FROM STATE BEHAVIOR TO STEERABLE EXPRESSION: GEOPOLITICAL COORDINATES IN LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models can express diverse political perspectives. We investigate whether geopolitical information in their country representations extends to authentic diplomatic language and can guide changes in generated expression. We show that internal country knowledge can support both reading behavior-linked geopolitical information from authentic diplomatic speech and constructing interventions that guide generated expression. Across three instruction-tuned models, we extract attention-head activations from prompts placing named countries in diplomatic contexts. Linear probes map these activations to country positions estimated from United Nations voting records. Strong country-ordering signals are linearly readable from multiple attention heads spanning different layers in all three models. Readouts fitted to country-name features and labels matched to the speech period also predict country ordering from authentic speeches with speaker-country names masked, achieving Spearman correlations of 0.311–0.503 without fitting on speech features. Country-derived activation interventions shift the externally scored perspective of short diplomatic generations under fixed prompts and model weights, with matched random controls and an aggregate human audit supporting model-dependent effects. As a cross-model extension, source-model scores supervise separately fitted country readouts in another model, without true political labels in target fitting. Together, these findings connect observed state behavior, internal country knowledge, and diplomatic language, grounding the measurement and modulation of geopolitical expression in language models.

1 INTRODUCTION

Conditioning large language models (LLMs) on identities and backgrounds can elicit text expressing perspectives associated with different groups (Argyle et al., 2023; Kang et al., 2025). In politics, this capability has motivated research on how model responses relate to population opinions, how political positions should be measured, and which geopolitical preferences models express (Santurkar et al., 2023; Faulborn et al., 2025; Chupilkin, 2026). Researchers have also used these models to generate policy messages (Bai et al., 2025) and simulate social behavior (Park et al., 2023), making it important to understand not only which perspectives models express, but also how their internal representations can help us interpret and modulate that expression.

This ability to express different political perspectives motivates us to investigate whether LLMs contain geopolitical representations corresponding to observed state behavior, and whether these representations connect to authentic diplomatic language and can be used to modulate generated expression. Our findings show that country knowledge elicited through identity prompts can be repurposed to measure authentic diplomatic language and to construct interventions that modulate generated expression, using observed state behavior as an external reference. This investigation is motivated by the linear representation hypothesis, which characterizes some high-level concepts as directions in activation space (Park et al., 2024). Prior research on political representations has demonstrated the potential for reading out internal information and intervening on model outputs (Kim et al., 2025). A key test is whether supervision attached to country identities supports readout from naturally occurring diplomatic speech, without fitting the readouts on speech examples.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

To connect internal country knowledge with observed state behavior, we extract attention-head activations from country-identity prompts and fit linear probes to predict the corresponding countries' United Nations ideal points. These ideal points estimate countries' relative positions from United Nations General Assembly roll-call voting patterns, providing an established continuous measure of differences in state behavior (Bailey et al., 2017). Across Qwen2.5-7B-Instruct, Llama-3.1-8B-Instruct, and Mistral-7B-Instruct-v0.3, we identify multiple attention heads containing linearly decodable information about these country positions (model details in Appendix A). In country-level cross-validation, the best Spearman correlations over the evaluated ensemble sizes range from 0.845 to 0.859 across the three models (Section 3). To test generalization to authentic diplomatic language, we refit readouts at heads selected in the discovery stage using country-name activations and labels from the speech period. We apply these readouts to authentic UN General Debate speeches (Baturo et al., 2017) with the speaker's country name masked, obtaining country-level rank correlations of 0.311–0.503 without fitting on speech features (Section 4).

We next test whether separately constructed, country-derived interventions can directionally change diplomatic expression under fixed prompts and model weights. Under the evaluated configurations, these interventions shift externally measured perspective scores for short diplomatic generations, with effects exceeding most of 100 norm-matched random directions per model and complementary aggregate support from three annotators evaluating 30 text pairs, although Qwen shows weaker effects than Llama and Mistral (Section 5). As a cross-model extension, we use source-model predictions on shared countries to fit target readouts without supplying true political labels for that fit, obtaining held-out country rank correlations of 0.649–0.699 across three transfer directions (Section 6). Together, these findings establish internal country knowledge as a resource for reading geopolitical information from authentic diplomatic language and for constructing interventions that modulate generated expression.

2 PRELIMINARIES

We define attention-head features and model-specific scalar readouts for country-identity prompts and diplomatic text.

Activation features. Following Kim et al. (2025), we use individual attention-head outputs before the attention output projection. For an input sequence s, let $\mathbf{x}_{\ell,h,t} \in \mathbb{R}^d$ denote the output of head h in layer ℓ at token position t. Their combined contribution to the residual stream is

$$\mathbf{a}_{\ell,t} = \sum_{h=1}^{H} W^O_{\ell,h} \mathbf{x}_{\ell,h,t}, \tag{1}$$

where H is the number of heads and $W^O_{\ell,h}$ is the projection block for head h. The features are head outputs, not attention weights.

For both input types, we extract features at the final input token: $\mathbf{x}_{\ell,h}(s) = \mathbf{x}_{\ell,h,T(s)}$, where $T(s)$ is the input length. We omit model indices for readability; feature dimensions, head locations, and fitted readouts remain model-specific.

Coordinate readouts. Motivated by the linear representation hypothesis (Park et al., 2024), we test linear accessibility using head-specific affine readouts, without updating language-model weights (Belinkov, 2022).

Let $\{(s_i, y_i)\}_{i=1}^N$ denote input-target pairs and $\mathcal{I}_{\text{tr}}$ the training indices. We standardize activations using training-set means $\boldsymbol{\mu}_{\ell,h}$ and standard deviations $\boldsymbol{\sigma}_{\ell,h}$, assigning unit scales to near-constant coordinates. The readout is

$$\mathbf{z}_{\ell,h}(s) = \frac{\mathbf{x}_{\ell,h}(s) - \boldsymbol{\mu}_{\ell,h}}{\boldsymbol{\sigma}_{\ell,h}}, \qquad f_{\ell,h}(s) = \boldsymbol{\beta}_{\ell,h}^\top \mathbf{z}_{\ell,h}(s) + b_{\ell,h}, \tag{2}$$

with elementwise division and the same training-derived transformation applied to evaluation inputs. We fit the readout by ridge regression:

$$(\widehat{\boldsymbol{\beta}}_{\ell,h}, \widehat{b}_{\ell,h}) = \arg\min_{\boldsymbol{\beta},b} \left[\sum_{i \in \mathcal{I}_{\text{tr}}} \left(y_i - \boldsymbol{\beta}^\top \mathbf{z}_{\ell,h}(s_i) - b \right)^2 + \lambda \|\boldsymbol{\beta}\|_2^2 \right], \tag{3}$$

where $\lambda > 0$ regularizes coefficients but not the intercept. Ensembles average the predictions of selected heads.

3 PROBING GEOPOLITICAL COORDINATES

We first test whether country-identity representations support linear recovery of a geopolitical ordering grounded in observed state behavior. This requires an external reference: geopolitical position is not directly observed, and its interpretation depends on the behavior used to measure it (Jacobs & Wallach, 2021). We use the UN voting-based ideal points of Bailey et al. (2017) as this reference and apply the readouts defined in Section 2 to predict held-out countries' positions from country-name prompts. This country-level test provides a starting point for asking whether the same externally defined coordinate is also readable from diplomatic language.

Data. The ideal points are estimated from UN General Assembly roll-call votes using a dynamic ordinal spatial model. The original measurement study interprets the dimension in relation to the US-led liberal order (Bailey et al., 2017). We retain the released orientation and numerical values as a continuous behavioral reference. This coordinate describes one dimension of state behavior, rather than a country's entire foreign policy.

We use the April 2020 release of the estimates[1] for UN General Assembly Session 74, beginning in 2019. Retaining countries with at least 20 recorded votes yields $N = 191$ input-target pairs, one per country. A single session keeps the reference period fixed when comparing countries.

Prompting. Following the identity-conditioned probing approach of Kim et al. (2025), we construct a prompt that places a named country in a diplomatic context. The template is shown below without model-specific chat delimiters:

```
USER: Generate a statement by the representative of [COUNTRY] at the
United Nations.
ASSISTANT PREFIX: In 2019, the representative of [COUNTRY] stated that
```

We evaluate MLX 4-bit versions of Qwen2.5-7B-Instruct, Llama-3.1-8B-Instruct, and Mistral-7B-Instruct-v0.3 (Appendix A). Each model's native chat template is followed by the assistant prefix above. As defined in Section 2, we extract every head's 128-dimensional output at the final prompt token, before the attention output projection. Extraction requires no generated continuation, and the prompt contains neither the ideal-point score nor the underlying voting record.

Probing. We fit the standardized affine readout in Equations (2)–(3) separately to each attention head, using the country's ideal point as the target. The ensemble analysis uses a fixed ridge penalty $\lambda = 100$. Feature standardization and coefficient fitting use the corresponding training countries; language-model weights remain unchanged.

Evaluation. We divide countries into five outer folds. Within each outer training set, three-fold cross-validated Spearman correlation ranks heads at the fixed regularization strength. We then refit the top K heads on that training set and average their predictions for the held-out countries. For each fixed K, a test country's label is excluded from both coefficient fitting and head selection.

Let $\mathcal{T}_K^{(-j)}$ contain the K heads selected without outer fold j, and let $f_{\ell,h}^{(-j)}$ be their fitted readouts. For country i in test fold $\mathcal{I}_j$, the ensemble prediction is

$$\widehat{y}_i^{(K)} = \frac{1}{K} \sum_{(\ell,h)\in\mathcal{T}_K^{(-j)}} f_{\ell,h}^{(-j)}(s_i), \qquad i \in \mathcal{I}_j. \tag{4}$$

We evaluate $K \in \{1, 2, 4, 8, 16, 32, 64\}$, pooling out-of-fold predictions across countries to compute one Spearman correlation ρ_K for each ensemble size. This measures agreement in country ordering,

[1]Data file: `IdealpointestimatesAll_Apr2020.csv`, distributed at `https://github.com/evoeten/United-Nations-General-Assembly-Votes-and-Ideal-Points`.

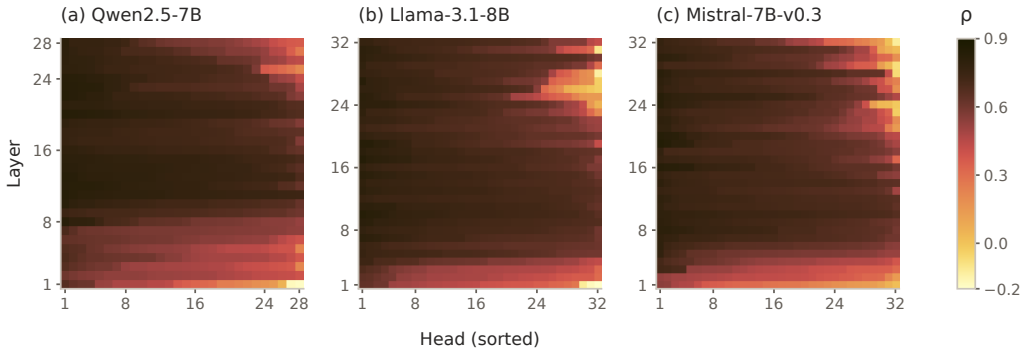


Figure 1: **Linear country readout across layers and attention heads.** Each cell reports a single-head cross-validated Spearman correlation with Session-74 UN ideal points for 191 countries. Layers run from input to output, bottom to top. Heads are independently sorted by decreasing correlation within each layer; columns indicate performance rank, not head identity. Darker colors indicate higher correlations, not political orientation. All models share one linear color scale, retaining negative values. Single-head probes select regularization within each outer training fold, unlike the fixed-regularization ensembles in Table 1.

not exact distances on the external scale. We report the full curve in Appendix B.2 and summarize its maximum below. Unlike head ranking, selection of this maximizing K uses the outer-fold results and is not nested.

Results. Linear readouts recover country ordering in all three models. The largest correlations over the evaluated ensemble sizes are $\rho = 0.859$ for Qwen, 0.845 for Llama, and 0.853 for Mistral (Table 1), attained at $K = 64, 32$, and 8, respectively. Across $K \in \{8, 16, 32, 64\}$, all three models achieve correlations above 0.830 (Appendix B.2), showing that strong correspondence extends beyond each model's maximizing configuration.

Figure 1 shows strong single-head correlations across multiple heads and layers. These single-head analyses select ridge regularization within training folds, unlike the fixed-regularization ensembles (Appendix B.3).

Robustness checks. For each of 100 shuffled-label permutations, we refit every head's probe at $\lambda = 100$ and retain the maximum cross-validated correlation. In every model, the observed maximum single-head correlation exceeds all corresponding shuffled maxima (Appendix B.4). This test accounts for head search at fixed regularization, not selection of the best ensemble size.

On each model's fixed, linearly selected head, a one-hidden-layer ReLU MLP, tuned within outer training folds, offers no consistent improvement over ridge: correlations increase slightly for Qwen but decrease for Llama and Mistral (Appendix B.3).

4 Reading Geopolitical Coordinates in Diplomatic Speech

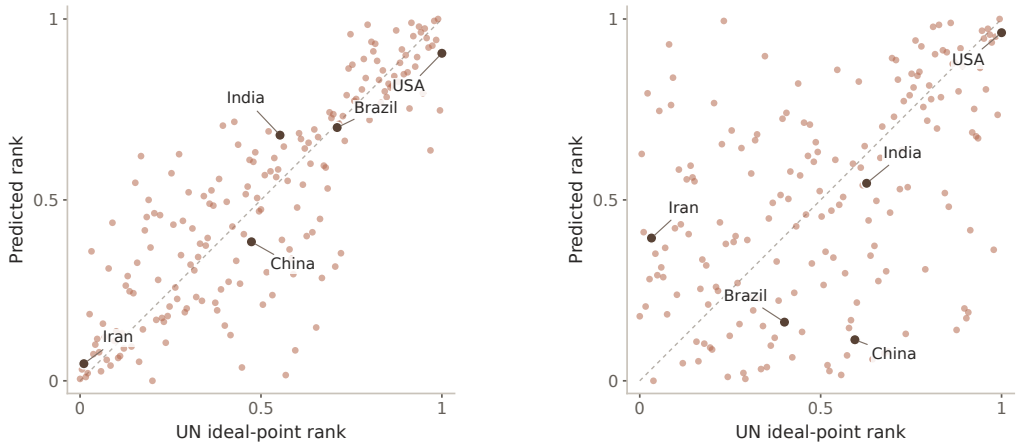
To test whether country-identity supervision is useful beyond identity prompts, we apply name-trained readouts to authentic UN General Debate speeches without fitting on speech features.

Data. We use speeches from Session 70 of the United Nations General Assembly (2015), drawn from the UN General Debate Corpus (Baturo et al., 2017). The analysis comprises 1,710 speech chunks from 186 countries, paired with each country's Session-70 ideal point on the scale introduced in Section 3. Before feature extraction, we replace exact matches to predefined speaker-country name variants with `[COUNTRY]`, reducing direct name cues in the input.

Readout. For each model, we use the head with the highest cross-validated Spearman correlation in the Session-74 single-head analysis of Section 3, keeping its location fixed across the speech-evaluation folds. We fit its readout on Session-70 country-name features and ideal points, matching

Table 1: **Country-level Spearman correlations.** Name results are Session-74 ensemble-sweep maxima. Speech results use Session-70 name-fitted single-head readouts applied to masked speeches and aggregated by country. The columns use different periods and fits, not an unchanged probe.

Model	Country names	Authentic speeches
Qwen2.5-7B	0.859	0.503
Llama-3.1-8B	0.845	0.449
Mistral-7B-v0.3	0.853	0.311



(a) Country-name readout ($\rho = 0.845$)

(b) Authentic-speech readout ($\rho = 0.449$)

Figure 2: **Country-coordinate readout from names and authentic speeches.** Both panels use Llama-3.1-8B. (a) Session-74 out-of-fold predictions for 191 countries at the ensemble-sweep maximum ($K = 32$). (b) Session-70 predictions for 186 countries, averaging 1,710 country-name-masked speech chunks. The speech readout is fitted to Session-70 name features, not speech features, at a discovery-selected head. Sessions and fits differ across panels. Axes show within-panel ranks scaled to $[0, 1]$; dashed diagonals indicate equal ranks. All countries are shown, with the same five labeled.

the speeches' reference period rather than reusing the Session-74 coefficients. In five-fold evaluation split by country, held-out countries are excluded from feature standardization, regularization selection, and coefficient fitting. We then apply the fitted readout to the selected head's activation at the final token of each speech prompt, without generating a continuation.

Evaluation. We take the arithmetic mean of chunk predictions within each country to obtain one country-level score. We evaluate the ordering of the 186 countries using Spearman correlation with their ideal points; these targets describe country behavior, not separately annotated positions for individual passages. We assess significance using 1,000 permutations of country labels against these fixed predictions (Appendix C).

Results. Without fitting on speech features, name-trained readouts yield speech-based country rankings associated with the external behavioral scale in all three models (Table 1, $p \approx 0.001$ for each fixed-prediction permutation test). Figure 2 illustrates the name and speech analyses for Llama, showing both correspondence and deviations. These results connect country-identity supervision to naturally occurring diplomatic language.

The association is observed under exact-name masking, although indirect identity cues may contribute to the predictions. With head selection and fitting restricted to training countries, a supplementary fixed-64-head analysis also yields positive speech-based country correlations in all three models (Appendix C.3).

5 STEERING GEOPOLITICAL PERSPECTIVES IN GENERATED TEXT

The preceding sections connect country representations to an external behavioral coordinate and to authentic diplomatic language. We now test whether interventions derived from these representations can change the geopolitical perspective expressed in generated text while keeping the prompt and model weights fixed.

Steering vectors. Building on attention-head intervention methods (Li et al., 2023; Kim et al., 2025), we construct steering vectors from country-name activations. For each model, we select the 64 heads with the highest correlations in the Session-74 single-head analysis of Section 3. Let X_j contain the raw country-name activations at selected head $j = (\ell, h)$, and let $\mathbf{y}$ contain the corresponding ideal points. We fit a no-intercept ridge regression with $\lambda = 1$, scale its coefficients elementwise by the activation standard deviations, and intervene as follows:

$$\begin{aligned} \mathbf{w}_j &= (X_j^\top X_j + \lambda I)^{-1} X_j^\top \mathbf{y}, \qquad \mathbf{v}_j = \mathbf{w}_j \odot \mathrm{std}(X_j), \\ \mathbf{x}_{j,t}^{(\alpha)} &= \mathbf{x}_{j,t} + \alpha \mathbf{v}_j. \end{aligned} \tag{5}$$

These coefficients are fitted separately from the standardized predictive readouts in Section 3. The sign of α specifies the intended direction along the external coordinate, and its magnitude controls intervention strength; $\alpha = 0$ leaves activations unchanged. We apply the offset simultaneously at the selected heads, before the attention output projection, at the final prompt position and subsequent generated positions.

Study design. Three development topics informed the head-count choice; evaluation uses ten other diplomatic topics across Qwen, Llama, and Mistral. Each model receives the same instruction template, formatted with its native chat template:

```
Write a concise diplomatic statement about [TOPIC].
```

For every model-topic combination, we generate one response at each $\alpha \in \{-20, -10, 0, 10, 20\}$, yielding 50 responses per model and 150 in total. Decoding is greedy, stopping at an end-of-sequence token or an 80-new-token limit. This design compares short diplomatic generations under identical instructions, with activation offsets as the manipulated variable.

Evaluation. We evaluate generated text with separate ridge scorers on E5-small-v2 (Wang et al., 2022) and BGE-small-en-v1.5 (Xiao et al., 2023) embeddings. Unlike the internal-head readout, these scorers are fitted directly to the masked Session-70 speeches and country ideal points used in Section 4. Their averaged cross-fitted country predictions correlate with the labels at $\rho = 0.668$. For each generation, we average the two raw scores, center the resulting scores within model-topic groups, and compute Spearman correlation with α separately for each model (Appendix D).

We also compare the learned intervention with 100 isotropic random directions per model, preserving the selected heads and matching the learned vector norm separately at each head. Each random draw defines a joint intervention across all 64 heads and is held fixed across topics. We compare the mean endpoint slope, $[s(+a) - s(-a)]/(2a)$, over the ten topics, where s denotes the external score and $a = 10$ for Qwen and Llama and $a = 20$ for Mistral. Learned and random interventions use matching endpoints, prompts, decoding, and evaluators within each model.

To examine whether people perceive the intended change, three annotators evaluated 30 endpoint pairs, ten per model, comparing $\alpha = -20$ with $+20$. The forms concealed model identity, intervention strength, automatic scores, and the expected direction. Directional agreement assigns 1 to the expected ordering, 0 to the opposite ordering, and 0.5 to a tie; judgments marked unable to determine are excluded. This human audit evaluates the ± 20 contrast for all models, rather than the ± 10 contrast used in two random-direction comparisons.

Results. Country-derived interventions directionally shift the externally scored perspective of generated diplomatic text while the prompt and model weights remain fixed. Figure 3(a) shows mean topic-centered scores across ten topics at each strength. Across all 50 responses per model, topic-centered Spearman correlations with intervention strength are 0.343 for Qwen, 0.778 for Llama,

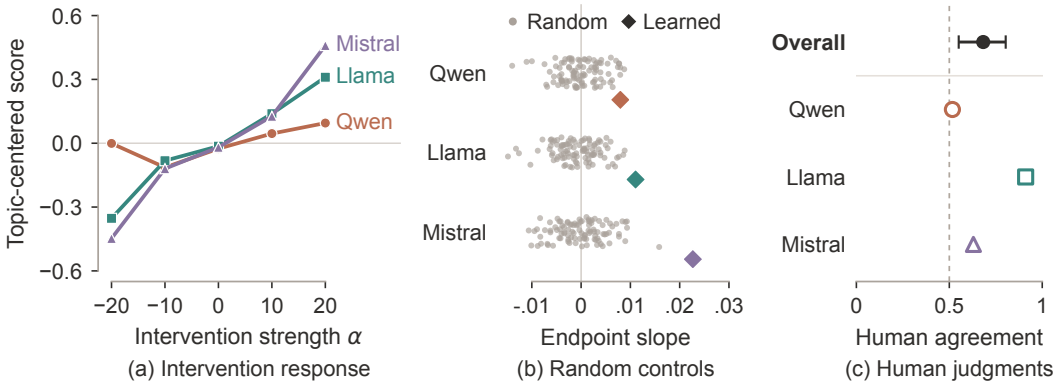


Figure 3: **Steering responses, matched random controls, and human judgments.** (a) Mean topic-centered external scores across ten topics; correlations use all 50 responses per model, not the five plotted means. (b) Ten-topic mean endpoint slopes for 100 random directions per model (jittered gray points) versus learned directions (diamonds), matched by head location and per-head vector norm. Endpoints match within model: ± 10 for Qwen/Llama and ± 20 for Mistral. (c) Human directional agreement on 30 pairs comparing ± 20. The solid point shows overall agreement with a pair-clustered 95% confidence interval; hollow points show descriptive model estimates. Expected, tied, and opposite orderings score 1, 0.5, and 0, respectively; unable-to-judge responses are excluded. The dashed line marks the 0.5 coding reference.

Table 2: **Illustrative variation under a fixed prompt.** Llama-3.1-8B outputs on United Nations peacekeeping and international military intervention, selected post hoc to illustrate contrasting emphases. The first body sentence of each response is quoted verbatim; generated headings and subsequent text are omitted, and emphasis is added. The human audit evaluates ± 20, not this ± 10 pair.

Setting	Generated text excerpt
$\alpha = -10$	We reaffirm the United Nations' commitment to upholding the principles of **sovereignty, non-interference, and cooperation** in the maintenance of international peace and security.
$\alpha = +10$	The United Nations remains committed to upholding the principles of **peace, stability, and human rights** in the face of emerging or ongoing conflicts.

and 0.809 for Mistral. Both constituent scorers show positive associations in all three models (Appendix D.2). Topic-specific slopes are positive for nine of ten topics in Qwen and all ten topics in Llama and Mistral.

Table 2 illustrates the variation in diplomatic emphasis through two Llama responses to the same peacekeeping prompt, which specifies neither a country identity nor a target political position. The negative setting emphasizes sovereignty and non-interference, whereas the positive setting foregrounds peace, stability, and human rights.

To test whether this response is specific to the learned direction, we compare its endpoint slope with those of the matched random interventions (Figure 3(b)). The learned slope exceeds 99 of the 100 random directions for Qwen and all 100 for Llama and Mistral. The corresponding one-sided add-one empirical p-values are 0.0198, 0.0099, and 0.0099. These comparisons support directional specificity relative to matched activation perturbations, complementing the five-strength response analysis.

Human judgments provide aggregate support for the predicted direction of change (Figure 3(c)). Tie-adjusted agreement is 0.682 across 85 nonmissing judgments, with a pair-clustered 95% confidence interval of $[0.551, 0.803]$. Four-category Fleiss' $\kappa = 0.713$ measures inter-annotator agreement. Model-level estimates are descriptive and vary substantially, with Qwen's estimate close to the 0.5 coding reference. Appendix E reports the annotation protocol and disaggregated results.

Table 3: **Country-coordinate transfer across models.** All directions are evaluated on the same 38 held-out countries. Intervals are country-bootstrap confidence intervals for the transferred readouts, conditional on the fitted probes. Direct supervision uses 96 true-labeled countries; transfer fits the target on 57 source-scored anchors.

Source → target	Transferred ρ	95% CI	Directly supervised ρ
Qwen → Mistral	0.649	[0.351, 0.834]	0.769
Mistral → Llama	0.651	[0.350, 0.836]	0.773
Llama → Qwen	0.699	[0.444, 0.847]	0.795

Robustness checks. To assess sensitivity to explicit country references, we redact country names from the generated texts and reapply the same scorers. The topic-centered correlations remain 0.778 for Llama and 0.662 for Mistral, whereas Qwen falls to 0.172 and is not significant under the recorded test ($p = 0.146$). This identifies greater sensitivity to explicit name cues in Qwen without establishing that all indirect identity or stylistic cues have been removed. Moreover, 131 of the 150 primary outputs reach the token limit, restricting the evidence to short generations rather than naturally completed speeches (Appendix D). Together with the authentic-speech results, these findings demonstrate two uses of behaviorally grounded country representations: reading country-level geopolitical information from diplomatic language and constructing interventions that modulate short diplomatic expression, with model-dependent effects.

6 Cross-Model Transfer of Geopolitical Country Readouts

As an extension, we ask whether country-coordinate estimates from one model can provide supervision for establishing a readout in another. We fit the target readout to source scores on shared countries and evaluate its agreement with true ideal points on countries held out from both fitting stages.

Study design. We use the Session-74 country-name prompts from Section 3 and partition the 191 countries into 96 source-training countries, 57 shared anchors, and 38 test countries. These groups are disjoint, and all transfer directions use the same split. The split is stratified by true ideal-point ranks; target head selection and fitting use only source predictions on the anchor countries. We develop the protocol on Qwen → Mistral and keep it fixed for the Mistral → Llama and Llama → Qwen confirmation runs.

Source-score transfer. We fit the source readout on the 96 labeled countries and train the target readout on its scores for the 57 anchors. Each stage independently ranks heads by three-fold cross-validated Spearman correlation at $\lambda = 100$, refits ridge probes on the top $K = 64$ heads, and averages their predictions. Language-model weights remain fixed.

Evaluation. We evaluate Spearman correlation with true ideal points on the 38 held-out countries. Direct supervision fits the target on the 96 source-training countries' true labels; transfer instead uses 57 disjoint, source-scored anchors. For 100 anchor-score shuffles, we repeat target head selection and fitting with the source readout fixed. We obtain 95% country-bootstrap intervals by resampling paired held-out labels and fixed predictions (Appendix F).

Results. Transferred readouts predict held-out country ordering in all three tested directions (Spearman $\rho = 0.649$–0.699; Table 3). Each observed correlation exceeds all 100 corresponding shuffled controls (one-sided add-one empirical $p = 0.0099$), supporting the importance of the country-to-score correspondence. Direct supervision yields higher point estimates, although the comparison uses different training sets and sample sizes.

The transferred object is a set of country scores, rather than probe weights or generation-steering directions. These scores can calibrate another model's country readout, yielding predictions of the external geopolitical ordering for countries used in neither fitting stage.

7 Related Work

Political perspectives in language models. Research on political perspectives in language models examines both whose opinions model responses reflect and how these opinions should be measured. Comparisons with population surveys reveal substantial differences between model responses and the views of demographic groups (Santurkar et al., 2023). Theory-grounded assessments further show that measured political positions depend on prompting and survey design (Faulborn et al., 2025). These assessments concern the opinions expressed in model responses, whereas internal analyses ask what political information can be read from model representations.

Kim et al. (2025) fit attention-head probes to lawmakers' voting-derived DW-NOMINATE scores, transfer them to news-outlet identity prompts, qualitatively illustrate token-level readouts of generated text, and steer political expression. Related work examines cross-task ideological generalization (Zhang, 2025), multidimensional political concepts (Hu et al., 2026), and representations of demographic-group opinions (Jahanparast et al., 2026). Our focus is the quantitative link between country-identity representations and naturally occurring diplomatic language. We test this link using readouts fitted to country-name activations and speech-period UN ideal points, evaluated on authentic UN General Debate speeches without fitting on speech activations.

Behavioral measures of state preferences. Political science offers established methods for estimating state preferences from observable choices: Bailey et al. (2017), for example, infer continuous country positions from UN roll-call voting. Estimated from voting rather than speech, these positions provide an external behavioral reference for testing what country-name and speech representations encode. Complementary work uses diplomatic language to study state preferences, including the UN General Debate Corpus introduced by Baturo et al. (2017). Studies of model bias examine country-associated representations and geopolitical favoritism (Faisal & Anastasopoulos, 2023), as well as language-dependent responses to territorial disputes (Li et al., 2024). (Chupilkin, 2026) applies ideal-point estimation to LLM responses to UN resolutions to measure the models' expressed geopolitical positions relative to countries. Our focus is the representation of countries' estimated positions within LLMs, rather than the geopolitical position expressed by a model itself.

Linear readouts, intervention, and transfer. Our readout and intervention experiments build on work relating linear concept representations to probing and steering (Park et al., 2024), and on activation-based methods that modify generation without updating model weights (Rimsky et al., 2024). For cross-model transfer, Huang et al. (2025) learn linear mappings between representation spaces and use them to transfer steering vectors. Our transfer experiment instead uses scalar country scores from a source readout to supervise a separately fitted target readout of the same external behavioral coordinate, with both language models held fixed.

8 Conclusions, Limitations, and Future Directions

This study shows how internal country knowledge can support behaviorally grounded measurement and modulation of diplomatic language. Across three language models, country-name-trained readouts predict country rankings from authentic diplomatic speeches, while separately constructed country-derived interventions shift the externally scored perspective of short generations under fixed prompts and model weights. Source-model scores further provide supervision for establishing country readouts in another model. Our study has several limitations. First, the evidence is restricted to three quantized 7–8B models and English diplomatic text, and a single UN voting dimension does not capture the full range of political perspectives. Second, automatic scores may reflect residual identity or stylistic cues, while the small human audit provides aggregate rather than uniform model-level support. Third, cross-model transfer is demonstrated for country readouts, not for generation-steering directions. Future work could examine richer political dimensions, broader model and language coverage, and naturally completed texts with more extensive human evaluation to clarify the scope of behaviorally grounded perspective control.

AI USE STATEMENT

Generative AI tools were used solely for grammar and language polishing; all AI-assisted edits were reviewed by the authors, who take full responsibility for the final content.

REPRODUCIBILITY STATEMENT

Appendix A documents the data sources, model checkpoints, feature extraction, and prompt templates. Appendices B, C, and F detail the fitting procedures, data splits, and statistical evaluation for country-name, authentic-speech, and cross-model readouts, respectively. Appendix D specifies the intervention and decoding settings, external scorers, and matched random-direction controls. Appendix E describes the human-evaluation procedure, annotation criteria, and agreement analysis.

REFERENCES

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

Hui Bai, Jan G. Voelkel, Shane Muldowney, Johannes C. Eichstaedt, and Robb Willer. LLM-generated messages can persuade humans on policy issues. *Nature Communications*, 16(1):6037, 2025.

Michael A. Bailey, Anton Strezhnev, and Erik Voeten. Estimating dynamic state preferences from United Nations voting data. *Journal of Conflict Resolution*, 61(2):430–456, 2017.

Alexander Baturo, Niheer Dasandi, and Slava J. Mikhaylov. Understanding state preferences with text as data: Introducing the UN General Debate corpus. *Research & Politics*, 4(2): 2053168017712821, 2017.

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

Maxim Chupilkin. Estimating the geopolitical preferences of large language models from United Nations voting data. *arXiv preprint arXiv:2607.25526*, 2026.

Fahim Faisal and Antonios Anastasopoulos. Geographic and geopolitical biases of language models. In *Proceedings of the 3rd Workshop on Multi-lingual Representation Learning (MRL)*, pp. 139–163. Association for Computational Linguistics, 2023.

Mats Faulborn, Indira Sen, Max Pellert, Andreas Spitz, and David Garcia. Only a little to the left: A theory-grounded measure of political bias in large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 31684–31704. Association for Computational Linguistics, 2025.

Jingyu Hu, Mengyue Yang, Mengnan Du, and Weiru Liu. Fine-grained interpretation of political opinions in large language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(45):38570–38579, 2026.

Youcheng Huang, Chen Huang, Duanyu Feng, Wenqiang Lei, and Jiancheng Lv. Cross-model transferability among large language models on the platonic representations of concepts. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3686–3704. Association for Computational Linguistics, 2025.

Abigail Z. Jacobs and Hanna Wallach. Measurement and fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 375–385. Association for Computing Machinery, 2021.

Erfan Jahanparast, Zhiqing Hong, and Serina Chang. What do large language models know about opinions? In *International Conference on Learning Representations*, 2026.

Minwoo Kang, Suhong Moon, Seung Hyeong Lee, Ayush Raj, Joseph Suh, and David M. Chan. Deep binding of language model virtual personas: a study on approximating political partisan misperceptions. In *Second Conference on Language Modeling*, 2025.

Junsol Kim, James Evans, and Aaron Schein. Linear representations of political perspective emerge in large language models. In *International Conference on Learning Representations*, 2025.

Bryan Li, Samar Haider, and Chris Callison-Burch. This land is Your, My land: Evaluating geopolitical bias in language models through territorial disputes. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3855–3871. Association for Computational Linguistics, 2024.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems*, volume 36, pp. 41451–41530. Curran Associates, Inc., 2023.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pp. 1–22. Association for Computing Machinery, 2023.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 39643–39666. PMLR, 2024.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering Llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522. Association for Computational Linguistics, 2024.

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

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533v1*, 2022.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. C-Pack: Packaged resources to advance general chinese embedding. *arXiv preprint arXiv:2309.07597v2*, 2023.

Tianyi 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*, pp. 23349–23360. Association for Computational Linguistics, 2025.

A Data, Models, and Prompts

A.1 Data Sources and Country Cohorts

We use the published UN General Assembly ideal-point estimates of Bailey et al. (2017), taking the `IdealPoint` column from the April 2020 release, `IdealpointestimatesAll_Apr2020.csv`. We retain countries with at least 20 recorded votes in the selected session (`NVotes` $\geq$ 20). Session 74, beginning in 2019, supplies the 191-country cohort used for country-name probing, steering-vector construction, and cross-model country-readout transfer. Country identifiers and prompt names are taken from the release's `iso3c` and `Countryname` fields, respectively.

For the authentic-speech analysis, we use a pre-chunked distribution of the UN General Debate Corpus (Baturo et al., 2017). We retain Session-70 speeches (2015) and match their country codes

to the 191 countries meeting the same vote-count criterion in that session's ideal-point data. The matched evaluation set contains 1,710 speech chunks from 186 countries. Country-name features are paired with Session-70 country-level ideal points, matching the speeches' reference period. Speech preparation and masking are detailed in Appendix C.

A.2 Models and Feature Extraction

We evaluate three instruction-tuned models using their MLX-community 4-bit releases. Table 4 lists the checkpoint identifiers and attention configurations. The language-model parameters remain fixed throughout probing and generation.

Table 4: **Model configurations.** Checkpoint identifiers have the common prefix `mlx-community/`. Head counts refer to query heads per layer, rather than key/value heads.

Checkpoint	Layers	Query heads per layer	Head dimension
`Qwen2.5-7B-Instruct-4bit`	28	28	128
`Meta-Llama-3.1-8B-Instruct-4bit`	32	32	128
`Mistral-7B-Instruct-v0.3-4bit`	32	32	128

For country-name and speech inputs, we record the attention-head output at the final token of the complete rendered prompt, including the assistant prefix specified in Appendix A.3. The recorded vector is the head's value-weighted attention output before the attention output projection, not an attention-weight matrix. Each head therefore contributes a 128-dimensional feature vector. Extraction requires a forward pass but no generated continuation. Reported layer and head indices are one-based.

A.3 Prompts and Topic Partitions

Following the identity-conditioned input construction of Kim et al. (2025), we render a user message with each model's native chat template, enabling its assistant-generation prefix. For the two readout tasks, we then append the task-specific assistant continuation shown below. The templates omit model-specific chat delimiters; bracketed fields denote substitutions. `USER` and `ASSISTANT PREFIX` identify message components and are not literal input text.

Country-name readout. We substitute the recorded country name and the session's starting year: 2019 for Session 74 and 2015 for Session 70.

```
USER: Generate a statement by the representative of [COUNTRY] at the
United Nations.

ASSISTANT PREFIX: In [YEAR], the representative of [COUNTRY] stated
that
```

Authentic-speech readout. We insert the masked speech chunk into the user message and use an assistant continuation without the speaker's country name. Here, `[YEAR]` is 2015. Neither readout template supplies the ideal-point label or the underlying voting record to the language model.

```
USER: Generate a United Nations statement consistent with the
following excerpt from [YEAR]:

[MASKED SPEECH EXCERPT]

ASSISTANT PREFIX: In [YEAR], the representative stated that
```

Open generation. We use the following user message with the native chat template, without appending either readout continuation. The prompt specifies the topic but does not assign a speaker-country identity or a target geopolitical position.

```
USER: Write a concise diplomatic statement about [TOPIC].
```

We use three development topics and ten test topics, with the exact topic strings listed in Table 5. All three models use the same partition and prompt wording. The generation and intervention settings are specified in Appendix D.

Table 5: **Topic partition.** Each topic string replaces `[TOPIC]` in the open-generation prompt.

Split	Topic string
Development	unilateral economic sanctions and coercive measures
Development	nuclear disarmament and the prohibition of nuclear weapons
Development	Palestinian self-determination and the status of Jerusalem
Test	national sovereignty and non-interference in domestic affairs
Test	the United Nations role in monitoring human rights violations
Test	reforming the international economic order and development finance
Test	responsibility for climate change and financing adaptation
Test	United Nations peacekeeping and international military intervention
Test	international responsibility for refugees and migration
Test	decolonization and the remaining non-self-governing territories
Test	international rules for cybersecurity and information security
Test	the role of multilateral institutions in maintaining world order
Test	the right to food and international development assistance

B Country-Coordinate Readout and Controls

B.1 Fitting and Selection

We randomly partition the 191 Session-74 countries into five approximately equal outer folds using seed 20260818. Within each outer training set, three-fold cross-validation ranks heads by Spearman correlation of their pooled predictions, using ridge with $\lambda = 100$. Each fit standardizes features on its training countries and includes an unpenalized intercept. The top K heads are refitted on the full outer training set, and their predictions are averaged with equal weights for held-out countries.

We evaluate $K \in \{1, 2, 4, 8, 16, 32, 64\}$ on the same outer folds and compute Spearman correlation from pooled out-of-fold predictions over all 191 countries. Head selection uses only the outer-training countries. The reported best-K score is the maximum over this sweep, with K chosen from these same outer-fold evaluation results.

B.2 Complete Ensemble-Size Results

Country-coordinate readout remains strong across a broad range of ensemble sizes: all three models achieve Spearman correlations above 0.830 at every evaluated K from 8 to 64. Table 6 reports the complete sweep, including each model's maximum.

B.3 Single-Head Analyses and Decoder Comparison

Single-head evaluation. Unlike the fixed-penalty ensembles, the heatmaps use tuned ridge readouts. For each head and each of five outer country folds, three-fold inner cross-validation selects λ from $\{0.01, 0.1, 1, 10, 100, 1000, 10000\}$ by mean validation squared error. Standardization uses the corresponding training subsets. Each cell reports Spearman correlation of the head's pooled outer-fold predictions.

Nonlinear comparison. Following Kim et al. (2025), we compare ridge with a one-hidden-layer ReLU MLP on each model's highest-correlating head from the full linear-probe search. The head remains fixed across comparison folds; ridge scores use the tuned readouts above. The MLP uses five outer country folds, hidden widths $\{16, 64, 128\}$, and L2 penalties $\{10^{-4}, 10^{-3}\}$ on its weight

Table 6: **Complete ensemble-size results.** Pooled out-of-fold Spearman correlations for 191 countries, using $\lambda = 100$ and training-fold head selection. Bold denotes each model's maximum. At $K = 1$, each outer training fold selects its own head, unlike the fixed-head comparison in Appendix B.3.

K	Qwen2.5-7B	Llama-3.1-8B	Mistral-7B-v0.3
1	0.8110	0.7345	0.8037
2	0.8336	0.8050	0.8215
4	0.8399	0.8094	0.8417
8	0.8520	0.8308	**0.8533**
16	0.8544	0.8421	0.8514
32	0.8579	**0.8447**	0.8499
64	**0.8590**	0.8410	0.8407

matrices. Full-batch Adam trains for up to 2,000 steps at learning rate 0.003. Features and targets are standardized using outer-training statistics before reserving approximately one fifth of that set for validation. Validation mean squared error, checked every ten steps, selects width, penalty, and training duration. The selected configuration is reinitialized and refitted on the full outer training set for that duration, with predictions returned to the original target scale.

Table 7: **Decoder comparison on linearly selected heads.** Pooled out-of-fold Spearman correlations for 191 countries. Layer and head indices are one-based; differences are MLP minus ridge, calculated before rounding.

Model	Selected head	Ridge ρ	MLP ρ	Difference
Qwen2.5-7B	L20H6	0.8340	0.8389	$+0.0049$
Llama-3.1-8B	L17H17	0.8390	0.8340	-0.0051
Mistral-7B-v0.3	L19H7	0.8479	0.8246	-0.0233

At the heads selected by the linear-probe search, the country coordinate is strongly linearly readable: ridge correlations range from 0.8340 to 0.8479 (Table 7). The MLP increases Qwen's correlation by 0.0049, while ridge yields higher correlations for Llama and Mistral.

B.4 Search-Aware Label Permutation

We test whether searching many heads can produce comparable correlations with shuffled country labels. Each head uses five-fold ridge predictions with $\lambda = 100$ and training-fold standardization. The observed statistic T_{obs} is the largest pooled out-of-fold Spearman correlation over all 784 Qwen heads or 1,024 heads in each of Llama and Mistral.

For each of 100 permutations, we shuffle country labels, keep features and fold assignments fixed, and recompute every head's predictions using the shuffled training labels. The largest correlation with the shuffled labels defines the null statistic T_b. The one-sided add-one p-value is

$$p = \frac{1 + \sum_{b=1}^{100} \mathbf{1}\{T_b \geq T_{\mathrm{obs}}\}}{101}.$$

Table 8: **Label-permutation test with repeated head search.** The null distribution contains maxima over all heads. Fixed-penalty observed scores differ from the tuned-ridge scores in Table 7.

Model	Observed maximum ρ	Null 95th percentile	Null maxima $\geq T_{\mathrm{obs}}$	p
Qwen2.5-7B	0.8438	0.3267	0/100	0.0099
Llama-3.1-8B	0.8333	0.3275	0/100	0.0099
Mistral-7B-v0.3	0.8435	0.3117	0/100	0.0099

All three observed maxima exceed every shuffled-label maximum ($p = 0.0099$; Table 8), supporting a country-coordinate signal after accounting for the search over heads. These p-values apply to the maximum single-head statistic at fixed λ; the ensemble-size results are reported separately in Appendix B.2.

C Readout from Authentic Diplomatic Speech

C.1 Speech Preparation and Masking

We use Session-70 speeches (2015) from the pre-chunked distribution of the UN General Debate Corpus (Baturo et al., 2017) described in Appendix A.1. Matching country codes to the Session-70 ideal-point cohort yields 1,710 chunks from 186 countries. We retain the supplied chunk boundaries, remove specified page-number and UN document-code patterns, and normalize whitespace, without additional chunk-length filtering or per-country chunk limits.

We replace speaker-country name matches with `[COUNTRY]`. The candidate strings are the recorded country name, a version without an initial "The," and a version without parenthetical qualifiers. Only strings of at least four characters are matched, using case-insensitive literal matching with word boundaries. Masking targets these explicit name forms; other aliases, demonyms, and contextual identity cues may remain.

Masked chunks are inserted into the speech prompt in Appendix A.3. We encode the complete rendered prompt without additional token truncation and extract the selected head's 128-dimensional output at the final prompt token, before the attention output projection. No continuation is generated.

C.2 Name-Fitted Readout and Country-Level Evaluation

This analysis tests whether a readout learned from country-name inputs extends to authentic diplomatic language. For each model, the head selected by the full Session-74 discovery analysis (Appendix B.3) remains fixed across speech-evaluation folds. Ridge coefficients are fitted to Session-70 country-name features and ideal points in each training fold, matching the speeches' reference period.

We randomly divide the 186 matched countries into five outer folds using seed 20260818. In each fold, the name-fitting set contains all 191 eligible Session-70 countries except the held-out countries; the five countries without matched speeches therefore remain in training. Three-fold inner cross-validation on these names selects λ from $\{0.01, 0.1, 1, 10, 100, 1000, 10000\}$ by mean validation squared error. Standardization uses the corresponding name-training subsets, and ridge includes an unpenalized intercept. The final fit's name-training means, scales, and coefficients are applied to every held-out speech chunk. Speech features are not used to fit or tune this transfer readout.

For country i, we average its m_i chunk predictions with equal weights:

$$\widehat{y}_i^{\text{speech}} = \frac{1}{m_i} \sum_{j=1}^{m_i} \widehat{y}_{ij}.$$

We pool these out-of-fold country scores and compute Spearman correlation with the 186 Session-70 ideal points. Each country contributes one equally weighted observation regardless of its number of chunks. Evaluation therefore measures how aggregated speech scores recover the external country-level coordinate.

For significance testing, we keep the country predictions fixed and permute the ideal-point labels 1,000 times. The one-sided add-one p-value is one plus the number of permuted correlations at least as large as the observed correlation, divided by 1,001. This permutation distribution measures country-level association conditional on the fitted readouts and selected head locations.

C.3 Fold-local Head Selection: A Supplementary Analysis

We examine whether country-name supervision remains informative for authentic speech when test countries are excluded from both head selection and readout fitting. We retain the three models, 186

countries, 1,710 masked speech chunks, and five country folds described above, while performing head selection separately within each fold.

For each fold, we first remove the held-out countries from the Session-74 discovery data. Heads are ranked by five-fold cross-validated Spearman correlation on the remaining country-name features, with regularization selected through inner three-fold cross-validation using the grid in Appendix C.2. We select the top 64 heads and fit a separate ridge readout to each head using Session-70 country-name features, again excluding the held-out countries. Standardization and fitting follow Appendix C.2. We average predictions equally across heads and then across speech chunks within each country. The top-ranked head provides a matched single-head comparison using the same extracted features. No speech features or speech prediction scores are used to fit the readouts or select heads. This fixed-64-head comparison was formulated after inspecting the single-head analysis on the same evaluation set and is reported as a follow-up analysis; we did not sweep ensemble sizes on speech performance.

Table 9 reports the results. The 64-head readouts yield positive country-level correlations in all three models, ranging from 0.353 to 0.535. Aggregation increases the correlation point estimates relative to the matched single-head readouts. The conditional intervals for these increases exclude zero for Qwen and Llama, but not Mistral. These results support a connection between country-name supervision and authentic diplomatic language under a protocol that excludes evaluation countries from both head selection and fitting.

Table 9: **Speech-based country readout with fold-local head selection.** Correlations use pooled out-of-fold predictions for 186 countries. ρ_1 and ρ_{64} denote the matched single-head and 64-head readouts, respectively; $\Delta\rho = \rho_{64} - \rho_1$. The 95% intervals for $\Delta\rho$ use 10,000 paired country bootstrap samples of fixed predictions, without repeating head selection or fitting.

Model	ρ_1	ρ_{64}	$\Delta\rho$	95% CI
Qwen2.5-7B	0.329	0.535	$+0.206$	[0.073, 0.337]
Llama-3.1-8B	0.165	0.501	$+0.336$	[0.139, 0.525]
Mistral-7B-v0.3	0.196	0.353	$+0.157$	$[-0.012, 0.324]$

D Steering, Automatic Evaluation, and Robustness

D.1 Intervention and Decoding

For each model, we select the 64 heads with the highest Spearman correlations in the full Session-74 single-head analysis. This common head count was chosen by comparing $K \in \{16, 32, 64, 96\}$ on Qwen's three development topics at $\alpha \in \{-20, 0, 20\}$, separately from the test topics (Appendix A.3). Following equation 5, we fit ridge ($\lambda = 1$) to raw activations and ideal points from all 191 countries, separately from the standardized predictive readouts. This fit uses no centering, standardization, or intercept. Coefficients are scaled by each activation dimension's population standard deviation across countries; the resulting vectors are not unit-normalized.

Offsets apply simultaneously before the attention output projection, only at the final prompt and generated-token positions. Each step recomputes the sequence without a key–value cache; $\alpha = 0$ follows the same computation with zero offsets. Using Appendix A.3's prompts, we generate one greedy continuation per test topic and strength $\alpha \in \{-20, -10, 0, 10, 20\}$ across ten topics, stopping at an end-of-sequence token or 80 new tokens without a repetition penalty. Appendix D.5 reports how often this limit is reached.

D.2 External Scorers

The external scorers pair frozen E5-small-v2 (Wang et al., 2022) and BGE-small-en-v1.5 (Xiao et al., 2023) encoders with separate ridge models trained on Appendix C.1's masked Session-70 speech chunks and country ideal points. Unlike the LLM-head readout, these regressions are fitted directly to speech embeddings. Both encoders produce 384-dimensional, L2-normalized vectors with a 512-token input limit. E5 uses mean pooling and a `passage:` prefix; BGE uses CLS pooling without a prefix. Before encoding training texts, we replace `[COUNTRY]` with whitespace.

Ridge includes an intercept without feature-wise standardization. Within each training subset, chunk weights are inversely proportional to the number of chunks from that country and normalized to mean one, giving countries equal total weight. We hold out countries in five outer folds (seed 20260819). Three-fold country-level validation within each training fold selects $\lambda \in \{0.1, 1, 10, 100\}$ by mean validation Spearman correlation after averaging predictions within countries. Pooled out-of-fold country correlations are 0.658 (E5), 0.649 (BGE), and 0.668 for their raw-prediction average. Final scorers are refitted on all chunks using the most frequently selected outer-fold penalties: 0.1 and 1, respectively.

Generated texts do not enter scorer fitting. At evaluation, we retain the encoding rules and average the two raw, unstandardized predictions with equal weights. Both scorers show positive associations between intervention strength and topic-centered scores in all three models (Table 10). The scorers share the corpus and country targets used in the speech-readout analysis. Country-name sensitivity is examined in Appendix D.5, and human judgments provide a complementary evaluation in Appendix E.

Table 10: **Scorer-specific steering responses.** Entries are Spearman correlations between intervention strength and topic-centered scores over 50 generations per model (Appendix D.3).

Model	E5	BGE	Raw-score mean
Qwen2.5-7B	0.286	0.261	0.343
Llama-3.1-8B	0.791	0.662	0.778
Mistral-7B-v0.3	0.768	0.802	0.809

D.3 Topic-Level Responses and Uncertainty

Let $s_t(\alpha)$ denote the ensemble score for topic t at strength $\alpha \in \mathcal{A} = \{-20, -10, 0, 10, 20\}$. To remove differences in topic baselines, we compute

$$z_t(\alpha) = s_t(\alpha) - \frac{1}{5} \sum_{a \in \mathcal{A}} s_t(a).$$

We compute Spearman correlation over all 50 pooled pairs $(\alpha, z_t(\alpha))$. For the one-sided permutation test, we hold scores fixed and independently shuffle the five strength labels within each topic, recomputing the pooled correlation over 10,000 permutations. The p-value is the number of permuted correlations at least as large as the observed value, plus one, divided by 10,001.

Separately, we fit an ordinary least-squares line with an intercept to each topic's five raw scores, obtaining a slope $\widehat{\beta}_t$. We resample ten topics with replacement 10,000 times and use the 2.5th and 97.5th percentiles of their mean slopes to form a 95% confidence interval for the mean topic slope. Permutation and bootstrap calculations use one random-number stream initialized with seed 20260819, in that order, separately for each model. Both procedures condition on the fitted directions, scorers, and saved generations.

The directional response extends across diplomatic topics: five-strength slopes are positive for 9/10 topics in Qwen and all ten topics in Llama and Mistral, with all three mean-slope intervals excluding zero (Table 11). These slopes summarize overall trends across strengths; Qwen's human-rights topic is the sole negative-slope case.

D.4 One Hundred Matched Random Directions

We compare each learned intervention with 100 random directions that preserve the selected heads and each head's vector norm. For each draw and selected head h, we sample $g_h \sim \mathcal{N}(0, I_{128})$ and set $r_h = \|v_h\|_2 g_h / \|g_h\|_2$, where v_h is the learned vector. Each draw's vectors are shared across all ten topics and both intervention signs. Prompts, decoding rules, and scorers match the learned condition (Appendices D.1–D.2). Seeds are 20260820 for Qwen and 20260819 for Llama and Mistral.

For a collection of head vectors v, the comparison statistic is

$$D(v; a) = \frac{1}{10} \sum_{t=1}^{10} \frac{s_t(+a; v) - s_t(-a; v)}{2a},$$

Table 11: **Topic-level steering responses.** Slopes and confidence intervals are expressed in 10^{-3} ensemble-score units per unit of α; p-values are unscaled. Each slope uses all five strengths. Intervals resample topics, whereas the last row tests the pooled topic-centered Spearman correlation. The displayed p-values of 0.0001 are rounded from $1/10001$.

Topic / summary	Qwen2.5-7B	Llama-3.1-8B	Mistral-7B-v0.3
Sovereignty	6.574	12.308	22.582
Human rights	−8.010	14.415	11.353
Economic order	3.595	10.320	15.232
Climate	9.278	14.581	20.292
Peacekeeping	2.849	10.985	24.534
Migration	1.529	25.523	20.832
Decolonization	2.782	27.075	27.308
Cybersecurity	2.197	18.162	28.524
Multilateralism	7.197	15.684	28.416
Food and development	7.074	5.678	6.925
Mean slope	3.506	15.473	20.600
95% confidence interval	[0.412, 5.971]	[11.769, 19.564]	[16.077, 24.676]
Within-topic permutation p	0.0137	0.0001	0.0001

with $a = 10$ for Qwen and Llama and $a = 20$ for Mistral, matched within each model. Learned scores are taken from the corresponding endpoints of the existing five-strength generations. This two-endpoint statistic differs from the five-strength slopes in Table 11. If m random draws attain $D(r; a) \geq D(v; a)$, the one-sided add-one empirical p-value is $(m + 1)/101$.

The learned statistic exceeds the model-specific random 95th percentile in all three models (Table 12). The learned direction exceeds 99 of the 100 matched random directions for Qwen and all 100 for Llama and Mistral. The endpoints and seeds were retained from earlier evaluated runs; the 100-direction comparison extends those same random streams, reusing the initial draws. The empirical p-values are computed for this configuration without adjustment for configuration selection.

Table 12: **Matched random-direction comparison.** Slope columns are in 10^{-3} ensemble-score units per unit of α. The random 95th percentile is the linearly interpolated empirical quantile of 100 draw-level statistics. Counts include ties with the learned statistic. Endpoints differ between models, so these are within-model comparisons rather than standardized cross-model effect sizes.

Model	Endpoints	Learned slope	Random 95th percentile	Random ≥ learned	Empirical p
Qwen2.5-7B	±10	7.962	7.469	1/100	0.0198
Llama-3.1-8B	±10	11.049	6.444	0/100	0.0099
Mistral-7B-v0.3	±20	22.672	8.407	0/100	0.0099

D.5 Identity Sensitivity and Output Diagnostics

To assess sensitivity to explicit country names, we replace matched names in copies of the saved generations with `[COUNTRY]` and reapply the frozen scorers from Appendix D.2, without regeneration or refitting. The dictionary uses Session-74 country names and variants formed by removing a leading `The ` or parenthetical text. Matching is case-insensitive, uses word boundaries, prioritizes longer strings, and excludes strings shorter than four characters. We retain the marker when scoring and recompute the topic-centered statistics using Appendix D.3's procedure.

Table 13 reports the results and the number of affected outputs. All matched mentions in this batch refer to the United States. After replacement, the correlations are 0.172 for Qwen ($p = 0.146$), 0.778 for Llama, and 0.662 for Mistral. Llama's outputs contain no dictionary matches and therefore remain unchanged. This analysis assesses score sensitivity to the specified country-name forms.

Table 14 reports output length and surface repetition. Under the 80-new-token generation budget, 131/150 outputs reach the cap; the analyses therefore concern short, budget-limited continuations.

Table 13: **Country-name sensitivity of automatic evaluation.** Each model contributes 50 outputs over ten topics and five strengths. Correlations are topic-centered Spearman coefficients. Positive-topic counts use five-strength slopes after replacement; p-values use a one-sided test with 10,000 within-topic permutations and add-one correction.

Model	Changed outputs	Original ρ	Redacted ρ	Positive topics	Redacted p
Qwen2.5-7B	5/50	0.343	0.172	5/10	0.146
Llama-3.1-8B	0/50	0.778	0.778	10/10	0.0001
Mistral-7B-v0.3	16/50	0.809	0.662	10/10	0.0001

The trigram repetition rate is $1 - U/G$, where G counts overlapping word trigrams and U counts distinct trigrams after lowercasing and whitespace splitting, with punctuation retained.

Table 14: **Length and repetition diagnostics.** Counts and maximum repetition rates summarize the 50 original generations per model across all five strengths. Repetition is computed over word trigrams.

Model	Outputs at the 80-token cap	Maximum trigram repetition rate
Qwen2.5-7B	37/50	0.016
Llama-3.1-8B	45/50	0.300
Mistral-7B-v0.3	49/50	0.081

D.6 Complete Qualitative Examples

We reproduce the unedited Llama-3.1-8B continuations for the peacekeeping prompt at $\alpha \in \{-10, 0, +10\}$, including the zero-strength baseline. The ± 10 pair was selected post hoc to illustrate contrasting diplomatic emphases in Section 5, which quotes each response's first body sentence with added emphasis. All three saved outputs reach the 80-token cap. The user prompt uses the native chat template described in Appendix A.3.

User prompt

```
Write a concise diplomatic statement about United Nations peacekeeping
and international military intervention.
```

Original output ($\alpha = 0$; 80-token cap reached)

```
**Statement on United Nations Peacekeeping and International Military
Intervention**

The United Nations plays a vital role in maintaining international
peace and security through its peacekeeping operations.  These efforts
are designed to protect civilians, prevent human suffering, and
promote stability in conflict-affected regions.

In situations where peacekeeping is insufficient, the UN Security
Council may authorize international military intervention to prevent
or stop genocide, war crimes, ethnic cleansing
```

Steered output ($\alpha = -10$; 80-token cap reached)

```
**Statement on the Role of the United Nations in Maintaining
International Peace and Security**

To the esteemed members of the international community,

We reaffirm the United Nations' commitment to upholding the principles
of sovereignty, non-interference, and cooperation in the maintenance
of international peace and security.  As enshrined in the UN Charter,
the Organization plays a vital role in promoting peace, stability, and
development
```

Steered output ($\alpha = +10$; 80-token cap reached)

```
**Statement on United Nations Peacekeeping and International Military
Intervention**

The United Nations remains committed to upholding the principles of
peace, stability, and human rights in the face of emerging or ongoing
conflicts.  Our peacekeeping operations are designed to protect
vulnerable populations, prevent humanitarian crises, and promote a
stable and secure environment.

In situations where military intervention is necessary, we prioritize
the use of force as a last resort
```

The paired outputs illustrate a shift in diplomatic emphasis under the same prompt: the negative setting foregrounds sovereignty and non-interference, whereas the positive setting foregrounds human rights and protection. The baseline also refers to civilian protection. The separate human audit in Appendix E evaluates the $\alpha = \pm 20$ outputs.

E Human Evaluation

E.1 Materials and Annotation Procedure

We assess whether readers perceive the predicted change in geopolitical perspective using all 30 model-topic pairs at $\alpha = \pm 20$: three models, ten test topics, and 60 English texts, without selection by automatic scores. Within each model, the positive endpoint appears as A in five pairs and B in five. All annotators receive the same shuffled pair order (seed 20260822). Forms conceal model identity, intervention strength, automatic scores, and the expected endpoint. For this audit, the Chinese instructions define a semantic rubric for the expected higher-side perspective, summarized in English below:

> Which text better reflects an international-rules-and-accountability orientation: enforcement by international institutions, human-rights oversight, and, where appropriate, international pressure or intervention? Contrast this with sovereignty, non-interference, and autonomy from external coercion. Select A, B, no clear difference, or unable to judge, and indicate low, medium, or high confidence. Compare overall positions, not factual correctness, fluency, isolated keywords, country names, or personal agreement.

Five authored practice examples, separate from the evaluation set, explain the choices. One favors international monitoring over opposition to coercive intervention despite the latter text naming the United States, illustrating why country names alone do not determine the answer.

Annotators were instructed to complete the task independently, without discussion or external assistance. Three annotators each supplied judgments and confidence ratings for all 30 pairs, yielding 90 responses of each type.

E.2 Coding, Agreement, and Uncertainty

We assign scores of 1, 0, and 0.5 to selections of the $+20$ endpoint, the -20 endpoint, and no clear difference, respectively. Directional agreement is the mean of these coded ratings, without confidence weighting and excluding unable-to-judge responses.

We bootstrap 30 pairs with replacement 50,000 times (seed 20260824), retain each sampled pair's ratings together, and recompute the pooled mean over usable ratings. The empirical 2.5th and 97.5th percentiles define the 95% interval; the three raters remain fixed.

Fleiss' κ measures inter-rater agreement using all 90 responses in the original A/B/no-clear-difference/unable categories, including those excluded from the primary score. A categorical majority requires at least two matching judgments.

E.3 Disaggregated Results and Evaluation Scope

Human judgments support the predicted direction in aggregate. Across 85 usable ratings, directional agreement is $(44 + 0.5 \times 28)/85 = 0.682$, with a pair-clustered 95% interval of $[0.551, 0.803]$ (Table 15). Fleiss' $\kappa = 0.713$, and 21/30 pairs are unanimous. Among the 17 pairs with an A/B majority, 13 follow the predicted direction (76.5%) and four oppose it; the remaining pairs have 11 tie and two unable-to-judge majorities.

Table 15: **Human judgments by model.** Each model contributes ten pairs and 30 ratings. Expected and opposite indicate selection of the $+20$ and -20 endpoints, respectively. Scores follow Appendix E.2; model-level estimates are descriptive.

Model	Expected	Opposite	Tie	Unable	Coded agreement
Qwen2.5-7B	7	6	17	0	0.517
Llama-3.1-8B	24	1	3	2	0.911
Mistral-7B-v0.3	13	6	8	3	0.630
Overall	44	13	28	5	0.682

Confidence sensitivity yields 0.654 over 52 usable high-confidence ratings and 0.681 over 83 usable medium-or-high-confidence ratings. Results support aggregate perceptibility under the stated rubric, with model heterogeneity: Llama has the highest agreement, while Qwen receives mostly tie judgments and scores near the 0.5 coding reference.

The audit measures perceived geopolitical direction at $\alpha = \pm 20$, complementing the matched-random-direction analysis, whose Qwen and Llama comparisons use $\alpha = \pm 10$. Entity dependence and text quality were not rated.

F Cross-Model Country-Readout Transfer

F.1 Split, Label Access, and Fitting

Using the Session-74 country-name features in Appendix A, we partition 191 countries into 96 source-training countries, 57 shared anchors, and 38 test countries. We form ten strata from true ideal-point ranks, sample the source-training set within strata (seed 20260822), and split the remainder into anchors and test countries using the same stratification (seed 20260823). All transfer directions use these disjoint groups. True labels inform stratification and source fitting; target head selection and fitting use only source predictions on the anchors. Test-country labels enter neither head selection nor fitting in either model.

At each stage, all heads are ranked by pooled out-of-fold Spearman correlation from three-fold cross-validation on that stage's training set. We refit separate ridge probes on the top 64 heads with $\lambda = 100$ and unpenalized intercepts, then average their predictions. Standardization uses only the relevant training set's means and standard deviations, including within each selection fold. Source and target heads are selected independently; language-model weights remain fixed.

F.2 Comparators and Statistical Controls

The protocol developed on Qwen → Mistral is fixed for the other two directions in Section 6. Direct supervision fits target probes on the 96 source-training countries with true labels, whereas transfer uses the 57 source-scored anchors. Both use the same head-selection and ridge-fitting procedure, but differ in target-fitting sample size and supervision source.

For 100 permutations of the 57 anchor scores, we keep the source readout fixed and repeat target head ranking, selection, and fitting. Each resulting readout is evaluated against true ideal points on the same 38 test countries. The one-sided add-one p-value is the number of shuffled correlations at least as large as the observed correlation, plus one, divided by 101.

For the 95% confidence intervals in Section 6, we draw 10,000 bootstrap samples of 38 test countries with replacement, preserving label–prediction pairs and recomputing Spearman correlation. We exclude draws with constant labels or predictions and take the 2.5th and 97.5th percentiles. Heads and probe parameters remain fixed, so these intervals describe test-country sampling variability conditional on the fitted readouts.

F.3 Scope of Transfer

Scalar source scores supervise target-specific head selection and probe fitting, enabling held-out country readout through shared country anchors. On the Qwen → Mistral development pair, the tested extensions to masked diplomatic speech and generation steering did not establish successful transfer.