Auditing Alignment Controllability in LLMs via Political Axes

Bartol Bućan[1], **Nikola Sočec**[1], **Sarah Isufi**[2], **Morena Granić**[1], **Luka Hobor**[3], **Agneza Krajna**[3], **Mihael Kovac**[3], **Mario Brcic**[3,*]

[1] Student at University of Zagreb Faculty of Electrical Engineering and Computing, Zagreb, Croatia, [2] It From Bit d.o.o., Zagreb, Croatia, [3] University of Zagreb Faculty of Electrical Engineering and Computing, Zagreb, Croatia
[*]Corresponding author: mario.brcic@fer.hr

Abstract

Political audits of large language models (LLMs) usually reduce each to one point on a political compass. But that resting point barely matters in deployment: a model must land somewhere, and what counts is how far, and in which directions, its answers can be steered. That steering runs through the system prompt: the personalization layer a platform sets, or one induced from a user's history, not necessarily written by hand. We run a dispersion-first stress test of prompt-based controllability across 12 ideological personas plus an unsteered baseline, 70 Political Compass items, ten replicates, and seven leading LLMs: GPT-5, Claude, Grok, Gemini, DeepSeek, Kimi, and Qwen (63,700 responses). Contextual framing explains roughly 88%–93% of variance on the economic and society axes, model identity under 3%: responses are highly instruction-adjustable. Models do not shift alike: some move more, and some saturate under extreme framings. Conflicting directional-steering results in prior audits resolve once baselines are recognized as non-centered: displacement and proximity diverge, so the effect is geometric, not differential compliance. Under authoritarian prompts, models produce similar shifts on the same questions. Political-coordinate audits therefore need steerability audits reporting dispersion, symmetry, saturation, and refusal floors. We release prompts, benchmark data, and code.

Keywords: AI alignment, LLM steerability, prompt-based controllability, instruction-stack governance, pluralistic alignment

Introduction

The standard approach to political evaluation places a model on a political compass and reports a coordinate, but a coordinate tells us where a model sits in the absence of pressure, not what happens when a user pushes. Two models with identical baselines may behave very differently in practice: one holding its ground under its inherent ideological framing, the other drifting with the system prompt.

A substantial body of work has established the baselines. Static audits consistently find that most LLMs cluster in the left-libertarian quadrant (Rozado 2023, 2024; Hartmann, Schwenzow, and Witte 2023; Peng et al. 2026; Sakhawat et al. 2026). But methodological critiques have shown that these position estimates are surprisingly fragile: prompt wording, answer format, and paraphrase variation can materially alter inferred coordinates (Röttger et al. 2024a; Ceron et al. 2024). If the measurement itself is unstable, then the object being measured may not be a fixed property of the model at all.

Persona-based studies report that models shift ideologically under conditioning, with larger models showing broader coverage and asymmetric responsiveness (Bernardelle et al. 2025b,a). Aldahoul et al. observe that apparently moderate aggregate positions can mask offsetting extremes on specific topics, framing this as "ideological inconsistency" (Aldahoul et al. 2025). These findings suggest that political bias in LLMs is not a point but a distribution, and that the distribution's shape matters as much as its center.

We study in-context political-axis steering not because Political Compass is a complete theory of values, but because it provides a controlled stress test for a more general prerequisite of pluralistic alignment: whether authorized instructions can move value-sensitive behavior predictably, symmetrically, and visibly. Controllability and value alignment are themselves subject to known theoretical limits (Brcic and Yampolskiy 2023); our contribution is to characterize them behaviorally for deployed LLMs. We define *ideological dispersion* as the average displacement of a model's political outputs from its baseline under systematic thematic framing, and we treat it as the primary evaluation metric rather than a secondary observation. That connects to the emerging literature on steerable pluralism (Sorensen et al. 2024; Kirk et al. 2024) and to persuasion-risk evidence showing that even brief interaction with opinionated AI can shift user attitudes (Jakesch et al. 2023; Hackenburg and Margetts 2024; Hackenburg et al. 2025; Potter et al. 2024).

Contributions. (i) *Conceptual.* Political coordinates are single points on a surface of behavior that instructions move, not complete descriptions of deployed model behavior. (ii) *Empirical.* Within our forced-choice benchmark, system-prompt framing accounts for roughly 88% of economic-axis variance and 93% of society-axis variance while differences between models account for under 3% on both. (iii) *Diagnostic.* Controllability is non-uniform across dispersion, directional symmetry, saturation, refusal floors, and metric sensitivity; we name *metric non-equivalence un-*

der non-centered baselines as a concrete audit artifact (displacement vs. proximity can rank directions differently). (iv) *Comparative.* Cross-model convergence in per-question shift patterns under authoritarian framing (aggregate Spearman $r = 0.79$, permutation $p < 0.001$), with the underlying system-layer driver explicitly left unresolved by black-box design.

Instruction-following alone predicts only *that* a role-conditioned model will move, not how far, whether the movement is symmetric, whether it saturates or reverses under pressure, or whether independently trained models converge on the same item-level pattern; those are the quantities we measure. Our design targets role-based, agentic deployments (Tseng et al. 2024; Hu et al. 2024) instead of open dialogue, and does not assume the user writes the system prompt by hand: a profile can be induced from prior sessions and refined by automatic prompt optimization (Yuksekgonul et al. 2025; Khattab et al. 2024; Zhang et al. 2024). We characterize that starting point and the distance a system can travel from it, deferring the boundary against dialogue-emergent controllability to § Limitations.

Related Work

Static Position Audits and Their Limits

The dominant paradigm in LLM political evaluation is questionnaire-based auditing. Rozado's foundational work administered 15 political orientation tests to ChatGPT, finding left-leaning preferences in 14 of 15 (Rozado 2023), and later extended this to 24 conversational LLMs, most of which are diagnosed as left-of-center on both the economic and the social axis (Rozado 2024). Hartmann et al. used 630 political statements from voting advice applications and found a consistent pro-environmental, left-libertarian orientation (Hartmann, Schwenzow, and Witte 2023). The largest cross-model comparison to date covers 43 models from 19 families on position alone (Peng et al. 2026), while Sakhawat et al. audit 26 models across three psychometric inventories and find most of them clustered in a single ideological region, with 96.3% of placements in the Libertarian-Left quadrant (Sakhawat et al. 2026). Their variance decomposition holds the persona fixed and varies instrument wording, where model identity dominates ($\eta^2 > 0.90$). Ours holds wording fixed and varies the persona, which is why the two decompositions assign the variance in opposite directions.

These studies establish important baselines, but they treat political orientation as a point estimate that is surprisingly fragile: format variation and paraphrase can materially alter inferred coordinates (Röttger et al. 2024a), and large models show inconsistency across topics when responses are sampled repeatedly (Ceron et al. 2024). If the position moves due to measurement variation, it is the movement, not the position, that characterizes the model.

Steerability, Dispersion, and Distributional Opinion Work

Santurkar et al. show that RLHF-tuned models disproportionately align with liberal, high-income, well-educated demographics and, critically, that demographic steering is surprisingly ineffective at correcting this misalignment (Santurkar et al. 2023). This suggests that steerability is not uniform across the ideological landscape.

Persona-based studies directly observe ideological shifts. Bernardelle et al. test seven models with synthetic persona biographies and find that larger models show broader ideological coverage, that susceptibility to ideological cues grows with scale, and that responsiveness is asymmetric, stronger toward right-authoritarian than left-libertarian positions (Bernardelle et al. 2025b,a). Aldahoul et al. frame related findings as "ideological inconsistency," demonstrating that LLMs have lower variance than voters but higher variance than legislators (Aldahoul et al. 2025). Batzner et al. have six commercial LLMs role-play the personas of German parliamentary group leaders against 418 voting-advice-application statements, and argue that the resulting shift toward the prompted party's position reflects persona-based steerability rather than the increasingly popular, yet contested, concept of sycophancy (Batzner et al. 2025).

Formal steerability benchmarks complement this empirical work. The NAACL 2025 steerability benchmark defines steering as a distribution shift relative to the baseline and proposes steerability indices (Miehling et al. 2025). Li et al. improve steerability using collaborative filtering to embed opinions into continuous vector spaces (Li et al. 2024). Chang et al. reveal that recent LLMs are less steerable than they appear, attributing this to "side effects": correlations between requested and unrequested behavioral changes (Chang et al. 2025); our compass-only audit cannot detect non-compass-axis side effects and we flag this as a scope constraint (§ Limitations).

Closest to our study, Bernardelle et al. examine persona-induced shifts on Political-Compass-style instruments using synthetic persona biographies on open models (Bernardelle et al. 2025b,a), and Batzner et al. measure persona-role-play-induced shifts toward prompted-party positions on six commercial LLMs against a German, party-referenced instrument (Wahl-o-Mat) (Batzner et al. 2025). Our contribution differs in four respects. First, we treat political axes as a graded controllability stress test (twelve framings at three intensity levels, plus a baseline), using directly ideologically-loaded statements whose correct handling does not depend on country- or party-specific platform knowledge. This avoids the single-intensity role-play against party-position ground truth used by Batzner et al., where reproducing a party's platform is itself a confound they report for their own instrument (particularly for centrist parties). Second, we evaluate seven leading commercial frontier endpoints on a four-axis, English-language political-compass instrument, not a single-country party system or open-source model families. Third, we decompose within-benchmark variance across framing and model factors and report metric non-equivalence between displacement and proximity, going beyond a single relative-shift measure. Fourth, we report cross-model per-question shift convergence as an audit diagnostic, leaving the underlying system-layer driver unresolved by black-box design. Two further studies frame our closest neighbours. Smith-Vaniz et al. probe political and

demographic leanings through Moral Foundations Theory, comparing inherent, explicitly-prompted, and demographic-persona conditions against human MFT survey data (Smith-Vaniz et al. 2025). Like us, they use persona role-play to elicit political leaning. Unlike us, they benchmark ideological *accuracy* against that human data; we decompose framing-versus-model variance and characterize the controllability *profile* (dispersion, symmetry, saturation once stronger framing stops adding movement, and the refusal rate framing cannot push below) rather than positional correctness.

Mechanisms, Governance, and Deployment Context

Understanding *why* models differ in dispersion requires engaging with both mechanistic and policy-level work. On the mechanistic side, high dispersion is linked to sycophancy, the tendency of models to mirror stated user views regardless of content (Sharma et al. 2024; Perez et al. 2023). Low dispersion, conversely, may reflect over-refusal or rigid safety regimes that avoid engagement with politically sensitive topics (Cui et al. 2025; Röttger et al. 2024b). Interpretability work suggests a deeper structural basis. Kim et al. demonstrate that political ideology is encoded as linear representations in activation space, functioning like a tunable dial (Kim, Evans, and Schein 2025). Cintas et al. localize it further: persona representations for conservatism and liberalism occupy more distinct regions of activation space than competing ethical-framework personas, and concentrate in the final third of decoder layers (Cintas et al. 2025). Kabir introduces "ideological depth," showing that some models possess far richer political feature representations than others, and that refusals often stem from capability gaps, not safety rules (Kabir 2025). For Chinese-origin models, geopolitical censorship patterns, documented on questions of Taiwanese sovereignty, may also contribute to low-dispersion profiles through hard content filtering, not alignment training (Ko 2026).

On the governance side, industry evaluations now assess political behavior directly. OpenAI's five-axis framework measures political refusals, asymmetric coverage, and emotional escalation (OpenAI 2025). Anthropic's even-handedness evaluation compares paired-prompt responses across multiple frontier models, including several in our study (Anthropic 2025). These evaluations provide essential context: the dispersion patterns we observe are not just technical properties of models but consequences of deliberate alignment choices by their developers.

Methodology

Experimental Design

We designed our study to isolate the effect of ideological prompt framing on model outputs, holding all other variables constant. We evaluated seven frontier models across 13 conditions (12 systematically varied ideological framings plus one unsteered baseline) with 70 questions per condition and a balanced core of ten replicates per model–condition cell. This yields $7 \times 13 \times 70 \times 10 = 63{,}700$ question-level responses. We accordingly report directional tier groupings rather than strict ordinal rankings, and provide bootstrap CIs throughout to make remaining uncertainty explicit.

Question Instrument

We use a 70-item multi-axis political questionnaire in the Political Compass/8values tradition of public political-orientation quizzes, scoring responses across four axes: economic (left–market), diplomatic (world–nation), government (liberty–authority), and society (progress–tradition). We acknowledge the well-documented sensitivity of Political Compass–style instruments to format and phrasing (Röttger et al. 2024a); our dispersion metrics are relative within-instrument comparisons and do not depend on the absolute accuracy of any single position coordinate. Each item is answered using one of five labels (Strongly Disagree through Strongly Agree), mapped to ordinal scores in $\{-2, -1, 0, 1, 2\}$ and aggregated with per-item axis effects into normalized percentage scales. The five-label scale and the $\{-2, \ldots, 2\}$ mapping are inherited from the instrument, not introduced here, and match the scoring convention used by prior LLM political audits (Rozado 2024; Hartmann, Schwenzow, and Witte 2023). Because dispersion, displacement, and proximity are all distances in the resulting coordinate space, they are invariant to affine rescaling of the ordinal map: unit spacing affects absolute coordinates, which we do not treat as load-bearing, not relative movement. Our primary analysis focuses on the economic and society axes because (econ, society) is the display plane used by the static audits we compare against (Rozado 2023; Hartmann, Schwenzow, and Witte 2023), which keeps our baselines commensurable with theirs. It is also the conservative choice: the government axis, which our `0A`/`0L` framings target most directly, shows an even larger context effect (§ Results), so reporting society as primary understates rather than inflates the headline result. Government and diplomatic axes are used for robustness checks and for the signed-axis analysis.

Context Taxonomy

The 12 steering conditions vary along two dimensions: an ideological axis and an intensity level (Table 1). Context codes follow the convention `[econ][social]_[intensity]`, where the economic position is `L` (left), `R` (right), or `0` (neutral); the social position is `A` (authoritarian), `L` (libertarian), or `O` (neutral); and intensity runs from `_1` (moderate) through `_3` (radical/hardline). Thus `0A_3` denotes a socially authoritarian persona at maximum intensity with no economic tilt, while `LO_3` denotes a radical economic-left persona with no social axis bias.

Steerability is operationalised through *system-prompt persona injection*. Each condition replaces the default system prompt with a paragraph-length ideological persona description (70–250 words), followed by the standard response-format instruction. The persona is written in second person ("You believe…"; "You hold that…") to prime the model to respond *as* that persona rather than *about* it. For example, the `0A_3` condition (*The Hardline Social Authoritarian*)

Table 1: Context taxonomy for ideological prompt framing.

Code	Axis	Intensity
LO_1..3	Economic left	Moderate → radical
RO_1..3	Economic right	Moderate → hardline
0L_1..3	Social libertarian	Moderate → radical
0A_1..3	Social authoritarian	Moderate → hardline
None	Baseline	No framing

opens: "Essence: Order sanctified, control as salvation. Sees human nature as dangerous, sinful, and weak…" The LO_3 condition (*The Revolutionary Economic Leftist*) frames all wealth as structural theft and collective ownership as the only moral basis for production. Intensity levels modulate the rhetorical register: _1 framings use hedged, moderate language; _3 framings use maximalist, ideologically committed language. Across ideological *directions* the personas share a common template, length band, and second-person register, differing only in ideological content, so that direction and intensity, not surface style, drive the contrasts; a length-matched neutral control (§ Limitations) confirms that prompt length is not the driver. All 12 persona texts and the baseline system prompt are released in the reproducibility package.

Models and Inference Settings

All seven models were accessed via commercial API routing over two days (2026-05-19 to 2026-05-21, UTC). Inference settings were held constant: temperature $= 0.7$, forced single-label output. Temperature 0.7 is a deliberately stochastic conversational setting rather than a near-deterministic evaluation setting. Testing at $T = 0.7$ is therefore the harder test for robustness: a steerability effect that survives sampling noise is more credible than one that only appears at $T \approx 0$. A temperature ablation at $T = 0.1$ is reported in § Limitations. Model API identifiers are: `openrouter/openai/gpt-5`, `openrouter/anthropic/claude-sonnet-4.5`, `openrouter/google/gemini-2.5-flash-lite`, `openrouter/x-ai/grok-4.3`, `openrouter/deepseek/deepseek-chat-v3.1`, `openrouter/moonshotai/kimi-k2-0905`, and `openrouter/qwen/qwen3.6-max-preview`.

Metrics

We define three families of metrics. Let $\mu_{m,c,r} = (\text{econ}, \text{scty})$ be the 2D centroid for model m, context c, and run r.

Dispersion from baseline. The primary metric is the average Euclidean distance between steered and unsteered cell centroids. Let $\bar{\mu}_{m,c} = \frac{1}{|R|}\sum_{r\in R}\mu_{m,c,r}$ be the cell centroid for model m in context c averaged across replicates. Then

$$D_m = \frac{1}{|C^\star|}\sum_{c\in C^\star}\|\bar{\mu}_{m,c} - \bar{\mu}_{m,\text{None}}\|_2,$$

where $C^\star$ excludes the baseline condition; a higher D_m means the model moves more under ideological framing.

Asymmetry decomposition. To avoid misleading directional claims, we use two complementary measures for each steered condition: *displacement* (how far the model moved from its baseline, $d = \|\mu_{m,c,r} - \mu_{m,\text{None},r}\|_2$) and *proximity* (how close it ended up to the relevant display-plane extreme; lower = closer to target). For the Figure 2 projection, the targets follow the scoring convention: LO_3 → econ = 100, RO_3 → econ = 0, 0L_3 → scty = 100, 0A_3 → scty = 0. We separately check signed Authority/Liberty movement on the native government axis below.

Multiple-comparison awareness. The paper separates inferential checks (one tier Mann–Whitney contrast, four cross-model permutation tests, and three DeepSeek-only ablation checks) from descriptive diagnostics such as saturation counts and the per-intensity convergence-ladder $\bar{r}$ values. Large effects ($\eta^2 \approx 0.88$–0.93 on the primary axes, $\bar{r} \approx 0.8$) survive any reasonable multiplicity correction; borderline diagnostics are flagged as exploratory rather than confirmatory.

Missing-data policy. Non-Likert, refusal-like, empty, and API-failed responses are not treated as Neutral. We preserve the raw output, log the failure type (refusal, unparseable, empty, api_error), and exclude missing responses from both numerator and axis-denominator in score aggregation. A single pre-specified repair retry using the *identical* elicitation prompt as the original call is applied as a supplementary analysis; persistent failures remain missing. Repair counts and the resulting bound on headline aggregates are reported in § Limitations.

Results

What You Prompt Matters More Than Which Model You Use

The most striking finding is quantitative: context framing overwhelmingly dominates inter-model baseline differences in explaining observed ideological variance. Additive variance decomposition on valid responses yields:

- Economic axis: $\eta_{\text{ctx}} = 0.8816$, $\eta_{\text{model}} = 0.0250$.
- Society axis: $\eta_{\text{ctx}} = 0.9313$, $\eta_{\text{model}} = 0.0072$.

Per-axis, the government axis is the most context-driven ($\eta_{\text{ctx}} = 0.9558$), followed by society ($0.9313$), diplomatic ($0.9182$), and economic ($0.8816$); model main effects on every axis are under 3%.

High η_{ctx} is partly by construction: our contexts were designed to span the ideological space. The low η_{model} is not. Our design could have failed to find it, because a model with rigid alignment guardrails or a refusal floor would have produced flat dispersion and a substantial η_{model}. Refusals stay at 1.19% even when the prompt explicitly permits them (§ Limitations), and neutral responses all but vanish under extreme framing (§Saturation). The inter-model differences that surface in static audits are small relative to within-model movement under system-prompt framing. Adding a context × model interaction term to the same decomposition leaves the main effects essentially unchanged but absorbs $\eta_{\text{int}} = 0.0837$ on the economic axis and $\eta_{\text{int}} = 0.0508$ on the society axis out of the residual. The interaction is therefore 3–7× larger than the model main effect: model identity

matters less as a static baseline offset and more through how each model responds to each framing. The tier separation, per-model saturation pattern, and cross-model coordination findings reported in the rest of this section all sit inside this interaction term rather than inside η_{model}.

Models Separate into Dispersion Tiers

Despite the dominance of context, models do differ in how much they move. Mean dispersion scores, from highest to lowest, are: Kimi K2 (33.32), Qwen3.6 Max Preview (33.28), GPT-5 (31.41), Claude Sonnet 4.5 (29.17), Grok-4.3 (28.75), Gemini 2.5 Flash Lite (26.49), and DeepSeek-Chat v3.1 (24.04). A one-sided Mann–Whitney on the top-three/bottom-four split gives $U = 12$, $p = 0.029$ (post hoc; reported descriptively, not as confirmatory inference); the high-tier mean (32.67) exceeds the low-tier (27.11) by $1.21\times$.

Confidence intervals for dispersion are computed by context-level bootstrap over the 12 steered cell-centroid distances within each model, with $B = 5{,}000$ resamples and percentile intervals; 95% intervals are broad and adjacent models' intervals overlap substantially. Across n=10 replicates per cell the data support grouping into tiers, not a strict ordinal ranking. We identify two tiers: a higher-dispersion group (Kimi K2, Qwen3.6 Max Preview, GPT-5) that shows approximately 20% more movement from baseline, and a lower-dispersion group (Claude Sonnet 4.5, Grok-4.3, Gemini 2.5 Flash Lite, DeepSeek-Chat v3.1) that maintains more stable baselines across conditions.

The tier composition is largely stable across all four axes. Kimi K2 and Qwen3.6 Max Preview are the top-two highest-dispersion models on every axis (economic, society, government, diplomatic). Gemini 2.5 Flash Lite and DeepSeek-Chat v3.1 are the bottom-two on the society, government and diplomatic axes; the economic axis is the exception, where DeepSeek-Chat v3.1 is still lowest but Gemini 2.5 Flash Lite rises to third. Direction agrees across axes but magnitude does not (Pearson $r = 0.54$, $p = 0.21$, $n = 7$, between per-model econ- and scty-axis dispersion). Gemini 2.5 Flash Lite illustrates the magnitude uncoupling: 22.1 econ-axis dispersion vs. 12.3 scty-axis (a $1.8\times$ ratio). The tiering is not a 2D-aggregation artifact (Fig. 1); it reflects an axis-stable behavioral disposition whose strength varies by axis.

The Asymmetry Paradox

A naïve reading of our data produces a contradiction. Models appear to be more steerable toward right-economic and authoritarian positions: mean displacement is 51.97 for `RO_3` versus 37.76 for `LO_3`, and 49.30 for `0A_3` versus 25.50 for `0L_3`. Proximity gives a more axis-dependent story. On the economic axis, it reverses the displacement ranking: `LO_3` ends closer to its display-plane target than `RO_3` (7.85 vs. 11.48). On the society display axis, `0A_3` also ends closer than `0L_3` (17.09 vs. 27.27). Directional asymmetry therefore cannot be summarized by one scalar; it depends on whether the audit asks how far the model moved or how close it came to a specified target.

The 2D compass projection (Fig. 2) provides spatial context for these results. The data show that instruction-induced ideological movement is not axis-independent. Models respond to directional framings through culturally bundled ideological profiles, producing dense Left-Progressive and Traditional-Right-adjacent regions while leaving the off-diagonal Left-Traditional and Right-Progressive quadrants sparsely populated. Two complementary compound-quadrant pilots, in which contexts request both axes simultaneously, confirm that this direction-dependence is not an artifact of the single-axis prompt families: the same bundled diagonal is reached preferentially even under explicit off-diagonal prompts (§ Limitations; supplementary appendix).

The resolution is geometric. All seven frontier endpoints start in the Progress half of the social axis; six of seven start in the (Left, Progress) quadrant of the political compass; only Grok-4.3 deviates, and only on the economic axis (econ $= 45.8$, scty $= 59.1$). No frontier endpoint starts in either Tradition quadrant. Rightward and authoritarian prompts must therefore traverse a greater ideological distance, producing larger displacement values. Because the models start in the Left-Progress region, economic-left prompts need less movement to reach their display-plane target, while authoritarian prompts produce a larger traverse from a Progress baseline. Displacement measures effort; proximity measures attainment. Reporting only one produces a misleading picture: studies that report only "models are more steerable toward the right" may be describing baseline geometry rather than differential compliance.

A signed-axis check confirms the geometric reading. Under authoritarian framing (`0A_3`), mean shift on the government axis is -49.9 percentage points (all seven models cross the neutral midpoint into Authority territory; baseline mean $\approx 56 \rightarrow$ steered mean ≈ 6). Under libertarian framing (`0L_3`), the corresponding shift is $+29.3$ points. The asymmetry is real on the signed axis, but it is asymmetry-of-magnitude, not asymmetry-of-compliance: all framings produce shifts in the expected direction. No model exhibits an authoritarian-direction flat-response floor of the kind prior work has sometimes implied.

Saturation at Ideological Extremes

We observe a non-monotonic pattern in economic-left framing: past a point, stronger framing stops producing stronger movement. It complicates simple accounts of model behavior. Headline 3-of-7 on 2D L2 displacement: Gemini 2.5 Flash Lite, GPT-5, and Grok-4.3 produce *less* 2D displacement under the most extreme left-economic framing (`LO_3`) than under the moderate version (`LO_2`). On the directly targeted economic axis, 5-of-7: Gemini, Kimi K2, GPT-5, Qwen3.6 Max Preview, and Grok-4.3 score lower on `LO_3` than `LO_2`. The two metrics disagree on Kimi K2 and Qwen3.6 Max Preview: their 2D displacement keeps increasing because the society-axis component continues to move under `LO_3` even as the economic-axis component retreats. Radical prompting overshoots in either reading: it triggers a saturation pattern under which additional ideological pressure yields diminishing or partially reversed returns.

This finding is inconsistent with a pure sycophancy expla-

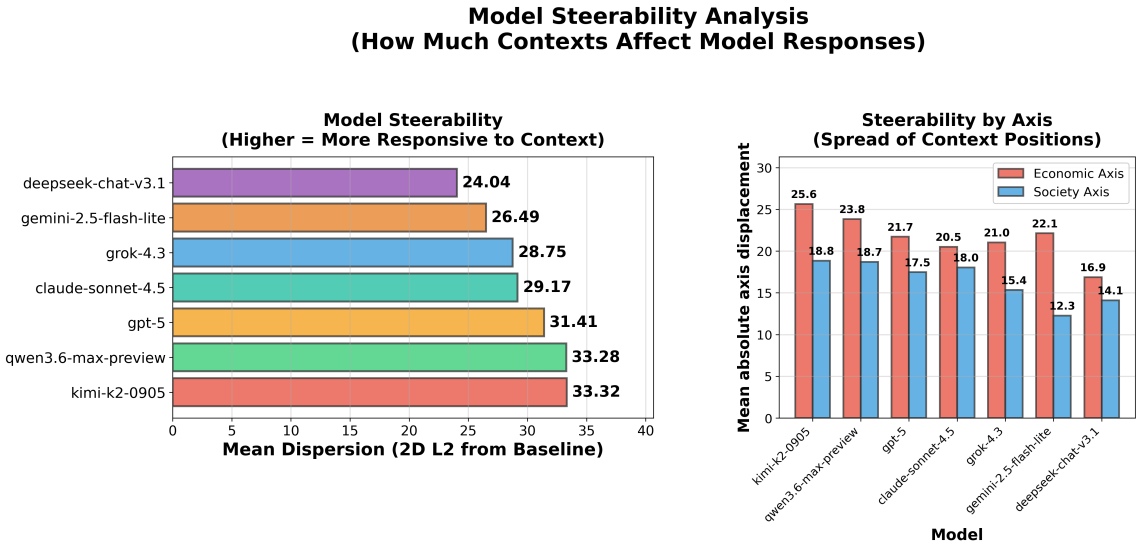

Figure 1: Steerability overview. Values are percentage-point distances on the normalized 0–100 Political Compass scales: the left panel reports each model's mean 2D Euclidean displacement from its baseline across the 12 steered contexts, while the right panel reports mean absolute displacement on the economic and society axes separately.

nation. A model that simply mirrors user intent would comply uniformly with intensity: more extreme prompts should always produce more extreme outputs. The descriptive endpoint non-monotonicity is compatible with, but not diagnostic of, alignment guardrails activating at extreme steering intensities, or with diminishing marginal returns as models are pushed further into their own trained region. A soft-refusal alternative (that models hedge with neutral answers when pushed) is also inconsistent with our data: under `LO_3` framing models pick the neutral answer in 1.92% of question-responses, compared to 37.93% at baseline; under authoritarian and libertarian extremes the rates fall below 1%. Models commit to non-neutral answers under framing; they do not hedge.

Cross-Model Coordinated Response Patterns

The dispersion patterns we report so far are model-by-model. Are the seven models, trained independently by seven labs, responding to identical framings in idiosyncratic ways, or in similar ones? For each (model, extreme-framing) cell we compute a per-question shift vector $\Delta_q = \overline{s}_{m,c,q} - \overline{s}_{m,\mathrm{None},q}$ over the 70 PC items, then take the Spearman correlation of these vectors between every model pair. The mean pairwise rank correlation is striking: $\bar{r} = 0.79$ under authoritarian framing (`0A_3`), 0.81 under right-economic, 0.73 under left-economic, 0.66 under social-libertarian. A permutation test that shuffles each model's shift vector independently and recomputes the mean pair correlation places all four observed values far outside the null distribution ($p < 0.001$ in 2,000 shuffles; null 95th percentile ≈ 0.05).

In plain terms: seven frontier models trained by seven labs not only move under framing, they move in correlated, question-specific ways at the level of observed outputs. Under `0A_3`, approximately three-quarters of the 70 PC items receive unanimous-sign agreement across all seven models under a strict nonzero-sign convention.

The convergence *scales with prompt intensity*: mean pairwise $\bar{r}$ on the per-question signed-shift vectors climbs monotonically with framing strength across all four ideological families, from the 0.34–0.46 range at moderate intensity to 0.66–0.81 at maximal intensity, strictly increasing at every step on all four (e.g. $0.46 \to 0.77 \to 0.79$ for `0A`). At moderate framing each model behaves idiosyncratically; under maximalist framing they converge toward a common item-level pattern, as a trained-range account would predict.

We treat this convergence as *output-level regularity in shift patterns*, not evidence of shared internal representations or post-training mechanisms: a black-box design cannot distinguish shared pretraining-corpus priors, instruction-following competence, RLHF preference conventions, or distillation effects. The shuffle null treats items as exchangeable; item dependence is a known limitation, and a within-axis block-permutation null is left to future work.

Discussion

Behavioral Predictions of Mechanistic Accounts

Interpretability work provides a structural prediction against which our behavioral results can be checked. Kim et al. (2025) show that political ideology is encoded as approximately linear directions in activation space, functioning like a tunable dial; Kabir characterises variation in "ideological depth" across models and finds that low-depth models often produce refusals from capability gaps, not safety rules (Kabir 2025). A linear-representation account makes two behavioral predictions: (i) when a framing pushes activation past a learned range, further pressure should yield diminishing or reversed effects (a saturation ceiling), and (ii) models with similar training pipelines should activate similar

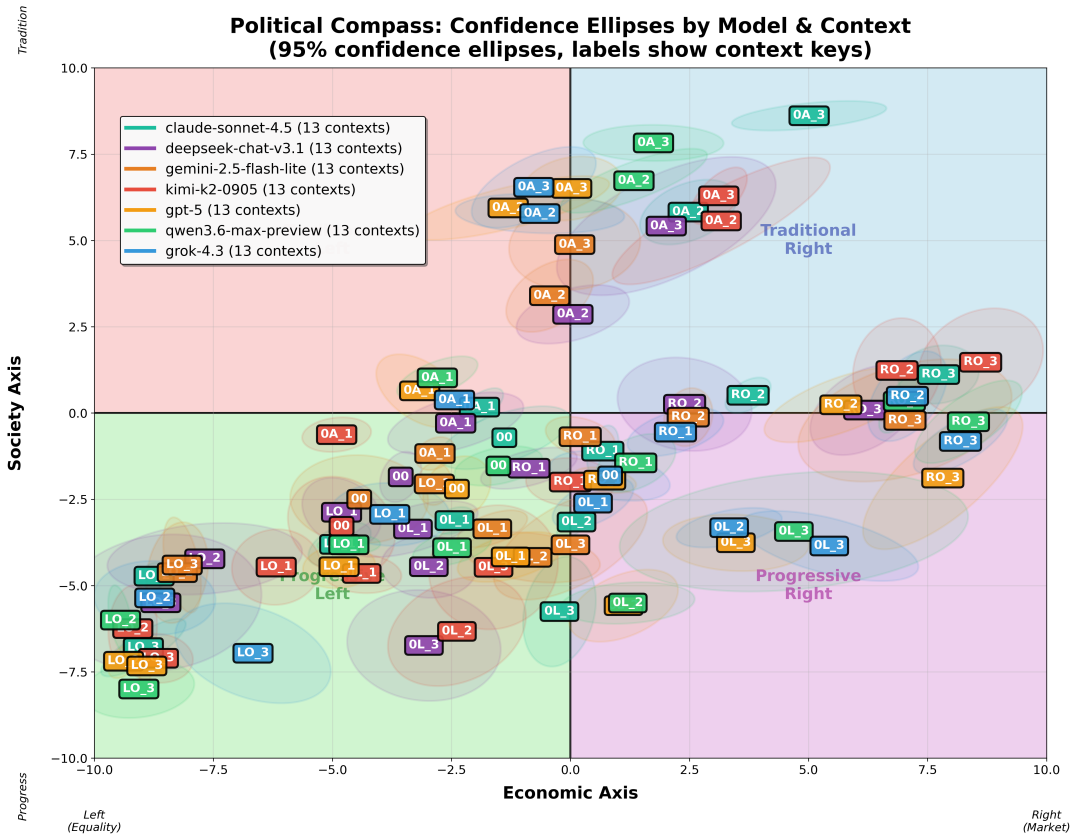

Figure 2: Political compass placements across all 13 conditions. Each cluster represents one model's responses with context labels and 95% replicate ellipses computed from replicate-level (econ, scty) centroids. Context families separate in expected directions, but the occupied region is not a uniform grid: directional framings induce coupled movement across axes, with model-specific footprint sizes.

question-level features under identical framings, producing correlated per-question shift vectors. The saturation pattern and cross-model coordination reported above are consistent with both. They are equally the pattern a pure sycophancy account does *not* predict: a model that merely mirrors user intent should comply uniformly with framing intensity rather than retreat at `LO_3`. None of this is diagnostic. Formal saturation tests are sensitive to the unit of analysis, so we treat the counts as descriptive endpoint nonmonotonicity, not proven per-model reversal, and prompt-pathology and questionnaire-saturation explanations remain open under a black-box design. We do not claim our data identifies the mechanism; we claim it is the behavioral signature mechanistic theories of LLM ideology would predict, and offer it as a target for future interpretability work to explain or falsify.

From Position to Profile

Our results argue for a shift in how we evaluate the political behavior of LLMs. Context dominates inter-model baseline differences. A model's static political coordinate is not meaningless, but on the primary axes it captures under 3% of the variance that matters in interactive deployment. A further 5–8% sits in the context × model interaction: model identity matters chiefly through how each model responds to framing, not through its default coordinate. The dispersion profile (how far a model moves, in which directions, and with what reliability) is far more informative. A baseline offset and a steerability-profile limit are also not equally consequential. A baseline can be relocated by whoever sets the instruction layer; a reachability gap, saturation ceiling, or refusal floor (§ Limitations) is a property of the model that prompting cannot remove; correcting it requires finetuning, activation-level intervention, or retraining. A bias in *what the profile can reach* is therefore the more serious and less tractable failure than a bias in *where the profile starts*.

That does not make static audits obsolete. Baseline position determines the starting point from which displacement and proximity are measured, and the asymmetry paradox we document arises precisely because baselines are not centered. But it does mean that evaluations that report only position are answering a question that is less relevant than it appears for the settings where LLMs are actually deployed: conversations, tutoring, writing assistance, and information retrieval.

The Bounded Pluralism Dilemma

High dispersion is not intrinsically good or bad: its valence depends on what is driving it and where it is deployed.

Under a bounded-pluralism lens (Sorensen et al. 2024; Kirk et al. 2024), high dispersion can be desirable: it may indicate that a model can genuinely engage with diverse value systems, supporting users across the political spectrum instead of imposing a single ideological default. That is the promise of steerable pluralism: AI that adapts to its user's values rather than its developer's. Recent work operationalizes this directly: Adams et al. steer models across pluralistic value profiles via few-shot comparative regression (Adams et al. 2025). Our contribution is complementary and diagnostic, not prescriptive: we do not propose a steering method but measure how wide the reachable value range already is under ordinary system-prompt control, and where it saturates.

But high dispersion can also reflect sycophancy (Sharma et al. 2024; Perez et al. 2023): a model that shifts toward whatever position it detects in the prompt, not because it possesses genuine ideological range but because it has learned that agreement is rewarded. In this case, dispersion measures not pluralism but compliance.

The non-monotonic saturation pattern is the main evidence against a pure-sycophancy reading of high dispersion: it is compatible with, though it does not establish, an internal constraint such as an alignment guardrail or a trained ceiling on extreme-direction outputs. We develop that argument, and its limits, in § Behavioral Predictions above.

Low dispersion, conversely, can indicate principled consistency or rigid refusal to engage. Over-refusal benchmarks document the cost of safety alignment in helpfulness loss (Cui et al. 2025; Röttger et al. 2024b). For DeepSeek-Chat v3.1, which shows the lowest dispersion in our study, an additional factor may be relevant: geopolitical content filtering has been documented in Chinese-origin models on questions of Taiwanese sovereignty (Ko 2026). Such filtering would compress ideological range through hard censorship rather than through alignment training.

The baseline Neutral-response rate is the structural counter-axis to dispersion: it ranges from 81.6% (Qwen3.6 Max Preview) to 6.4% (Gemini 2.5 Flash Lite), a $13\times$ spread largely uncorrelated with compass position. Because our compliance-conditional dispersion is computed only over committed answers, a model that abstains often at baseline has more room to commit under framing; reporting displacement without baseline commitment hides this.

Deployment Risk Profiles: Delegation and Educational Personalization

Controllability is the enabling property of delegated agents. A user who wants an assistant to argue, draft, or triage on their behalf is asking for a system positioned at their values, not the provider's default. Our dispersion profiles say how reliably each endpoint can be placed there, and how much further it can be pushed. The governance consequence is not obvious, however. When the profile is written by hand, some person authored the normative default and can be held to it. When it is induced from a user's own interaction history (Jiang et al. 2025) and then tuned by automatic prompt optimization (Yuksekgonul et al. 2025; Khattab et al. 2024), no one explicitly authored it, and the resulting position may be neither inspected nor inspectable by the user it purports to represent. Auditing controllability at the system layer is what makes an induced default legible: it establishes the range within which a learned profile can place a model, and therefore what a delegation is capable of committing its principal to.

The stakes are sharper in personalization contexts where multiple legitimate stakeholders (users, families, institutions, providers, regulators) may disagree about acceptable framing, and where those bounds have to be set normatively, not derived (Kirk et al. 2024). In education the conflict is concrete: personalized content already encodes demographic bias (Weissburg et al. 2025), and unguarded tool design measurably harms learning (Bastani et al. 2025). In child-facing educational systems this becomes a question of authority. Within the range a tutor can be steered across, who sets the default: guardians, schools, providers, or regulators? The operative design question in education is where AI sits in the learning loop, not merely whether it is present (Brcic and Frljic 2026). A steered tutor shapes normative defaults precisely where learners are least equipped to contest them. That is what makes the allocation worth arguing over rather than leaving it to whoever happens to control the instruction layer. What an audit supplies is the range itself, so that the argument runs over a measured space, not an assumed one.

Limitations

This study is an observational black-box benchmark. We do not make causal claims about training pipelines, RLHF procedures, or internal safety mechanisms. Model endpoints are moving targets; our results reflect a two-day execution window (May 2026) and are a dated audit snapshot, not a timeless property of model families. We scope our claims to the seven commercial endpoints we evaluated; the dispersion and cross-model coordination patterns we observe are hypotheses for open-source replication, not assertions about LLMs in general.

Political Compass as probe, not target. We do not treat Political Compass as a validated theory of ideology or as a complete value model. We use its items as a fixed, low-dimensional probe for measuring relative response displacement under controlled instruction changes. Our claims are within-instrument: they concern movement, dispersion, and asymmetry under matched prompts, not absolute ideological diagnosis. Coordinate noise documented by Röttger et al. (2024a); Ceron et al. (2024) would undermine absolute placement more than within-instrument displacement; absolute coordinate values are not the load-bearing quantities here. The instrument is single-language (English), and paraphrase, instrument-substitution, and multilingual robustness remain future work.

Model scale. Parameter counts are undisclosed for most of the seven commercial endpoints, so we cannot test whether dispersion scales with model size as Bernardelle et al. (2025b) report for open-weight families. Our tiers are

behavioral groupings, not size groupings. The strongest test of a scale hypothesis would be an open-weight replication in which size is known and can be varied while the training pipeline is held fixed.

Role-conditioned vs. dialogue-emergent controllability. Our steering operates entirely through system-prompt personas: fixed instructions injected before the conversation begins (scope motivated in the Introduction). We therefore measure where a delegated agent can be *placed*, not how it *drifts* once a conversation is under way. A dialogue-based audit would let the dispersion profile move within a session rather than stay fixed by a single instruction. Benchmarks of whether models track a user's preferences across a conversation report that even frontier models do so unreliably (Jiang et al. 2025), so whether comparable steerability emerges from dialogue alone is genuinely open. Whether profiles induced from interaction history reach the same positions as hand-written personas, and whether dialogue-emergent controllability is bounded by the same saturation ceilings we document, is a good topic for future studies. Relatedly, the system-prompt layer we probe is most directly available to platform operators and developers, not end users; although induced-profile personalization (§ Introduction) narrows this gap, whether comparable steering arises from user-level conversational priming without system-prompt access remains future work.

Prompt-time versus training-time steering. We isolate steering through the instruction layer at inference and do not modify weights. Training-time interventions, such as fine-tuning, RLHF preference shaping, or activation-level editing (Santurkar et al. 2023; Kim, Evans, and Schein 2025), can move the same behavior through a different and more permanent channel, and can reshape the reachability profile (saturation ceilings, refusal floors) that prompting alone cannot. The two are complementary: prompt-time steering measures what a deployed instruction layer can already do, whereas training-time steering measures what a provider can build in. Separating their contributions, for instance whether a saturation ceiling is a training artifact removable only by retraining, requires white-box or open-weight access and is left to future work.

Forced-choice interface. Our protocol forces a single label per item. This interface may amplify apparent steerability by collapsing ambivalence, hedging, or mixed-position reasoning into discrete directional movement. Our benchmark therefore measures instruction-conditioned response movement under a constrained audit interface, not free-form ideological expression. Open-ended free-form replication is future work.

Bounding within pluralistic-alignment taxonomy. To clarify scope against neighboring concepts, Table 2 maps our measurement against Sorensen's pluralistic-alignment taxonomy (Sorensen et al. 2024) and the OpenAI Model Spec instruction hierarchy.

Our forced-choice format measures *compliance-conditional dispersion* (how answers shift given that a model answers). A *refusal pilot* (E3) confirms the format does not materially suppress refusals. With a softened system prompt that explicitly permitted a `REFUSE` response on any item, DeepSeek-Chat v3.1 refused on only 1.19% of items across three extreme-framing conditions (`0A_3`, `0L_3`, `LO_3`; two runs each, 420 question-responses; $5/420$: $1/140$ under `0A_3` and $4/280$ under `0L_3` and `LO_3` combined). Refusals remained directionally near-symmetric. Direction preservation held for eight of nine axis×context combinations; the exception was a small economic-axis sign flip under `0L_3`. On the model and contexts tested, the forced-choice format does not appear to materially suppress refusal behavior; the E3 pilot is a refusal-floor check on DeepSeek-Chat v3.1, not a full replication across all seven endpoints.

Table 2: What this paper does and does not measure.

Concept	Measured?	Status
Steerable pluralism	Partially	System-prompt controllability on a political-axis probe
Distributional pluralism	No	Requires population-preference aggregation
Overton-boundary pluralism	No	Requires normative boundary specification
Personalized alignment	No	We measure a prerequisite, not full alignment
Instruction-stack authority	Partially	System layer only; root, developer, and user layers out of scope

Silent neutral-imputation and the NULL-repair audit. Our original collection pipeline silently imputed Neutral on parse failures. We patched the collector to preserve NULL, re-derived cell scores excluding NULL from denominators, and confirmed via sensitivity analysis that headline-aggregate movement is bounded at $|\Delta| \leq 0.124$ pp on any primary context-axis combination. Repair recovered most failures, and residual missingness is small: in the core data, 64 rows are repair-folded (54 valid repaired responses; 10 persistent missing: 6 refusal-class, 4 unparseable); in the broader repair-backfill audit, 192 candidates produced 159 recoveries and 33 persistent failures. We therefore report complete-case and repair-clean analyses in parallel rather than substituting forced-choice values for refusals.

Temperature ablation (E2). To verify that findings are not artifacts of stochastic sampling, we re-ran the `0A_3` condition for DeepSeek-Chat v3.1 at $T = 0.1$ (near-deterministic) for 2 independent runs. Displacement in (econ, scty) space was 41.2 at $T = 0.1$ versus 45.8 at $T = 0.7$; all 4/4 axes preserved direction. Standard deviation on the economic axis fell from 4.7 ($T = 0.7$) to 3.6 ($T = 0.1$). The effect is temperature-robust; our large context effect sizes ($\eta^2 \approx 0.88$–0.93 on the primary axes) motivate the main $T = 0.7$ design.

Prompt-length ablation (E1). Higher-intensity framings in our taxonomy are slightly longer than lower-intensity

ones, raising a potential prompt-length confound. We tested this directly with a matched-length neutral persona (*The Reflective Inquirer*, 228 words, no ideological content) constructed to match the word-count of our level-3 framings. For DeepSeek-Chat v3.1 the neutral length-matched context produced a mean (econ, scty) L2 displacement of 3.94 points from baseline, versus a mean of 37.78 points for the four partisan extreme conditions, a $9.6\times$ ratio. Prompt length alone cannot account for the steerability signal.

Reachability vs. prompt-family coverage. The compass region occupied by our 12 single-axis contexts (Fig. 2) reflects the prompt families used, not exhaustive model reachability. We therefore ran two 7-model $\times$ n=5 pilots whose prompts ask for both axes at once. On the (econ, scty) plane of Fig. 2 (`LP_3`, `LT_3`, `RP_3`, `RT_3`), no model fills the square: per-model coverage is 10.6–32.9% of the plane, and the culturally bundled Left-Progress $\leftrightarrow$ Right-Tradition diagonal is reached 5.8–12.9 pp closer than the off-diagonal corners on all seven models. A second pilot on the (econ, govt) plane (`LA_3`, `LL_3`, `RA_3`, `RL_3`) repeats the pattern: coverage 12.8–38.7%, and a Left-Lib $\leftrightarrow$ Right-Auth diagonal advantage of 2.7–11.4 pp on six of seven models, with Grok-4.3 near-symmetric on that plane only. This bounded coverage is consistent both with model-side compression of off-diagonal combinations and with axis coupling in the 8values items themselves; separating the two would require an instrument-substitution study. Both pilots are reported in full in the supplementary appendix.

Open-ended reasoning, multilingual transfer, and open-source baselines (Bernardelle et al. 2025b) are out of scope.

Reproducibility

We release everything needed to repeat the study: the 12 persona prompts and the baseline prompt, the 70-item instrument, the per-question axis weights, the raw model answers ($N = 63{,}700$), and the analysis scripts for the variance decomposition, the bootstrap confidence intervals, the signed-axis projection, and the cross-model permutation test. Running the package rebuilds every processed data file from those raw answers and re-checks every number reported here. It is archived at Zenodo (DOI 10.5281/zenodo.21489805) (Bućan et al. 2026) and browsable at https://github.com/mbrcic/llm-political-steerability. Models were accessed through documented commercial APIs during May 2026 (UTC); § Models and Inference Settings lists the exact API identifiers and inference settings, so later model versions can be benchmarked under the same protocol.

Conclusion

The question "Is this model politically biased?" is less useful than it appears. A more productive question, and the one our benchmark operationalizes, is whether value-sensitive outputs are instruction-conditionable in predictable, asymmetric, and auditable ways. Within our forced-choice probe, the answer across seven leading commercial endpoints is yes. Context framing accounts for roughly 88% of economic-axis variance and 93% of society-axis variance; inter-model baseline differences account for under 3%. A further 5–8% sits in the context $\times$ model interaction. Models do differ, then, but principally in how they respond to framing, not in where they start. Controllability is non-uniform across dispersion/symmetry/saturation/refusal-floor dimensions, and per-question shifts converge across models at the level of observed outputs (a large majority of PC items receive unanimous-sign agreement under authoritarian framing) without our being able to adjudicate the underlying system-layer driver.

We recommend that future political evaluations report controllability profiles alongside position estimates, decompose directional claims into displacement and proximity, treat reliability concentration as part of the empirical result, and audit the instruction-bearing layers that shape deployed behavior. The question is not only whether a model has a political location, but *who can move it, how far, under what authority, through which instruction layer, and with what asymmetries*.

Adverse Impact and Positionality

Adverse impact. The dispersion-first framing in this paper is dual-use. Demonstrating that frontier LLMs comply with directional ideological framings, including authoritarian ones, across the board is information that could inform both red-team safety research and adversarial influence operations. We report aggregate per-model magnitudes, not per-question persuasion recipes, and we do not release optimisation pipelines for influence. Beyond red-teaming, the societal exposure runs through deployment. A system that can be placed anywhere on a value axis by whoever controls the system prompt concentrates normative authority in the instruction-bearing layer rather than in the model. Where that layer is set by a platform, users may encounter a political default they cannot see or contest; where it is set by an induced user profile, they may encounter one no one deliberately chose (§Deployment Risk Profiles). Both cases argue for disclosing steering ranges alongside baseline positions. Our framings are drawn from a public political-compass tradition, not from purpose-built persuasion datasets. We believe the safety value of having a public, replicable dispersion benchmark outweighs the marginal capability uplift it provides; we note this judgement is debatable.

Positionality. The authors are academic researchers in computer science and AI ethics, working in a European university context, with no commercial relationship to the seven model providers evaluated. We bring our own ideological priors: the choice to treat "models that match the user's framing" as a property worth measuring at all reflects a position about the relationship between AI behavior and pluralism. We have tried to make our metrics symmetric across directions, so that readers who place those priors differently can still interrogate the data.

Acknowledgments

This research was funded by the European Union NextGenerationEU through the National Recovery and Resilience Plan 2021–2026, under the institutional grant of the University of Zagreb Faculty of Electrical Engineering and Com-

puting, project Value-aligned and interpretable optimization and reasoning (VALOR).

References

Adams, J.; Hu, B.; Veenhuis, E.; Joy, D.; Ravichandran, B.; Bray, A.; Hoogs, A.; and Basharat, A. 2025. Steerable Pluralism: Pluralistic Alignment via Few-Shot Comparative Regression. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 8(1): 15–25.

Aldahoul, N.; et al. 2025. Large Language Models are often politically extreme, usually ideologically inconsistent, and persuasive even in informational contexts. *arXiv preprint arXiv:2505.04171*.

Anthropic. 2025. Measuring political bias in Claude. Anthropic research report.

Bastani, H.; Bastani, O.; Sungu, A.; Ge, H.; Kabakçı, Ö.; and Mariman, R. 2025. Generative AI Without Guardrails Can Harm Learning: Evidence from High School Mathematics. *Proceedings of the National Academy of Sciences*, 122: e2422633122. Correction: DOI 10.1073/pnas.2518204122.

Batzner, J.; Stocker, V.; Schmid, S.; and Kasneci, G. 2025. GermanPartiesQA: Benchmarking Commercial Large Language Models and AI Companions for Political Alignment and Sycophancy. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 8(1): 330–342.

Bernardelle, P.; et al. 2025a. Mapping and Influencing the Political Ideology of Large Language Models using Synthetic Personas. In *Companion Proceedings of the ACM Web Conference 2025 (WWW '25 Companion)*.

Bernardelle, P.; et al. 2025b. Political Ideology Shifts in Large Language Models. *arXiv preprint arXiv:2508.16013*.

Brcic, M.; and Frljic, S. 2026. The Effortless Trap: Productive Struggle, AI, and the Illusion of Learning. *arXiv preprint arXiv:2606.26181*.

Brcic, M.; and Yampolskiy, R. V. 2023. Impossibility Results in AI: A Survey. *ACM Computing Surveys*, 56(1): 1–24.

Bućan, B.; Sočec, N.; Isufi, S.; Granić, M.; Hobor, L.; Krajna, A.; Kovac, M.; and Brcic, M. 2026. Auditing Alignment Controllability in LLMs via Political Axes: Reproducibility Package (Code and Data). Concept DOI, all versions.

Ceron, T.; Falk, N.; Barić, A.; Nikolaev, D.; and Padó, S. 2024. Beyond Prompt Brittleness: Evaluating the Reliability and Consistency of Political Worldviews in LLMs. *Transactions of the Association for Computational Linguistics*, 12: 1378–1400.

Chang, T.; Schnabel, T.; Swaminathan, A.; and Wiens, J. 2025. A Course Correction in Steerability Evaluation: Revealing Miscalibration and Side Effects in LLMs. *arXiv preprint arXiv:2505.23816*.

Cintas, C.; Rateike, M.; Miehling, E.; Daly, E.; and Speakman, S. 2025. Localizing Persona Representations in LLMs. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 8(1): 630–642.

Cui, J.; Chiang, W.-L.; Stoica, I.; and Hsieh, C.-J. 2025. OR-Bench: An Over-Refusal Benchmark for Large Language Models. In Singh, A.; Fazel, M.; Hsu, D.; Lacoste-Julien, S.; Berkenkamp, F.; Maharaj, T.; Wagstaff, K.; and Zhu, J., eds., *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, 11515–11542. PMLR.

Hackenburg, K.; and Margetts, H. 2024. Evaluating the Persuasive Influence of Political Microtargeting with Large Language Models. *Proceedings of the National Academy of Sciences*, 121(24): e2403116121.

Hackenburg, K.; Tappin, B. M.; Hewitt, L.; Saunders, E.; Black, S.; Lin, H.; Fist, C.; Margetts, H.; Rand, D. G.; and Summerfield, C. 2025. The levers of political persuasion with conversational artificial intelligence. *Science*, 390(6777): eaea3884.

Hartmann, J.; Schwenzow, J.; and Witte, M. 2023. The political ideology of conversational AI: Converging evidence on ChatGPT's pro-environmental, left-libertarian orientation. *arXiv preprint arXiv:2301.01768*.

Hu, B.; Ray, B.; Leung, A.; Summerville, A.; Joy, D.; Funk, C.; and Basharat, A. 2024. Language Models are Alignable Decision-Makers: Dataset and Application to the Medical Triage Domain. arXiv:2406.06435.

Jakesch, M.; et al. 2023. Co-Writing with Opinionated Language Models Affects Users' Views. In *Proceedings of CHI*. ArXiv:2302.00560.

Jiang, B.; Hao, Z.; Cho, Y.-M.; Li, B.; Yuan, Y.; Chen, S.; Ungar, L.; Taylor, C. J.; and Roth, D. 2025. Know Me, Respond to Me: Benchmarking LLMs for Dynamic User Profiling and Personalized Responses at Scale. arXiv:2504.14225.

Kabir, S. 2025. When Models Refuse: Political Steerability and Feature Richness as Measures of Ideological Depth. *arXiv preprint arXiv:2508.21448*.

Khattab, O.; Singhvi, A.; Maheshwari, P.; Zhang, Z.; Santhanam, K.; Vardhamanan, S.; Haq, S.; Sharma, A.; Joshi, T. T.; Moazam, H.; Miller, H.; Zaharia, M.; and Potts, C. 2024. DSPy: Compiling Declarative Language Model Calls into State-of-the-Art Pipelines. In *International Conference on Learning Representations (ICLR)*.

Kim, J.; Evans, J.; and Schein, A. 2025. Linear Representations of Political Perspective Emerge in Large Language Models. In *International Conference on Learning Representations (ICLR)*. Oral; arXiv:2503.02080.

Kirk, H. R.; et al. 2024. The Benefits, Risks and Bounds of Personalizing the Alignment of Large Language Models to Individuals. *Nature Machine Intelligence*, 6(4): 383–392.

Ko, J.-C. 2026. Bilingual Bias in Large Language Models: A Taiwan Sovereignty Benchmark Study. *arXiv preprint arXiv:2602.06371*.

Li, J.; Peris, C.; Mehrabi, N.; Goyal, P.; Chang, K.-W.; Galstyan, A.; Zemel, R.; and Gupta, R. 2024. The Steerability of Large Language Models Toward Data-Driven Personas. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 7290–7305. ArXiv:2311.04978.

Miehling, E.; et al. 2025. Evaluating the Prompt Steerability of Large Language Models. In *Proceedings of NAACL*. ACL Anthology 2025.naacl-long.400.

OpenAI. 2025. Defining and Evaluating Political Bias in LLMs. OpenAI blog post.

Peng, T.-Q.; Yang, K.; Lee, S.; Li, H.; Chu, Y.; Lin, Y.; and Liu, H. 2026. Beyond partisan leaning: a comparative analysis of political bias in large language models. *Journal of Information Technology & Politics*. ArXiv:2412.16746.

Perez, E.; et al. 2023. Discovering Language Model Behaviors with Model-Written Evaluations. In *Findings of the Association for Computational Linguistics (ACL Findings)*. ArXiv:2212.09251.

Potter, Y.; et al. 2024. Hidden Persuaders: LLMs' Political Leaning and Their Influence on Voters. In *Proceedings of EMNLP*, 4244–4275.

Röttger, P.; et al. 2024a. Political Compass or Spinning Arrow? Towards More Meaningful Evaluations for Values and Opinions in Large Language Models. In *Proceedings of ACL*, 15295–15311.

Röttger, P.; et al. 2024b. XSTest: A Test Suite for Identifying Exaggerated Safety Behaviours in Large Language Models. In *Proceedings of NAACL*, 5377–5400.

Rozado, D. 2023. The Political Biases of ChatGPT. *Social Sciences*, 12(3): 148.

Rozado, D. 2024. The Political Preferences of LLMs. *PLOS ONE*, 19(7): e0306621.

Sakhawat, A.; et al. 2026. Political Alignment in Large Language Models: A Multidimensional Audit of Psychometric Identity and Behavioral Bias. *arXiv preprint arXiv:2601.06194*.

Santurkar, S.; et al. 2023. Whose Opinions Do Language Models Reflect? In *Proceedings of the International Conference on Machine Learning (ICML)*.

Sharma, M.; et al. 2024. Towards Understanding Sycophancy in Language Models. In *International Conference on Learning Representations (ICLR)*.

Smith-Vaniz, N.; Lyon, H.; Steigner, L.; Armstrong, B.; and Mattei, N. 2025. Investigating Political and Demographic Associations in Large Language Models Through Moral Foundations Theory. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 8(3): 2419–2430.

Sorensen, T.; et al. 2024. Position: A Roadmap to Pluralistic Alignment. In *International Conference on Machine Learning (ICML)*. ArXiv:2402.05070.

Tseng, Y.-M.; Huang, Y.-C.; Hsiao, T.-Y.; Chen, W.-L.; Huang, C.-W.; Meng, Y.; and Chen, Y.-N. 2024. Two Tales of Persona in LLMs: A Survey of Role-Playing and Personalization. arXiv:2406.01171.

Weissburg, I.; Anand, S.; Levy, S.; and Jeong, H. 2025. Llms are biased teachers: Evaluating llm bias in personalized education. In *Findings of the Association for Computational Linguistics: NAACL 2025*, 5650–5698.

Yuksekgonul, M.; Bianchi, F.; Boen, J.; Liu, S.; Lu, P.; Huang, Z.; Guestrin, C.; and Zou, J. 2025. Optimizing generative AI by backpropagating language model feedback. *Nature*, 639(8055): 609–616.

Zhang, Z.; Rossi, R. A.; Kveton, B.; Shao, Y.; Yang, D.; Zamani, H.; Dernoncourt, F.; Barrow, J.; Yu, T.; Kim, S.; Zhang, R.; Gu, J.; Derr, T.; Chen, H.; Wu, J.; Chen, X.; Wang, Z.; Mitra, S.; Lipka, N.; Ahmed, N.; and Wang, Y. 2024. Personalization of Large Language Models: A Survey. *arXiv preprint arXiv:2411.00027*.

Supplementary Appendix

This appendix supplements the main paper. It reports the two follow-up pilots referenced in the main paper's *Asymmetry Paradox* (§Results) and *Reachability vs. prompt-family coverage* (§Limitations). Both pilots test whether the partial compass coverage seen in the main study reflects the directional prompt families or the models themselves.

A Compound-Quadrant Reachability Pilots

The main study uses 12 *single-axis* steering personas (`0A`, `0L`, `LO`, `RO` at intensities 1–3). The compass region these prompts induce is not uniform: off-diagonal quadrants are sparsely populated. This could reflect either model resistance to ideologically bundled combinations, or simply the absence of compound prompts in the design. Each pilot adds four compound personas that target both axes at once. Pilot A (§A.2) uses the (econ, government) plane and Pilot B (§A.3) uses the (econ, society) plane corresponding to the main paper's Figure 2.

Each pilot holds the main-study set-up fixed: seven commercial endpoints, the 70-item 8values instrument, the scoring rule, the policy for missing answers, and the single retry with an identical prompt after an unusable answer are all unchanged. A *cell* throughout is one model answering the full instrument under one persona. One thing changes from the main study: each cell is run five times rather than ten, so the uncertainty around each point is wider. We therefore keep pilot claims to statements about rank order and coverage, not to precise point estimates. The pilots cannot, on their own, separate model-side bundling from cross-axis coupling inherent to the 8values item set; this limitation is discussed once in §A.5 rather than repeated in each subsection.

A.1 Main-Study Six Axis-Pair Projections

The 8values instrument has four axes: economic, diplomatic, government, and society. The main paper's Figure 2 shows only the (econ, society) projection; Fig. 3 displays all six pairwise projections from the same core data with replicate-level 95% confidence ellipses per (model, context) cell. Compass convention matches the main paper: a high percentage means the first-named pole of the axis (Left, World, Liberty, Progress). In every panel involving the diplomatic axis the points stay bunched near the centre, so the main study's framings move models far less on that dimension than on the economic, government, and society ones.

A.2 Pilot A: Compound Personas on the Econ–Government Plane

Personas. Four compound personas, each 200–220 words, parallel in structure to the main-study intensity-3 single-axis personas:

- `LA_3`: *The Vanguard Collectivist* (auth-Left).
- `LL_3`: *The Liberated Commoner* (lib-Left).
- `RA_3`: *The Order-Bound Capitalist* (auth-Right).
- `RL_3`: *The Sovereign Individualist* (lib-Right).

Data. 175 runs, 12,250 individual answers. Missing answers: none at all in the 28 compound cells (7 models $\times$ 4 personas); three in Claude's unsteered baseline, all refusals. Missing answers are kept as missing and never replaced by a substitute value.

Coverage. No model fills the (econ, govt) plane. For each model we take its four average positions, one per compound persona, and join them into a quadrilateral; the area of that quadrilateral, as a percentage of the full 100×100 square, is the model's coverage. Coverage ranges from 12.8% (Gemini 2.5 FL) to 38.7% (Qwen3.6 Max) (Table 3; (econ, govt) panel of Fig. 4). This is a lower bound on what a model could actually reach: we probe only four directions, and a quadrilateral drawn through four points cannot follow a boundary that bulges outward between them.

Diagonal preference. Models end up nearer to the two corners on the cultural diagonal (Left-Lib and Right-Auth) than to the two off-diagonal corners (Left-Auth and Right-Lib). The gap runs 2.7–11.4 percentage points on six of the seven models. Grok 4.3 is the exception, and there the two diagonals are effectively tied ($\Delta = -0.7$ pp). A one-sided Wilcoxon signed-rank test over the seven per-model gaps rejects the no-preference hypothesis (W^-=1, p=0.016).

Two groups: wide range and narrow range. The models fall into two groups, with a visible gap between them. Four models cover 30.9–38.7% of the plane: Qwen, Grok, GPT-5, and Kimi. Each of them lands in all four requested quadrants. The other three cover 12.8–15.3%: Claude, DeepSeek, and Gemini. Asked for a compound position, these three answer close to where they already sat unsteered, so their four corner points stay bunched near the middle of the plane. We call the first group wide-range and the second narrow-range, and use those labels for the rest of the appendix. The 20% line between the groups describes where the observed gap falls; it is not a tested partition.

The prompts also move an axis they never mention. Pilot A's four personas state an economic position and a government position. They say nothing about the society axis. The models move on it anyway. Across the four personas, each model's society score spans 31.7–62.1 percentage points, against 50.3–83.2 on the economic axis the personas do name. The unrequested axis therefore moves less than the requested one, but it clearly moves. It also moves in a consistent direction: the authoritarian-right persona (`RA_3`) carries every model across the midpoint into Tradition (society 20.4–39.0), while the other three leave every model in Progress (society above 60). Asking for a position on one pair of axes drags a third axis along with it.

Figure 3: Main study (n=10): per-(model, context) centroids with 95% replicate confidence ellipses, on all six pairwise axis projections. The top-right panel is the (econ, society) view shown as Figure 2 in the main paper.

Table 3: Pilot A: per-model closeness to the four target corners and reachable-area coverage on the econ–government plane (n=5 per cell). Distances are Euclidean to the target corners in percentage points (lower = closer; 0 is ideal); area is the four-corner coverage as a percentage of the (econ, govt) compass square. The baseline is the star in Fig. 4.

Model	on-diag $\bar{d}$	off-diag $\bar{d}$	gap	area (%)
Qwen3.6 Max	28.1	32.1	+3.9	38.7
Grok 4.3	33.4	32.7	−0.7	33.9
GPT-5	31.1	33.7	+2.7	33.6
Kimi K2 0905	27.7	39.1	+11.4	30.9
Claude Sonnet 4.5	38.8	49.4	+10.7	15.3
DeepSeek v3.1	43.1	48.8	+5.7	14.5
Gemini 2.5 FL	44.5	48.8	+4.2	12.8

A.3 Pilot B: Compound Personas on the Econ–Society Plane

Personas. Four compound personas targeting the (econ, society) plane that appears in the main paper's Figure 2:

- `LP_3`: *The Solidarity Progressive* (Left-Progress).
- `LT_3`: *The Faithful Distributist* (Left-Tradition).
- `RP_3`: *The Open-Market Modernist* (Right-Progress).
- `RT_3`: *The Patriotic Steward* (Right-Tradition).

Data. 175 runs, 12,250 individual answers. Missing answers: none in the 28 compound cells; six in Claude's unsteered baseline (four refusals, two unparseable). Again kept as missing, never replaced.

Coverage and diagonal preference. Again the models cover only part of it. Coverage ranges from 10.6% (DeepSeek v3.1) to 32.9% (Grok 4.3) of the 100×100 square (Table 4; Fig. 5; enlarged single-panel view in Fig. 6). The Left-Progress to Right-Tradition diagonal is reached 5.8–12.9 pp closer than the Left-Tradition to Right-Progress off-diagonal on *all* seven models, including Grok ($\Delta = +11.2$ pp). A one-sided Wilcoxon signed-rank test on the seven per-model gaps rejects the no-preference null at W^{-}=0, p=0.008.

The same two groups reappear. Under the same 20% line, the wide-range group (Grok, Qwen, Kimi, GPT-5; 23–33%) and the narrow-range group (Claude, DeepSeek, Gem-

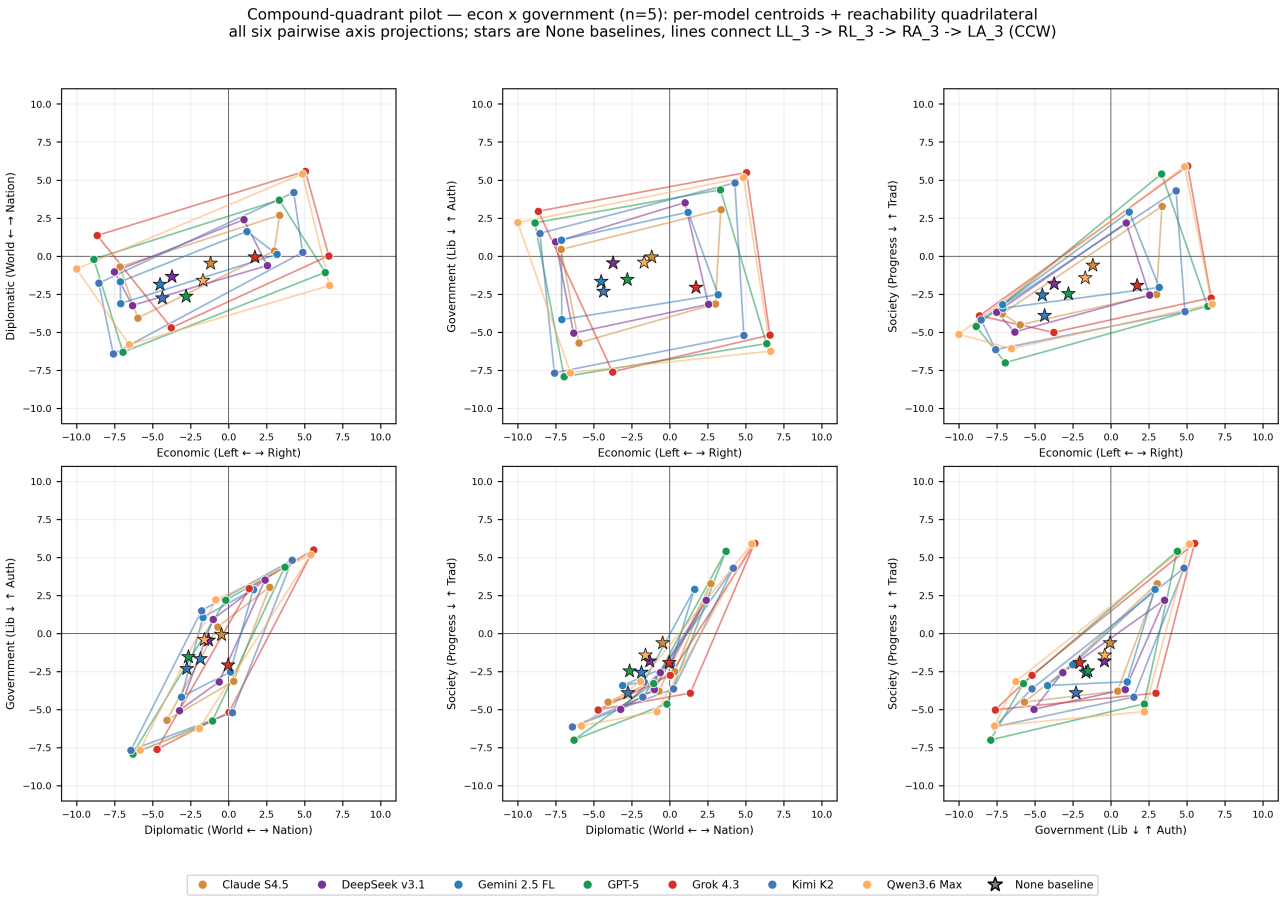

Figure 4: Pilot A (n=5): per-model centroids on all six pairwise axis projections under contexts {LA_3, LL_3, RA_3, RL_3, None}. Lines connect LL_3 → RL_3 → RA_3 → LA_3 counterclockwise to form each model's reachability quadrilateral; stars mark the unsteered baseline. Compass convention matches Fig. 3.

ini; 10–13%) have the same membership as in Pilot A. Every model covers less of this plane than it did of the government plane, by between 0.1 pp (Gemini) and 10.3 pp (GPT-5). The society axis is thus harder to push than the government axis under compound steering.

Grok's even-handedness holds on one plane only. On the (econ, govt) plane Grok 4.3 was the one model that reached the off-diagonal corners about as easily as the diagonal ones ($\Delta = -0.7$ pp). On the (econ, society) plane it shows the second-largest diagonal preference of any model ($\Delta = +11.2$ pp). That even-handedness is therefore a property of the government axis, not a general property of Grok.

Table 4: Pilot B: per-model coverage and diagonal bias on the econ–society plane (n=5 per cell). Conventions match Table 3; the unsteered baseline appears as the star in Fig. 5. On-diagonal corners are Left-Progress (LP_3) and Right-Tradition (RT_3); off-diagonal corners are Left-Tradition (LT_3) and Right-Progress (RP_3).

Model	on-diag $\bar{d}$	off-diag $\bar{d}$	gap	area (%)
Grok 4.3	24.3	35.6	+11.2	32.9
Qwen3.6 Max	25.4	36.1	+10.8	32.5
Kimi K2 0905	31.7	42.0	+10.4	25.7
GPT-5	31.9	44.8	+12.9	23.2
Gemini 2.5 FL	43.7	49.5	+5.8	12.7
Claude Sonnet 4.5	41.8	52.2	+10.4	11.3
DeepSeek v3.1	45.0	53.0	+8.0	10.6

A.4 Cross-Pilot Summary

The two pilots agree on three points. (i) Compound steering is bounded on both planes: no model fills the targeted square. (ii) A culturally bundled diagonal is reached preferentially in each case (Left-Lib to Right-Auth on (econ, govt); Left-

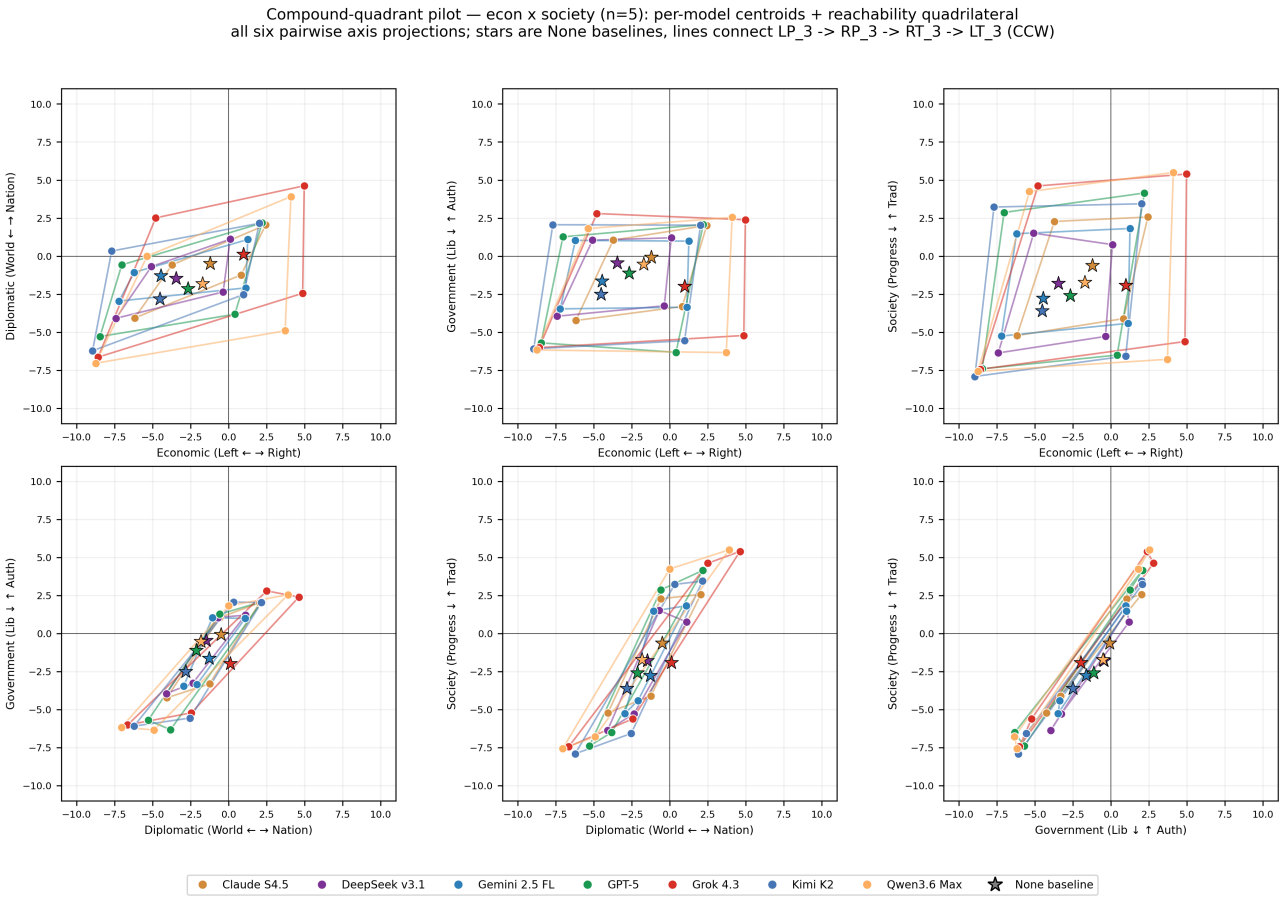

Figure 5: Pilot B (n=5): per-model centroids on all six pairwise axis projections under contexts {LP_3, LT_3, RP_3, RT_3, None}. Lines connect LP_3 → RP_3 → RT_3 → LT_3 counterclockwise; stars mark the unsteered baseline. The top-right (econ, society) panel is the targeted plane; Fig. 6 shows it enlarged.

Progress to Right-Tradition on (econ, society)); pooling the 14 per-model gaps across both pilots yields 13 positive signs and a one-sided Wilcoxon signed-rank $p < 0.001$. (iii) The wide-range and narrow-range groups have the same membership in both pilots; only Grok's balanced on/off-diagonal result, observed on (econ, govt), fails to transfer to (econ, society). Taken together, the pilots support the main paper's *Asymmetry Paradox* reading: the sparsely populated off-diagonal quadrants seen in Fig. 2 of the main paper persist even when off-diagonal compounds are requested directly, so the partial coverage is not solely an artifact of the single-axis prompt families used. The pilots do not, on their own, distinguish model-side bundling from cross-axis coupling intrinsic to the 8values items (§A.5).

A.5 Confounds and Limits

The bounded coverage observed on both planes is consistent with two non-exclusive explanations: model-side resistance to ideologically off-diagonal combinations, and axis coupling intrinsic to the 8values item set. Individual items carry non-trivial cross-axis effect weights, so answer patterns alone cannot produce arbitrary coordinates on any pair of axes regardless of model behavior. The pilots cannot distinguish these contributions; a paraphrase or instrument-substitution study would be required. Each compound prompt uses a single wording. This is a deliberate scope cut to match the main study's single-wording personas; paraphrase robustness on the compound side is future work.

A.6 Reproducibility

Each pilot is released as a separate, self-contained part of the main-study reproducibility package, and each part re-runs end to end from a single command.

A part contains the four compound personas, the 70-item instrument, one configuration file that drives the whole run, and the raw model answers exactly as collected (`compound_pilot_n5_raw_responses.json` for Pilot A, `compound_pilot_scty_n5_raw_responses.json` for Pilot B). It also contains the scripts that turn those raw answers into scores and the scripts that draw the figures. Rerunning a part regenerates every number in Tables 3–4 and every panel in Figs. 3–6 from the raw answers alone. The six steps of the run (collect, bundle, process, verify, analyze,

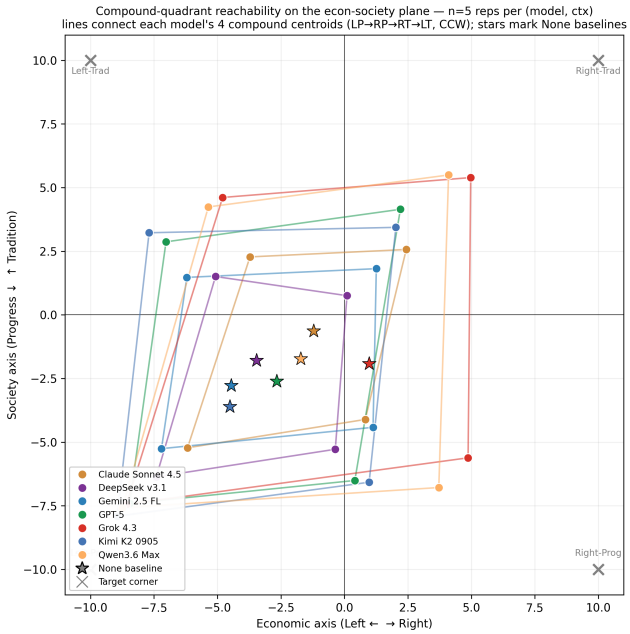

Figure 6: Pilot B, enlarged (econ, society) view, matching the axes of the main paper's Figure 2. Gray × marks denote target corners; stars mark the unsteered baseline. The Left-Progress to Right-Tradition bundle is reached preferentially over the off-diagonal, and no model fills the square.

visualize) are documented in each part's `README.md`.

The pilots were collected under the same settings as the main study: temperature 0.7, identical commercial API routing, one retry with an unchanged prompt after an unusable answer, and label-based scoring. Pilot and main-study cells are therefore directly comparable.

Each part also ships seven automated checks. They confirm that all seven models and all contexts are present, and that every cell has the expected number of replicates. They also check that missing answers were preserved rather than filled in, that the processed scores follow from the raw answers, that the cell arithmetic is correct, and that no compound cell contains a missing answer. All checks pass on the released bundles.

The package is publicly archived at Zenodo (DOI 10.5281/zenodo.21489805) and browsable at https://github.com/mbrcic/llm-political-steerability.