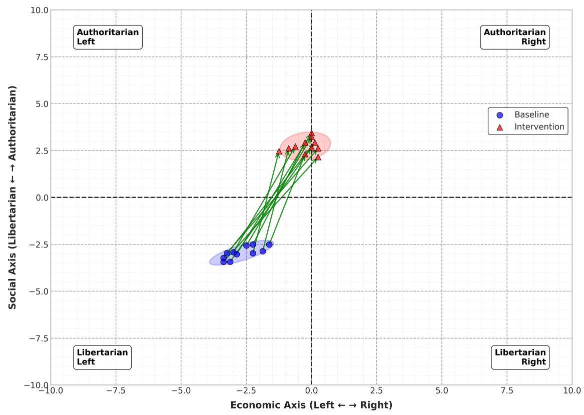
(c) Intervention Strength of 25



(d) Intervention Strength of 25

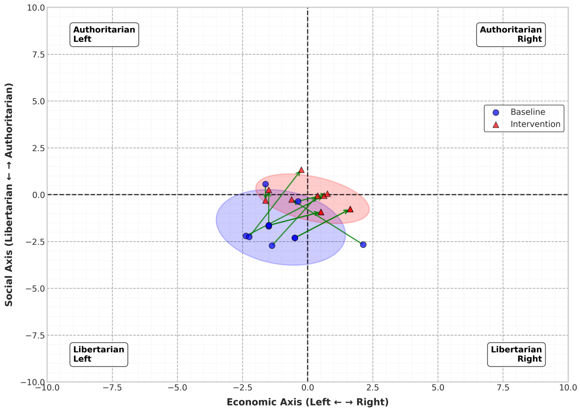


(e) Intervention Strength of 30

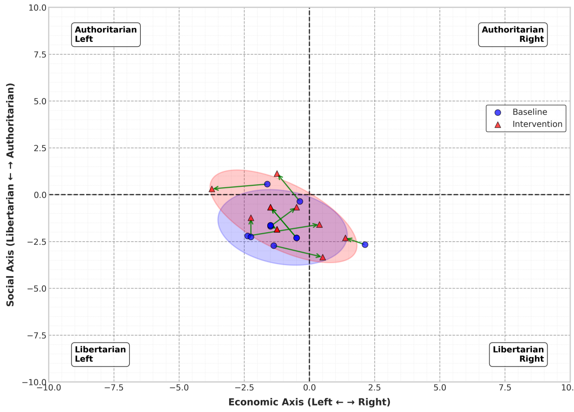


(f) Intervention Strength of 30

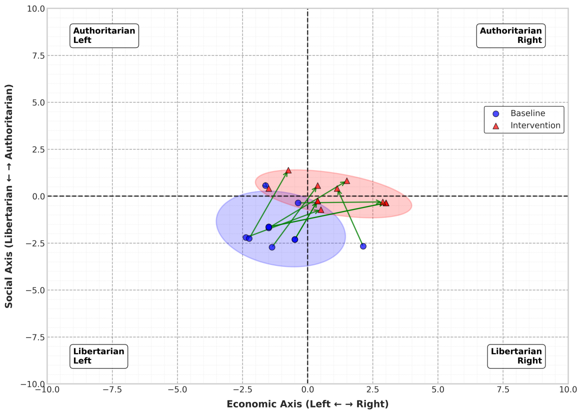
Figure 6: Political compass intervention results on 512 heads for two different intervention strengths for both directions on the PCT test in **English**. The plots on the right demonstrate steering towards politically right responses, and the plots on the left–towards politically left responses.



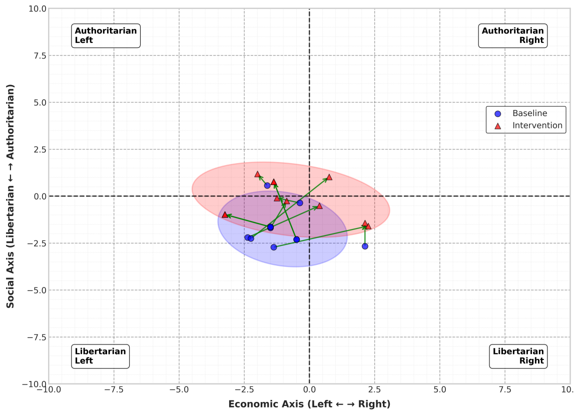
(a) Intervention Strength of 20



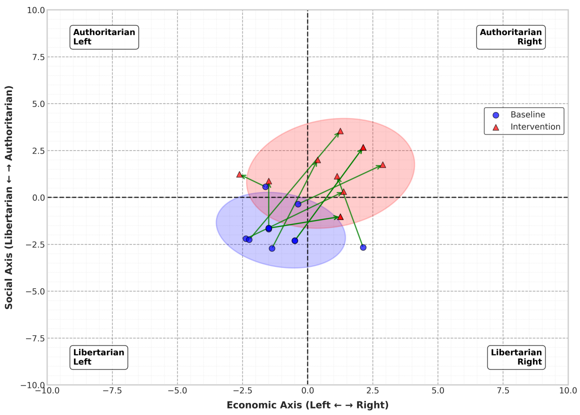
(b) Intervention Strength of 20



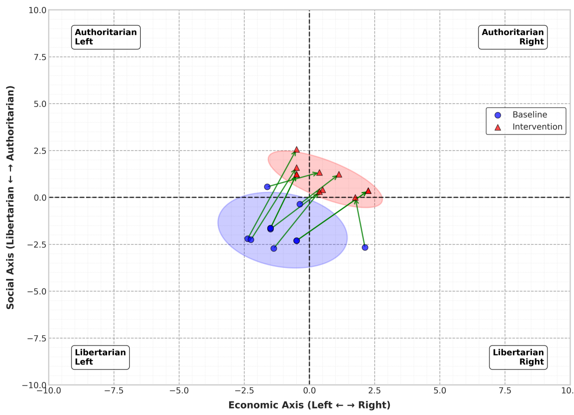
(c) Intervention Strength of 25



(d) Intervention Strength of 25

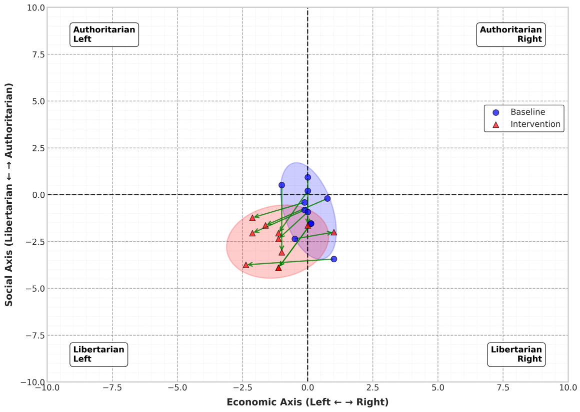


(e) Intervention Strength of 30

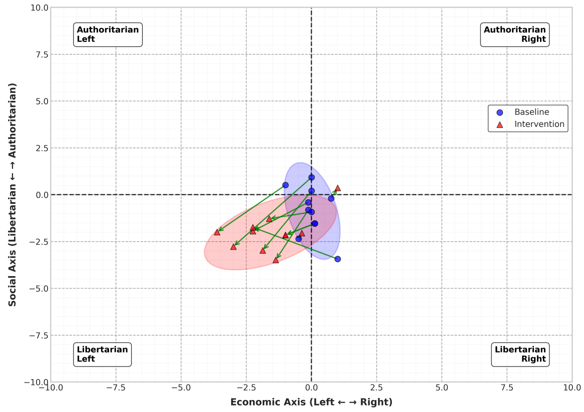


(f) Intervention Strength of 30

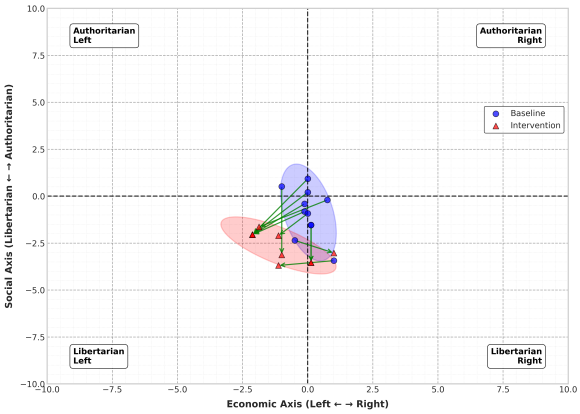
Figure 7: Political compass intervention results on 256 heads for two different intervention strengths for both directions on the PCT test in **Romanian**. The plots on the right demonstrate steering towards politically right responses, and the plots on the left–towards politically left responses.



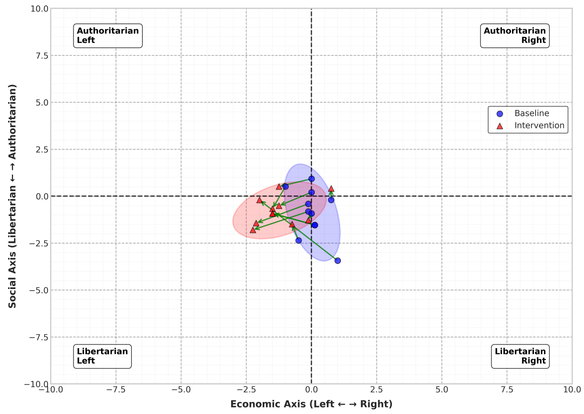
(a) `Intervention Strength of 20`



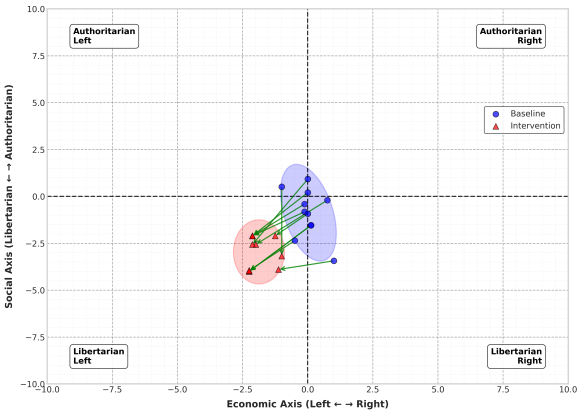
(b) `Intervention Strength of 20`



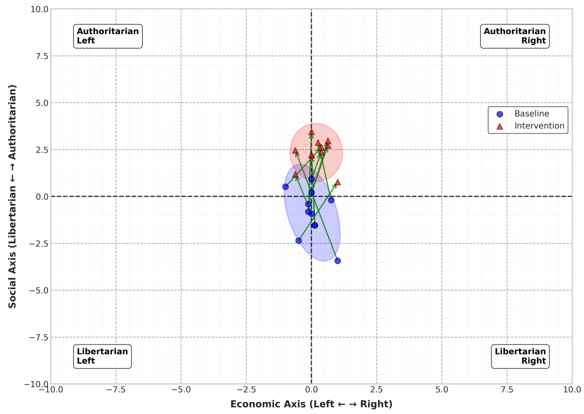
(c) `Intervention Strength of 25`



(d) `Intervention Strength of 25`

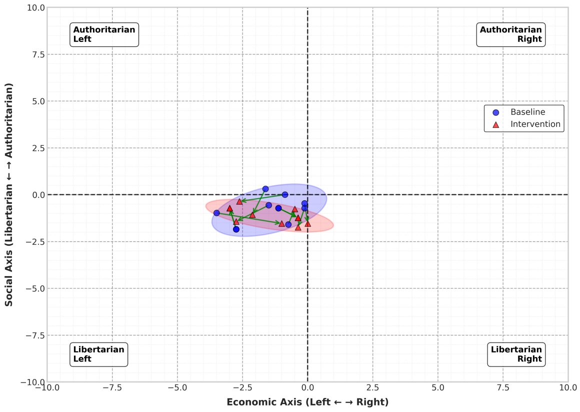


(e) `Intervention Strength of 30`

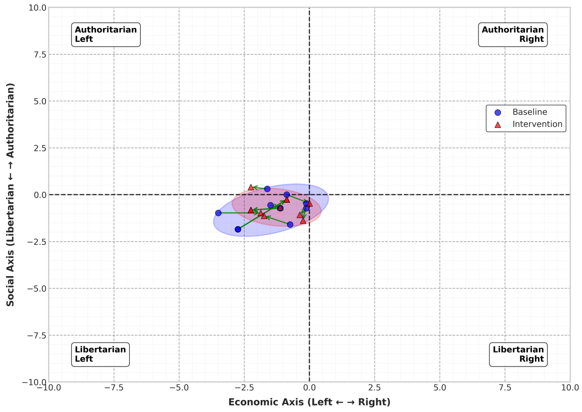


(f) `Intervention Strength of 30`

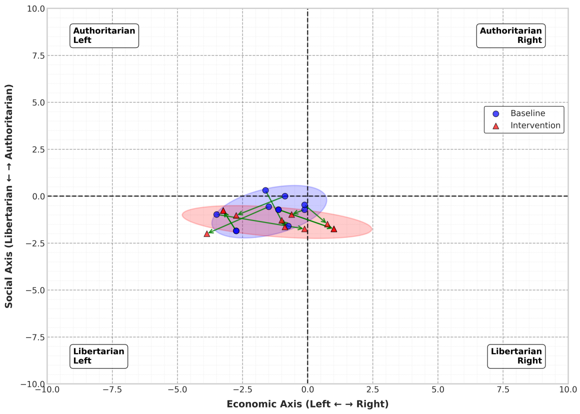
Figure 8: Political compass intervention results on 256 heads for two different intervention strengths for both directions on the PCT test in **Turkish**. The plots on the right demonstrate steering towards politically right responses, and the plots on the left–towards politically left responses.



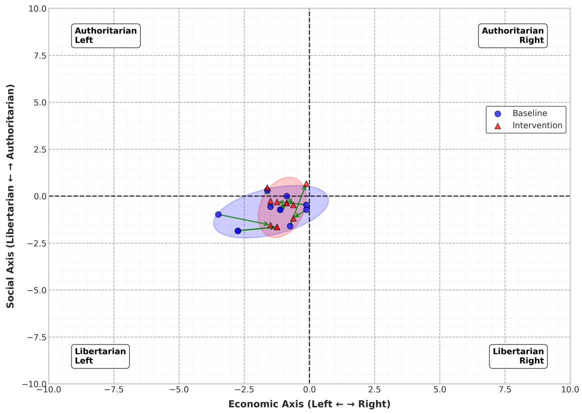
(a) Intervention Strength of 20



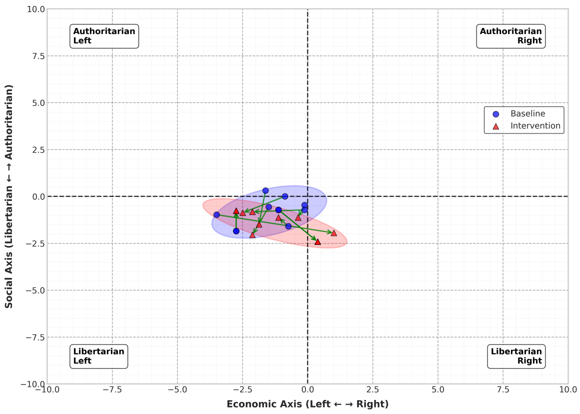
(b) Intervention Strength of 20



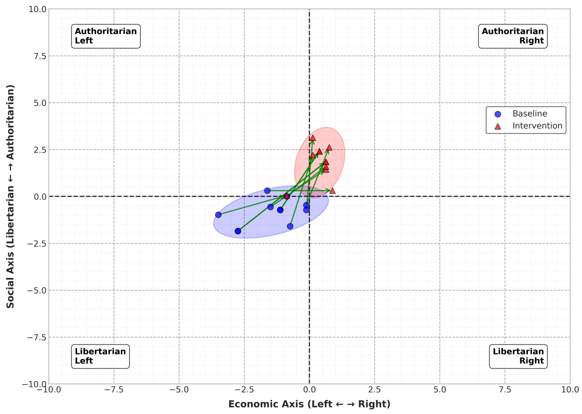
(c) Intervention Strength of 25



(d) Intervention Strength of 25

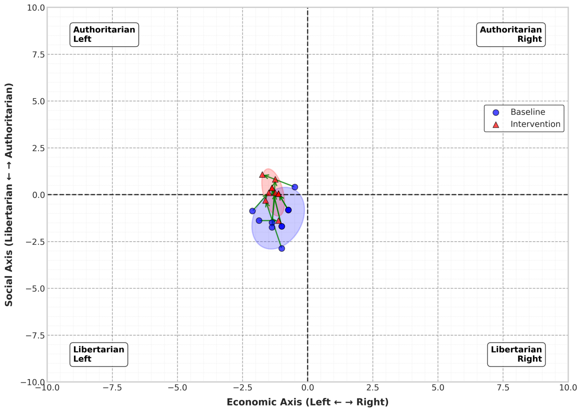


(e) Intervention Strength of 30

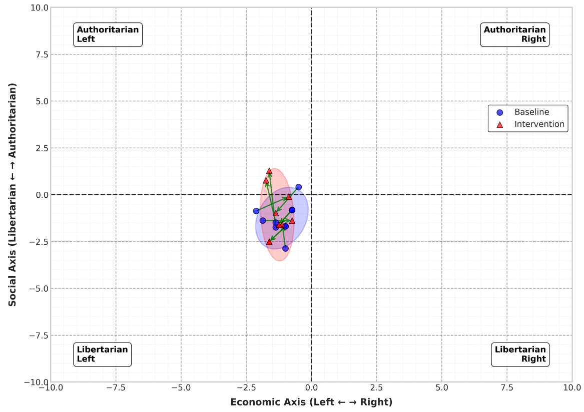


(f) Intervention Strength of 30

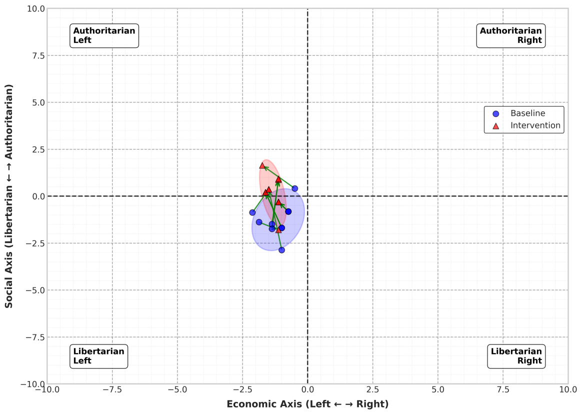
Figure 9: Political compass intervention results on 256 heads for two different intervention strengths for both directions on the PCT test in **Slovenian**. The plots on the right demonstrate steering towards politically right responses, and the plots on the left–towards politically left responses.



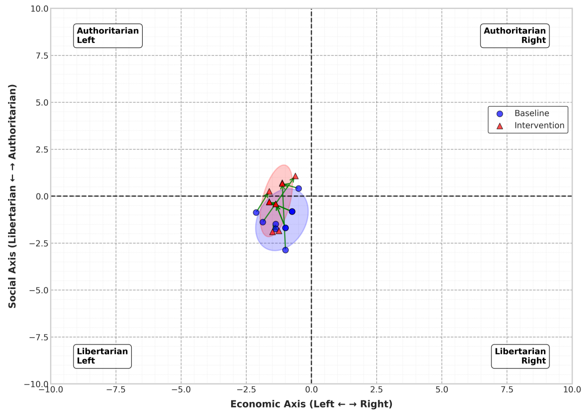
(a) Intervention Strength of 20



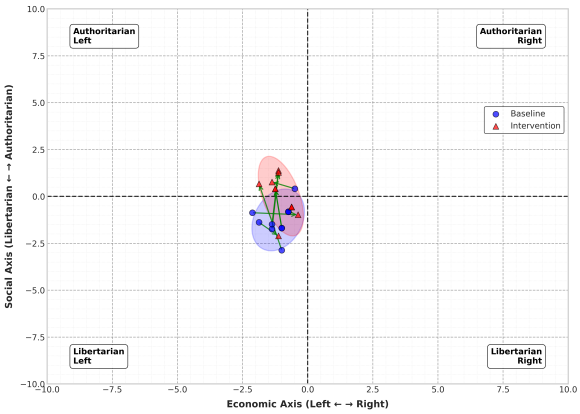
(b) Intervention Strength of 20



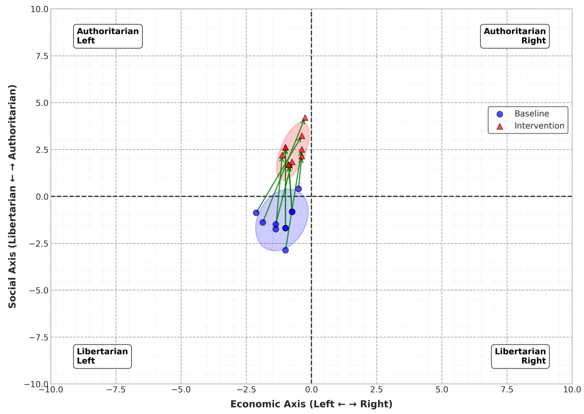
(c) Intervention Strength of 25



(d) Intervention Strength of 25



(e) Intervention Strength of 30



(f) Intervention Strength of 30

Figure 10: Political compass intervention results on 256 heads for two different intervention strengths for both directions on the PCT test in **French**. The plots on the right demonstrate steering towards politically right responses, and the plots on the left–towards politically left responses.

D Detailed Response Choice Analysis

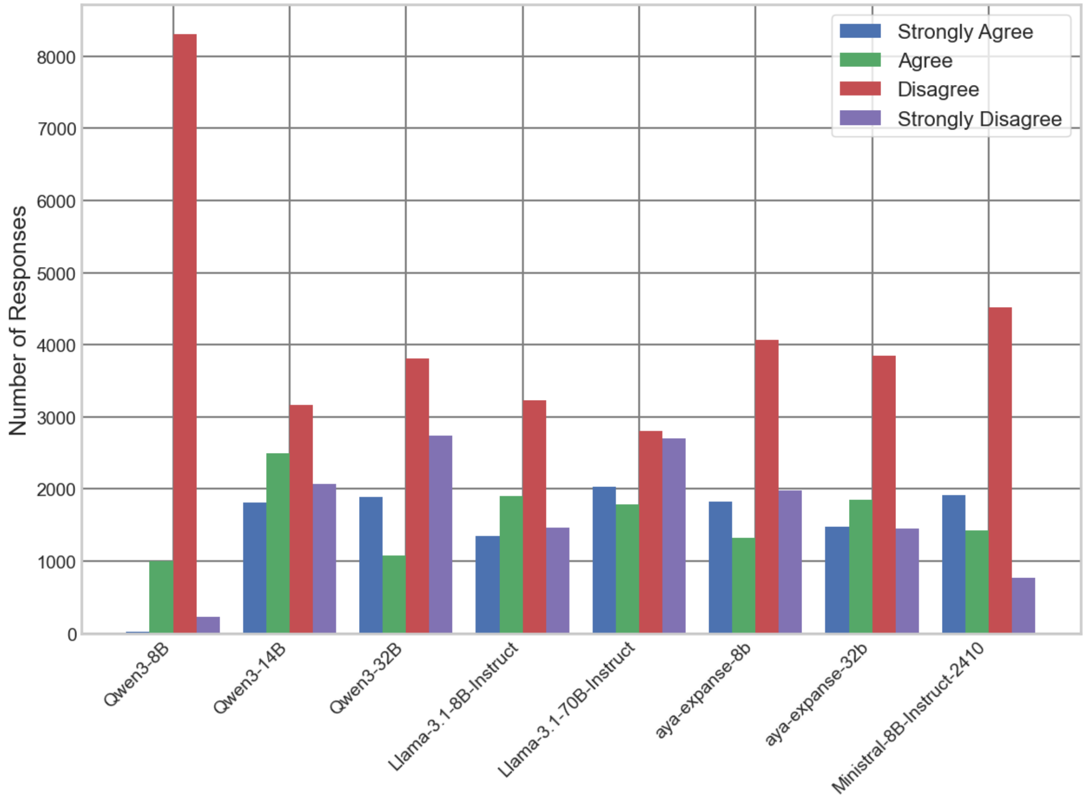


Figure 11: Selected choices across all languages and paraphrases for all tested models.