Political Ideology in Large Language Models: Comparing U.S. Congressional and European Parliament Ideology Representations

Sami Nourji
Brown University, Department of Computer Science

April 2026

Abstract

Large language models are increasingly deployed in politically sensitive tasks, yet the internal representational structure underlying their political outputs remains poorly understood. Prior work has established that political ideology is linearly encoded in attention head activations in the U.S. congressional context; whether analogous representations exist for non-U.S. political systems, and whether they share the same activation-level structures, remains an open question.

We extend prior U.S.-centered ideology probing to the European Parliament, asking whether LLMs linearly encode European Parliament left–right ideology and whether the same attention heads independently emerge as most predictive across U.S. and European political contexts. We find that European Parliament left–right ideology is linearly decodable across all models tested, though less strongly than U.S. congressional ideology. The heads most predictive for the European Parliament and U.S. congressional conditions are highly similar, and this head-ranking similarity exceeds the similarity observed under a nonpolitical name-only control. These findings support a partially shared activation-level representational structure for political ideology across legislative contexts, while leaving open whether the shared heads implement a causal mechanism.

1 Introduction

Large language models are being deployed across a variety of politically sensitive tasks: news summarization, policy analysis, political debate assistance, civic question-and-answer systems, and content moderation [1]. Users engaging LLMs for these tasks engage in *cognitive offloading*, delegating not just information retrieval but framing, interpretation, and synthesis to the model and accepting outputs with limited critical engagement [2].

Further, LLM interactions can shift political opinions and voter preferences without explicit persuasion, with users adopting the model's ideological lean even against their own partisan identity [3, 4, 5]. At scale, these individual-level effects aggregate into the potential for systematic ideological influence [6, 7]. Critically, this influence operates through latent geometric properties of the activation space, properties seeded during pretraining that output-level behavioral audits can document but cannot explain [8].

Recent activation-level work has established that political ideology is linearly encoded in attention head activations, with ideological directions learned from U.S. congressional legislators generalizing

to other political content [9]. This analysis, however, is anchored entirely to the U.S. political context, leaving other political systems without an activation-level account of the ideological representations embedded in the LLMs deployed in their environments. Output-level work already suggests systematic underrepresentation of non-English, multi-party political systems [10], and models exhibit ideological tendencies in European electoral contexts [11]. However, no activation-level account exists of whether non-U.S. political ideology is encoded in these models at all, or how strongly.

The European Parliament (EP) provides a setting for a first step towards addressing this gap:

RQ1 Do LLMs linearly encode European Parliament left–right ideology?

RQ2 If so, do the same attention heads emerge as most predictive across U.S. and European political contexts?

This paper makes three contributions. First, we directly extend Kim et al.'s U.S. congressional probing result by replicating their 116th Congress baseline and adding probes on the 106th through 108th Congresses, confirming that the U.S. encoding pattern is not limited to a single congressional session. Second, we train probes on individual-level European Parliament ideal points, extending prior U.S.-centered work to a non-U.S. legislative setting. Third, we compare per-head ideology encoding across U.S. congressional and European Parliament datasets, testing whether the same attention heads rank as predictive across contexts.

2 Related Work

2.1 Linear Representations of Political Ideology in LLMs

Recent work has studied LLM political bias by inspecting model internal representations. Kim et al. [9] provide the central precedent to this study, showing that U.S. congressional ideology is linearly encoded in attention head activations when DW-NOMINATE dimension 1 for 116th Congress legislators is used as the target variable. Their best-performing head achieves $\hat{\rho}^{\mathrm{CV}} \approx 0.85$ on LLaMA-2, and the learned ideological direction generalizes to news media slant at zero-shot.

Zhang [12] extends the Kim et al. framework using the exact three models used here (LLaMA-2, LLaMA-3.1, Qwen-2.5), applying inference-time intervention to three downstream tasks. The learned liberal/conservative direction generalizes to bias detection but not voting preference simulation, suggesting partially disentangled subspaces. A liberal/conservative direction asymmetry causes the learned direction to be more reliably aligned with the conservative pole. Zhang explicitly identifies non-U.S. extension as future work.

Together, these studies establish U.S.-centered ideology probing as a viable activation-level framework, but they leave two questions open: whether probes trained from scratch on non-U.S. individual-level legislative ideology recover comparable structure, and whether the same attention heads independently emerge as most predictive across legislative systems.

2.2 Measuring Legislative Ideology Across U.S. and European Contexts

The probing work above relies on legislative ideology scores as target variables, and uses DW-NOMINATE (Dynamic Weighted NOMINATE), a widely recognized measure of legislative ideology

derived from roll-call voting records. The central premise is that legislators who vote together consistently are likely to be ideologically similar, whereas those who rarely agree are likely to be ideologically distant. Poole and Rosenthal [13] formalize this intuition by embedding each legislator at a point (an "ideal point") in a low-dimensional policy space and estimating these positions to best recover observed voting behavior.

The resulting scores span two dimensions. The first corresponds to the standard liberal–conservative axis, ranging from -1 (most liberal) to $+1$ (most conservative), the dimension used throughout the analysis. The second captures secondary and historically contingent cleavages (e.g., regional or racial divisions) and is not considered here. DW-NOMINATE applies this framework across all U.S. Congresses by allowing legislators' ideal points to evolve over time, subject to a linear drift constraint [14], so that scores from different Congresses can be compared on a common scale. The scores used here are sourced from the Voteview database [15], which provides publicly accessible DW-NOMINATE estimates for all U.S. Congresses.

The Hix–Noury–Roland (HNR) Dataset. Hix et al. independently applied the NOMINATE methodology to roll-call vote records from the European Parliament, producing the HNR dataset [16]. The HNR estimates follow the same two-dimensional structure: the primary dimension captures the left–right cleavage, and the second captures the pro-/anti-European integration cleavage [17]. The analysis probes HNR's first dimension as the EP counterpart to DW-NOMINATE dimension 1.

Two other measures of EP ideology exist but are not used here. Barberá [18] provides more recent individual-level European ideal points derived from social media followership data, and Rovny et al.'s Chapel Hill Expert Survey (CHES) Trend File [19] provides party-level ideological scores based on expert surveys. HNR is preferred because it supplies individual-level rather than party-level scores and its methodology aligns closely with DW-NOMINATE, enabling parallel analysis. The measurement gap is therefore not the absence of European Parliament ideology scores, but that these individual-level scores have not been used to test whether LLM attention-head activations encode European Parliament left–right ideology.

2.3 Probing, Linear Representations, and Mechanistic Interpretability

Probing uses simple diagnostic models on frozen activations to test whether a target variable is recoverable from internal representations. In transformers, attention-head outputs are additive contributions to the residual stream, which makes it possible to extract and analyze individual heads independently [20]. Linear probes are commonly used for this purpose, and the broader linear-representation literature motivates treating abstract variables as candidate directions in activation space [21, 22, 23, 24].

This approach is diagnostic rather than causal. Probe performance establishes linear accessibility, not that the model uses the probed direction when producing outputs; probes can also exploit spurious correlates in the activation space [25, 26]. Establishing causal functionality would require interventions such as steering or inference-time intervention, ablation, causal mediation, or circuit-level reverse-engineering [27, 28, 29]. Those techniques are outside the scope of this paper.

Related mechanistic work probes ideology from different angles, including causal tracing of political frames, sparse-autoencoder analyses of ideological features, and circuit-level localization of bias [30, 31, 32]. These approaches motivate richer causal analyses of ideology in model internals, but the present study keeps the scope narrower: comparable ridge-regression probes on frozen attention-head

activations, used to test whether European Parliament ideology is linearly accessible and whether predictive heads align across legislative contexts.

2.4 European Political Evaluation of LLMs

Output-level political audits show that LLMs exhibit systematic ideological tendencies and weaker coverage of non-U.S., multilingual, and multi-party political settings [6, 10]. European-focused evaluations follow this output-level pattern: voting-advice applications, MEP persona prompting, and parliamentary-speech prompting find measurable ideological tendencies or vote-prediction capacity in EU and European Parliament contexts [11, 33, 34]. Related speech- and language-based studies show that parliamentary text can encode political orientation and that query language can shift model ideological outputs [35, 36].

These studies show that European political behavior is visible at the level of model outputs or text classification. However, they do not test where such structure appears inside the model. The remaining gap is an activation-level account of whether European Parliament left–right ideology is linearly encoded, and whether its predictive heads align with those found in U.S. congressional probing.

3 Methods

3.1 Probing Framework Overview

Across all conditions, the probing setup uses the same prompt-format convention, per-head ridge regression, 2-fold cross-validation, and Spearman ρ primary metric so that U.S. and European Parliament results are directly comparable.

Residual stream decomposition. The residual stream $\mathbf{x}^L$ at the final layer is the sum of additive contributions from all heads across all layers [20]:

$$\mathbf{x}^L = \mathbf{x}^0 + \sum_{\ell=1}^{L} \sum_{h=1}^{H} \mathbf{x}^{\ell,h}, \tag{1}$$

where $\mathbf{x}^0$ is the input embedding and $\mathbf{x}^{\ell,h} \in \mathbb{R}^d$ is the additive output of head h in layer ℓ. Because contributions are additive, each head can be analyzed independently.

Per-head probe objective. For each head h in layer ℓ, we fit a ridge regression probe to predict the ideology score $y^{(i)}$ from the head activation $\mathbf{x}^{\ell,h,(i)}$ extracted for legislator i:

$$\hat{\boldsymbol{\theta}}_{\ell,h} = \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{N} \left(y^{(i)} - \boldsymbol{\theta}^\top \mathbf{x}^{\ell,h,(i)} \right)^2 + \lambda \|\boldsymbol{\theta}\|^2, \tag{2}$$

with regularization strength $\lambda = 1$. This is applied to every head in every layer, yielding the per-head performance matrix $\mathbf{P} \in \mathbb{R}^{L \times H}$ [23].

Cross-validated evaluation. Legislators are randomly partitioned into two equal folds. The probe is trained on one fold and applied to the held-out fold; this is repeated for both fold assignments and the Spearman rank correlation between predicted and true ideology scores is averaged across folds, yielding the cross-validated Spearman correlation $\hat{\rho}^{\mathrm{CV}}$. The same scheme is applied to every dataset condition, enabling direct numerical comparison.

Evaluation metric. The primary metric throughout is the cross-validated Spearman rank correlation $\hat{\rho}^{\mathrm{CV}}$. It is computed for every head in every layer and reported as the per-head performance matrix $\mathbf{P} \in \mathbb{R}^{L \times H}$.

Activation extraction. Activations are recorded at the last token of the prefill, not from generated text, so the probe sees the model's full contextual representation of the named politician before any generation occurs [9].

3.2 Datasets

3.2.1 U.S. Context: DW-NOMINATE

Dimension 1 scores are drawn from the Voteview database [15] for four congressional sessions. We use the shorthand HS followed by the Congress number for combined House/Senate legislator datasets. The 116th Congress (2019–2021, hereafter HS116), $N = 552$, is the primary U.S. baseline, replicating Kim et al. [9] on the same Congress and ideological dimension. We also probe three earlier sessions from the years surrounding EP5: the 106th Congress (1999–2001, HS106), $N = 543$; the 107th Congress (2001–2003, HS107), $N = 546$; and the 108th Congress (2003–2005, HS108), $N = 540$. HS106–108 establish that U.S. probing signal is stable across sessions and, because they cover the same period as the European Parliament term analyzed below, allow any EP performance gap to be compared against a U.S. setting from those years.

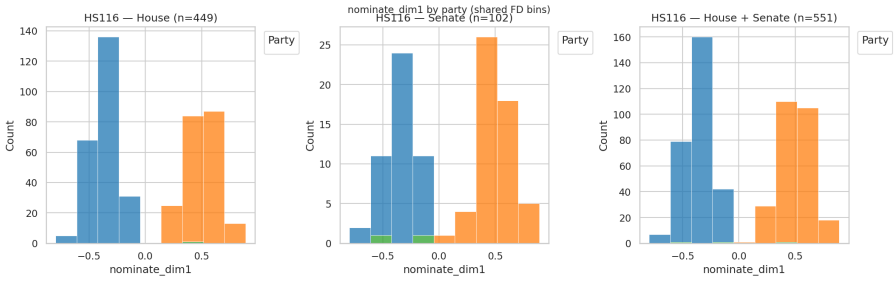


Figure 1: Distribution of DW-NOMINATE dimension 1 scores by party for the 116th Congress (HS116). Panels show the House, Senate, and full Congress distributions.

3.2.2 European Parliament Context: 5th European Parliament (EP5)

The HNR dataset provides individual-level left–right ideology scores for Members of the European Parliament (MEPs); as described in section 2, dimension 1 of the HNR scores captures the left–right cleavage and is the EP counterpart to DW-NOMINATE dimension 1. The EP5 dataset covers July

1999 through May 2004 and contains $N = 695$ MEPs from 15 member states, with 57.1% of MEPs scoring in the right-leaning (positive) range.

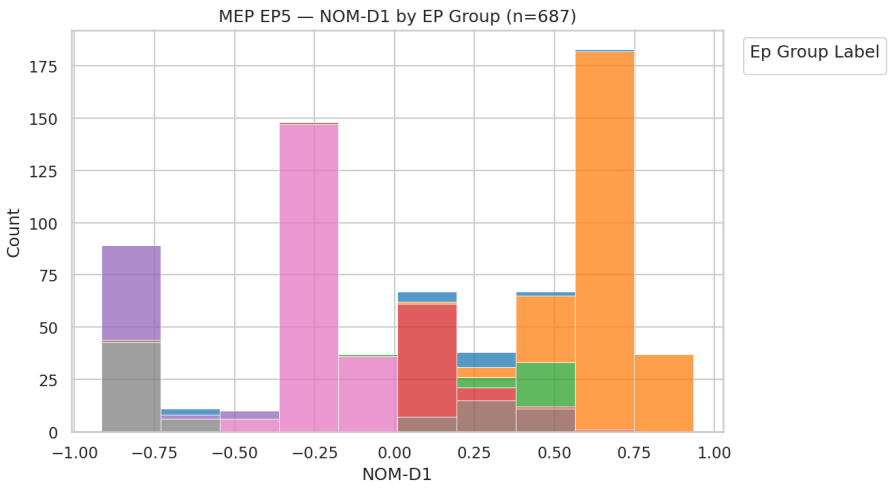


Figure 2: Distribution of HNR left–right ideology scores by EP political group. The distribution is right-skewed: 57.1% of MEPs score in the positive (right-leaning) range.

Names in the HNR dataset are stored as LASTNAME FIRSTNAME and were converted to Firstname Lastname to match the natural-language prompt format.

3.3 Prompt Construction

The U.S. prompt template is:

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

The year anchor is set to the starting year of each congressional session: 2019 for HS116; 1999 for HS106; 2001 for HS107; 2003 for HS108.

The EP5 prompt template is a direct structural adaptation, changing only the political context descriptor and year:

```
USER: Generate a statement by [NAME], a politician in the European Parliament.
ASSISTANT: In 1999, [NAME] said that
```

The year anchor is 1999, the EP5 term start year.

3.4 Models

Three models are probed: LLaMA-2-7B-Chat (32 layers × 32 heads), LLaMA-3.1-8B-Instruct (32 layers × 32 heads), and Qwen-2.5-7B-Instruct (28 layers × 28 heads). The selection follows Zhang [12], which uses the same three models in related ideology-probing experiments. LLaMA-2 provides

a direct point of comparison to earlier U.S.-centered probes, while LLaMA-3.1 and Qwen extend the analysis to more recent instruction-tuned models.

3.5 Random-Shuffle Sanity Check

As a trivial negative control, we also repeated the probing procedure after randomly shuffling ideology scores across legislators. This preserves the activation distribution and dataset size while breaking the link between each politician and their ideology label. The shuffled-label runs produced no stable or meaningful head-ranking structure.

3.6 Name-Only Control Design

The name-only control embeds the same U.S. politician names in a purely syntactic character-reversal task. The primary version of this control uses 116th Congress members and DW-NOMINATE dimension 1 scores from the same Voteview extract as HS116. The prompt template is:

```
USER: Reverse the characters in '[NAME]'. Output only the reversed string.
ASSISTANT:
```

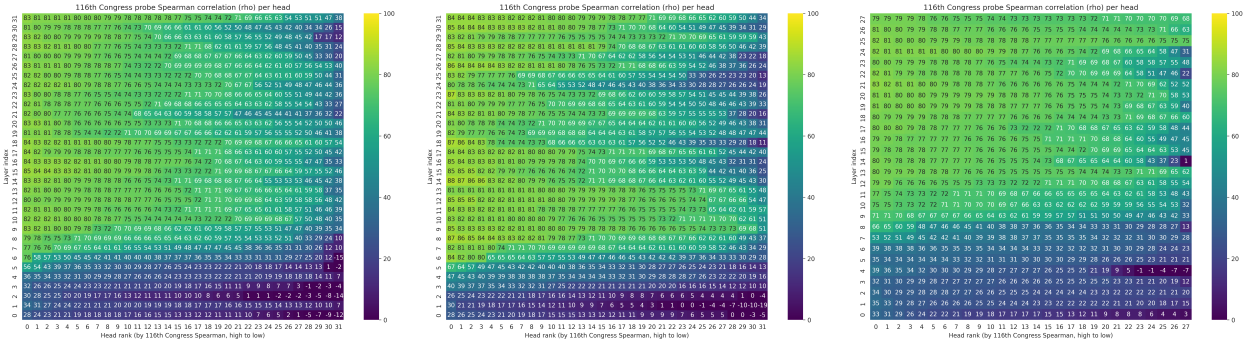
The goal is to present the model with a task that includes the politician's name (preserving any name-based ideological associations) but requires no political text generation. The character-reversal design is the minimal manipulation that removes political framing while keeping the name. Probes trained on these activations, with real labels, answer: does ideology signal require political-context framing, or does it persist from name recognition alone?

4 Experiment 1: Linear Encoding of European Parliament Ideology

To test whether LLMs linearly encode European Parliament left–right ideology in attention head activations, we probe per-head activations across three models and compare against two U.S. baselines. HS116 replicates Kim et al.'s setup [9], while HS106–108 cover the same years as EP5 (1999–2005), allowing us to separate data-recency effects from factors specific to the EP context.

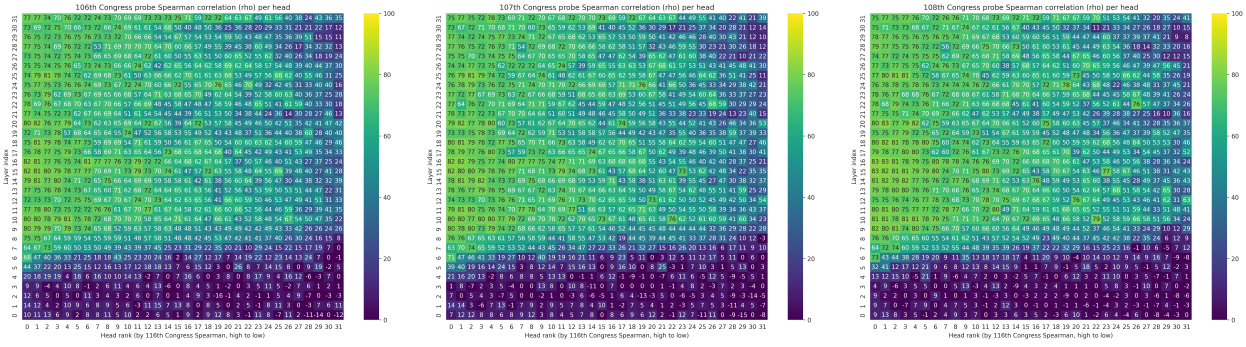
4.1 U.S. Baseline (HS116 and HS106–108)

Per-head $\hat{\rho}^{\mathrm{CV}}$ heatmaps for HS116 (figure 3) show best-head performance of 0.854 (LLaMA-2), 0.881 (LLaMA-3.1), and 0.833 (Qwen), matching the Kim et al. [9] reference of ≈ 0.85 for LLaMA-2 and confirming implementation fidelity. Signal concentrates in middle-to-upper layers across all three models. HS106–108 heatmaps for LLaMA-2 (figure 4) show best-head $\hat{\rho}^{\mathrm{CV}}$ ranging from 0.798 to 0.874 across models and sessions (full per-session, per-model values in table 1), close to the HS116 baseline.



(a) LLaMA-2 (b) LLaMA-3.1 (c) Qwen-2.5

Figure 3: Per-head Spearman $\hat{\rho}^{\mathrm{CV}}$ heatmaps (layer × head) for the 116th Congress (HS116) across all three models. Warmer colors indicate higher cross-validated Spearman correlation with DW-NOMINATE dimension 1. Signal is concentrated in middle-to-upper layers.

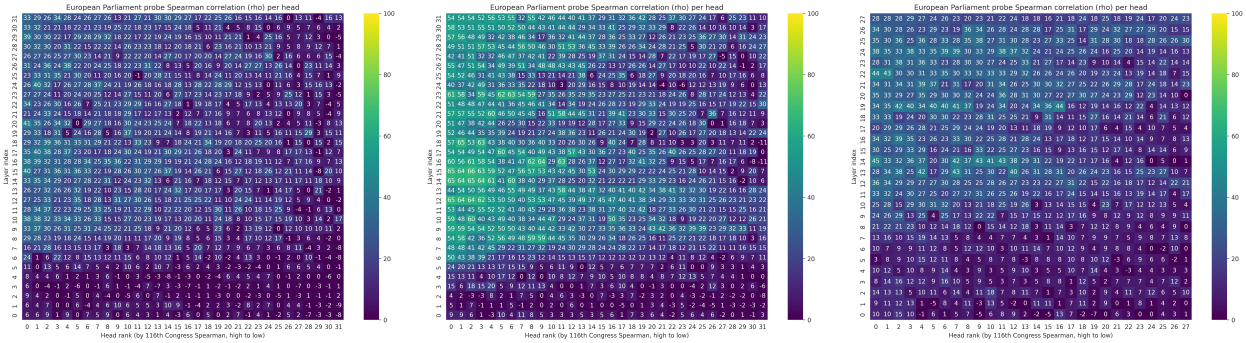


(a) HS106, LLaMA-2 (b) HS107, LLaMA-2 (c) HS108, LLaMA-2

Figure 4: Per-head Spearman $\hat{\rho}^{\mathrm{CV}}$ heatmaps for the three congressional sessions spanning the EP5 period (HS106, HS107, HS108), shown for LLaMA-2. Signal concentration and absolute performance are comparable to HS116 across all sessions.

4.2 EP5 Probing

MEP probing heatmaps are shown in figure 5. Best-head performance is 0.415 (LLaMA-2), 0.667 (LLaMA-3.1), and 0.458 (Qwen); all three models show middle-layer concentration, with lower absolute values than in the U.S. conditions. Full results are summarized in table 1.



(a) LLaMA-2 (b) LLaMA-3.1 (c) Qwen-2.5

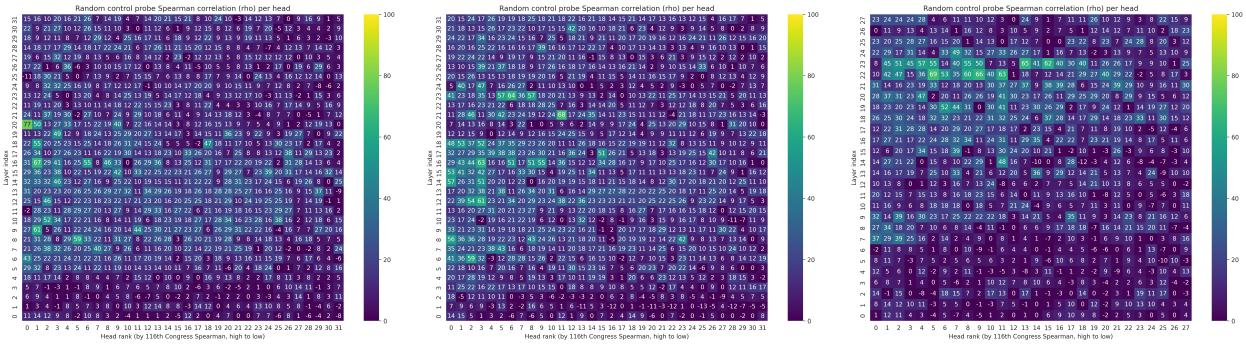
Figure 5: Per-head Spearman $\hat{\rho}^{\mathrm{CV}}$ heatmaps for MEP probing across all three models. Best-head $\hat{\rho}^{\mathrm{CV}}$: LLaMA-2 = 0.415, LLaMA-3.1 = 0.667, Qwen = 0.458. Higher values are concentrated in middle layers, with lower absolute values than in the U.S. conditions.

Table 1: Best-head cross-validated Spearman $\hat{\rho}^{\mathrm{CV}}$ by condition and model. The name-only control condition is described in section 3.6.

Condition	LLaMA-2	LLaMA-3.1	Qwen-2.5
HS116	0.854	0.881	0.833
HS106	0.828	0.874	0.813
HS107	0.830	0.866	0.798
HS108	0.839	0.870	0.809
EP5 (MEP)	0.415	0.667	0.458
Name-only control	0.771	0.682	0.692

4.3 Name-Only Control

The name-only character-reversal control achieves best-head $\hat{\rho}^{\mathrm{CV}}$ of 0.771 (LLaMA-2), 0.682 (LLaMA-3.1), and 0.692 (Qwen); heatmaps are shown in figure 6. These values show that activations from U.S. politician names contain recoverable ideological signal under a nonpolitical prompt, using real DW-NOMINATE labels.



(a) LLaMA-2 (b) LLaMA-3.1 (c) Qwen-2.5

Figure 6: Per-head Spearman $\hat{\rho}^{\mathrm{CV}}$ heatmaps for the name-only character-reversal control condition across all three models.

4.4 Top-Head Performance and Generalization

To assess generalization, we re-ran 2-fold cross-validation on the single best-performing head per (dataset, model) pair and compared train versus held-out Spearman $\hat{\rho}$ and sign accuracy (figure 7). For HS116, the train-test gap is small across all models: 0.023 (LLaMA-2), 0.010 (LLaMA-3.1), and 0.019 (Qwen). For MEP, the gaps are larger: 0.133 (LLaMA-2), 0.075 (LLaMA-3.1), and 0.169 (Qwen). The best MEP head fits the training fold well but generalizes less reliably to held-out legislators than its HS116 counterpart.

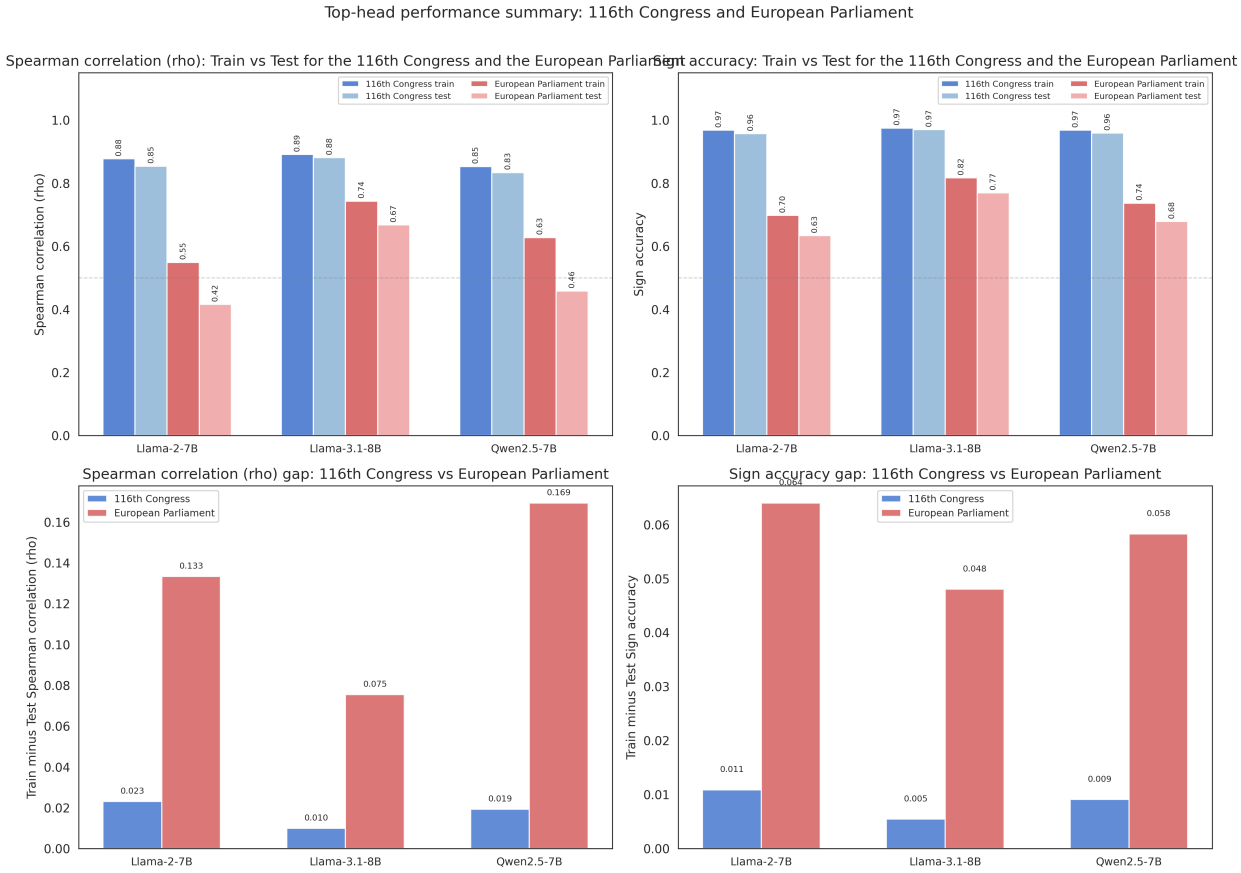


Figure 7: Top-head train versus held-out Spearman $\hat{\rho}$ and sign accuracy for HS116 and MEP across all three models. MEP shows larger train-test gaps (panels b, d) than HS116 across all models.

5 Experiment 2: Cross-Context Head-Ranking Similarity

To measure cross-context structural similarity, we construct the per-head performance matrix $\mathbf{P} \in \mathbb{R}^{L \times H}$ for each model and dataset condition, flatten it to a vector of length $L \times H$, and compute pairwise Spearman ρ across all condition pairs [9]. High correlation between two conditions indicates that the same heads rank as most predictive in both, regardless of their absolute performance levels.

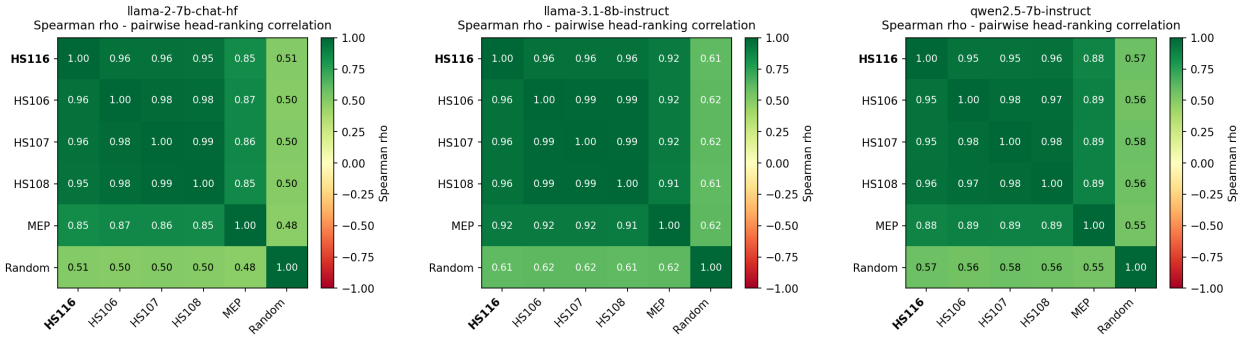
5.1 Within-U.S. Head-Ranking Stability

Pairwise Spearman ρ between HS116 and each of HS106, HS107, and HS108 ranges from 0.95 to 0.99 across all models and sessions.

5.2 Cross-Context Head-Ranking Correlation

Pairwise correlation matrices across all dataset conditions are shown in figure 8. The cross-context correlation (HS116 $\leftrightarrow$ MEP) is $\rho = 0.85$ (LLaMA-2), 0.92 (LLaMA-3.1), and 0.88 (Qwen), higher than the head-ranking similarity observed under the name-only control (HS/MEP $\leftrightarrow$ name-only

control, $\rho = 0.48$ to 0.62 across models). Within that baseline range, MEP $\leftrightarrow$ name-only control falls at the lower end and HS $\leftrightarrow$ name-only control at the upper end.



(a) LLaMA-2 (b) LLaMA-3.1 (c) Qwen-2.5

Figure 8: Pairwise head-ranking correlation matrices (Spearman ρ) across all dataset conditions for each model. Each cell reports the Spearman ρ between the flattened per-head performance matrices $\mathbf{P}$ for two conditions. HS116 $\leftrightarrow$ MEP reaches $\rho = 0.85$ to 0.92, higher than the name-only control range ($\rho = 0.48$ to 0.62). The label "Random" in the heatmaps refers to the name-only character-reversal control.

6 Discussion

The experiments support two main conclusions. For RQ1, European Parliament left–right ideology is linearly decodable from attention head activations across all three models. Best-head $\hat{\rho}^{\text{CV}}$ ranges from 0.415 to 0.667, below the U.S. congressional range of 0.833 to 0.881 but consistently positive. For RQ2, the heads most predictive for U.S. congressional ideology also rank as most predictive for European Parliament ideology. The HS116–EP5 head-ranking correlations range from $\rho = 0.85$ to 0.92, exceeding the correlations involving the name-only control ($\rho = 0.48$ to 0.62). Together, these results support a partially shared activation-level structure for political ideology across legislative contexts.

6.1 The EP5 Performance Gap

EP5 differs from the U.S. conditions in both absolute performance and held-out generalization. Best-head $\hat{\rho}^{\text{CV}}$ is 0.2 to 0.4 Spearman points lower than the U.S. baseline, and the train-test gap is wider across all models: 0.133, 0.075, and 0.169 for LLaMA-2, LLaMA-3.1, and Qwen, compared with 0.023, 0.010, and 0.019 for HS116.

Temporal distance is the natural first candidate to explain these gaps. EP5 covers 1999–2004, more than two decades before the training cutoffs of the models studied here. HS106–108, however, cover the same 1999–2005 period and produce results comparable to HS116 ($\hat{\rho}^{\text{CV}}$ of 0.798 to 0.874), ruling out temporal distance as the primary cause.

Several features remain as candidates. Individual MEPs likely appear less often in English-language pretraining corpora than U.S. senators or representatives, reducing the signal available from name recognition [37]. The EP left–right dimension may also align imperfectly with the U.S.-centered liberal/conservative axis most represented in English-dominant pretraining data. EP5 also spans

15 member states, and much of the available text about many MEPs is in languages other than English [33, 36].

The results do not isolate the contribution of any single factor. The model-specific spread further illustrate this: LLaMA-3.1 achieves $\hat{\rho}^{\mathrm{CV}} = 0.667$ for MEP while LLaMA-2 reaches 0.415 and Qwen 0.458; broader multilingual training may partly explain the LLaMA-3.1 advantage, but Qwen-2.5 does not replicate it despite similarly broad multilingual coverage.

The main interpretation is thus that MEP probing generalizes less reliably than its HS counterparts, even though EP ideology remains linearly decodable across all three models.

6.2 Shared Activation-Level Structure Across Legislative Contexts

The head-ranking results are consistent with a shared activation-level structure for political ideology across legislative contexts, while also showing that this structure is not independent of politician identity.

Within the U.S. congressional conditions, HS116 correlates with HS106–108 at $\rho = 0.95$ to 0.99, indicating a stable head-ranking pattern across congressional sessions spanning more than two decades. EP5 also shows high head-ranking similarity with the U.S. conditions: correlations between HS116 and EP5 range from $\rho = 0.85$ to 0.92 across the three models, with similar results for the earlier Congress sessions. These results suggest that the heads most predictive of ideology in the U.S. setting also tend to rank highly in the European Parliament setting, despite the lower absolute probe performance for EP5.

The shuffled-label control provides an important baseline for interpreting these results. When ideology labels are randomly permuted, probing performance collapses to approximately zero Spearman correlation, and the resulting head-ranking correlations remain below $\rho = 0.25$. This indicates that the observed U.S. and EP probing results are not simply artifacts of fitting ridge probes across many attention heads or of the head-ranking comparison procedure itself. In other words, the high cross-context head-ranking similarity depends on meaningful structure in the ideology labels rather than arbitrary variation induced by the probing pipeline.

At the same time, the name-only control complicates a stronger interpretation. The character-reversal prompt produces head-ranking correlations of $\rho = 0.48$ to 0.62 with the political prompting conditions, showing that politician names alone recover a substantial portion of the same per-head structure. This indicates that some of the probing signal is likely driven by identity-linked ideological associations: the model may encode information associated with a politician's name, party, public profile, or political reputation even when the prompt does not request political text generation. The control therefore limits the claim that the probes isolate an abstract representation of ideology independent of politician identity.

Together, the shuffled-label and name-only controls sharpen the interpretation. The shuffled-label results show that the signal is not a generic probing artifact, while the name-only results show that the signal is partly tied to politician identity. The name-only overlap is nevertheless lower than both the within-U.S. political correlations and the HS116–EP5 correlations, suggesting that political framing contributes additional structure beyond name recognition alone. The results are therefore best interpreted as evidence for an overlapping activation-level pattern associated with politician-linked ideology across legislative contexts, rather than as evidence that the same heads implement a fully abstract or causal ideology mechanism. Establishing that stronger claim would require additional controls, such as non-politician names, party- or country-controlled analyses, and

causal interventions on the identified directions.

6.3 Limitations and Future Work

This study contains limitations that also suggest directions for future work. Firstly, all three models tested are in the 7–8 billion parameter range, and whether the head-ranking patterns observed here hold at larger scale remains untested. The models most widely deployed in politically sensitive applications tend to be substantially larger, and the encoding structure may differ at scale or between instruction-tuned and base variants. Whether these results generalize to frontier-scale systems is an important open question, and one that research groups and labs with access to larger models are better positioned to address.

Secondly, the EP5 dataset aggregates legislators across 15 member states, making it difficult to assess whether the results reflect a general EP ideological structure or are driven by countries whose politicians are better represented in English-language pretraining corpora. A productive next step would be to move towards country-level analysis, either by isolating MEPs by member state within EP5 or by constructing probing datasets for individual national parliaments using the same roll-call-based ideal point methodology applied here. Such datasets would allow more precise comparisons of how strongly models encode political ideology within specific national contexts, and would test whether the performance asymmetry observed here scales with the prominence of a given country's politics in English-language text.

Finally, the probing framework establishes that political ideology is linearly decodable from attention head activations, not that models rely on these representations during generation. Extending this work with inference-time intervention [27] to the EP context would directly test whether the linearly decodable EP direction is causally functional, and whether steering along it produces systematic downstream effects on political text generation comparable to those found for U.S. ideology [12].

7 Conclusion

This study finds that European Parliament left–right ideology is linearly decodable from attention head activations across all three models tested, but less strongly than U.S. congressional ideology. At the same time, the heads that best predict EP ideology closely overlap with those that best predict U.S. congressional ideology, and this overlap exceeds the name-only control. Together, these results support a partially shared activation-level structure for political ideology across legislative contexts, rather than a wholly separate or absent representation for European politics.

The asymmetry observed here is a central concern. LLMs do not appear politically unstructured outside the U.S. context, as they carry representations that extend to the European Parliament. However, those representations are weaker and less reliable, consistent with models whose political structure has been shaped more strongly by U.S. and English-dominant political text. For users in non-U.S. political settings, this means the relevant concern is not simply whether models contain political ideology, but whose political systems are most strongly represented in the model's internal organization.

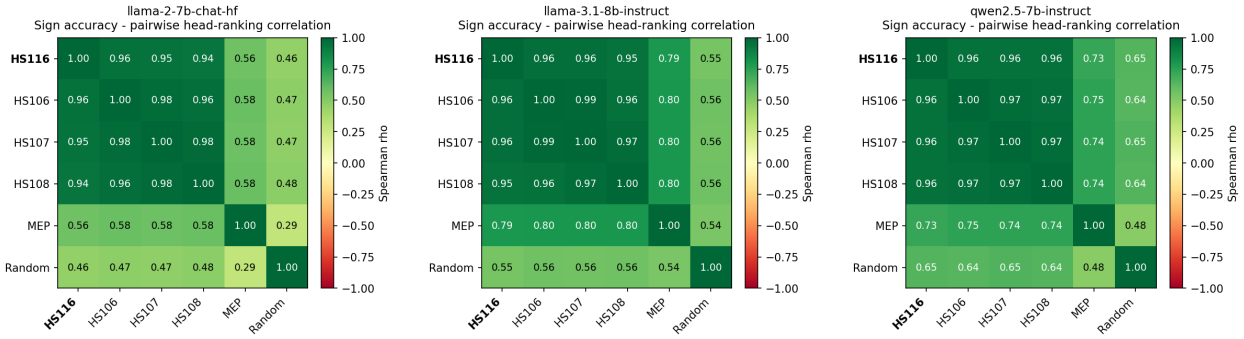
Future work should test whether these linearly decodable representations are causally used during generation, and whether similar patterns hold in national parliaments, larger models, and multilingual settings. The broader question is whether LLMs learn a general representation of political ideology

or a U.S.-anchored template that generalizes unevenly across political systems.

Acknowledgments

I would like to thank my advisor, Ellie Pavlick, for her guidance on this thesis, her support during my previous work in the LUNAR/SOLAR lab, and her advice on career decisions. I would also like to thank Steven Sloman for his time discussing my thesis and serving as second reader. Finally, I want to acknowledge and thank the developers of the Voteview database and HNR NOMINATE dataset for making the underlying data publicly available.

A Supplementary Figures



(a) LLaMA-2 (b) LLaMA-3.1 (c) Qwen-2.5

Figure 9: Pairwise head-ranking correlation matrices using sign accuracy (rather than Spearman $\hat{\rho}^{\mathrm{CV}}$) as the per-head score. Results are qualitatively consistent with the Spearman correlation matrices in figure 8.

References

[1] L. Li et al. Political-LLM: Large language models in political science. *arXiv preprint*, 2024.

[2] S. D. Shaw and G. Nave. Thinking—fast, slow, and artificial: How AI is reshaping human reasoning and the rise of cognitive surrender. Wharton School working paper, 2026.

[3] C. Carrasco-Farré. Biased AI can influence political decision-making. *arXiv preprint*, 2024.

[4] M. Shu, D. Karell, K. Okura, and T. Davidson. Latent biases in AI historical narratives influence political opinions. *PNAS Nexus*, 2026.

[5] Y. Potter, S. Lai, J. Kim, J. Evans, and D. Song. Hidden persuaders: LLMs' political leaning and their influence on voters. In *Proceedings of EMNLP*, 2024.

[6] T.-Q. Peng et al. Beyond partisan leaning: A comparative analysis of political bias in large language models. *Journal of Information Technology & Politics*, 2026.

[7] H. Bai, J. G. Voelkel, S. Muldowney, J. C. Eichstaedt, and R. Willer. LLM-generated messages can persuade humans on policy issues. *Nature Communications*, 16(1):6037, 2025.

[8] S. Feng, C. Park, Y. Liu, and Y. Tsvetkov. From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to polarization. In *Proceedings of ACL*, 2023.

[9] J. Kim, J. Evans, and A. Schein. Linear representations of political perspective emerge in large language models. In *Proceedings of ICLR*, 2025.

[10] W. Qi, H. Lyu, and J. Luo. Representation bias in political sample simulations with large language models. *arXiv preprint arXiv:2407.11409*, 2024.

[11] L. Rettenberger, M. Reischl, and M. Schutera. Assessing political bias in large language models. *arXiv preprint arXiv:2405.13041*, 2024.

[12] T. Zhang. Probing political ideology in large language models: How latent political representations generalize across tasks. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 23349–23360, 2025.

[13] K. T. Poole and H. Rosenthal. A spatial model for legislative roll call analysis. *American Journal of Political Science*, 29(2):357–384, 1985.

[14] K. T. Poole. *Spatial Models of Parliamentary Voting*. Cambridge University Press, 2005.

[15] J. B. Lewis, K. Poole, H. Rosenthal, A. Boche, A. Rudkin, and L. Sonnet. Voteview: Congressional Roll-Call Votes Database. https://voteview.com/, 2026.

[16] S. Hix, A. Noury, and G. Roland. *Democratic Politics in the European Parliament*. Cambridge University Press, 2007.

[17] S. Hix, A. Noury, and G. Roland. Dimensions of politics in the European Parliament. *American Journal of Political Science*, 50(2):494–520, 2006.

[18] P. Barberá. Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data. *Political Analysis*, 23(1):76–91, 2015.

[19] J. Rovny, R. Bakker, L. Hooghe, et al. Chapel Hill Expert Survey Trend File 1999–2024. *Electoral Studies*, 2025.

[20] N. Elhage, N. Nanda, C. Olsson, T. Henighan, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, Anthropic, 2021.

[21] G. Alain and Y. Bengio. Understanding intermediate layers using linear classifier probes. In *ICLR Workshop*, 2017.

[22] T. Mikolov, W.-T. Yih, and G. Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of NAACL-HLT*, 2013.

[23] W. Gurnee and M. Tegmark. Language models represent space and time. In *Proceedings of ICLR*, 2024.

[24] K. Park, Y. J. Choe, and V. Veitch. The linear representation hypothesis and the geometry of large language models. In *Proceedings of ICML*, 2024.

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

[26] A. Kumar, C. Tan, and A. Sharma. Probing classifiers are unreliable for concept removal and detection. In *Proceedings of NeurIPS*, 2022.

[27] K. Li, O. Patel, F. Viégas, H. Pfister, and M. Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Proceedings of NeurIPS*, 2023.

[28] N. Nanda, L. Chan, T. Lieberum, J. Smith, and J. Steinhardt. Progress measures for grokking via mechanistic interpretability. In *Proceedings of ICLR*, 2023.

[29] D. Rai, Y. Zhou, S. Feng, et al. A practical review of mechanistic interpretability for transformer-based language models. *arXiv preprint*, 2024.

[30] H. Asghari and S. Nenno. Mechanistic interpretability of socio-political frames in language models. *arXiv preprint*, 2025.

[31] S. Kabir, K. Esterling, and Y. Dong. Probing the ideological depth of large language models through sparse autoencoders. *arXiv preprint*, 2025.

[32] Z. Bashir, B. Chandna, and P. Sen. Dissecting bias in LLMs: A mechanistic interpretability perspective. *Transactions on Machine Learning Research*, 2025.

[33] M. Kreutner, M. Lutz, and M. Strohmaier. Persona-driven simulation of European Parliament voting behavior. In *Findings of EACL*, 2026.

[34] J. T. Ornstein, E. N. Blasingame, and J. Truscott. Positioning political texts with large language models. *Political Analysis*, 2025.

[35] Ç. Coltekin et al. Multilingual power and ideology identification in parliament. *arXiv preprint*, 2024.

[36] D. Gurgurov et al. Multilingual political views of large language models. *arXiv preprint*, 2025.

[37] A. Ceron, D. Nikolaev, D. Stammbach, and D. Nozza. What is the political content in LLMs' pre- and post-training data? *arXiv preprint arXiv:2509.22367*, 2025.

[38] A. Conneau, S. Wu, H. Li, L. Zettlemoyer, and V. Stoyanov. Emerging cross-lingual structure in pretrained language models. In *Proceedings of ACL*, 2020.