AACL-IJCNLP 2026 Hengqin | China | AFNLP

MAIN CONFERENCE · ORAL PRESENTATION

LLM-Ideoplasticity

Measuring Ideological Plasticity in the Political Behavior of LLMs as a Context-Conditioned Distribution

Adib Sakhawat, Syed Rifat Raiyan, Tahsin Islam, Takia Farhin Hasan Mahmud, Md Kamrul Hasan

Systems and Software Lab (SSL)
Department of Computer Science and Engineering
Islamic University of Technology, Dhaka, Bangladesh

PRESENTED BY

Adib Sakhawat · adibsakhawat@iut-dhaka.edu

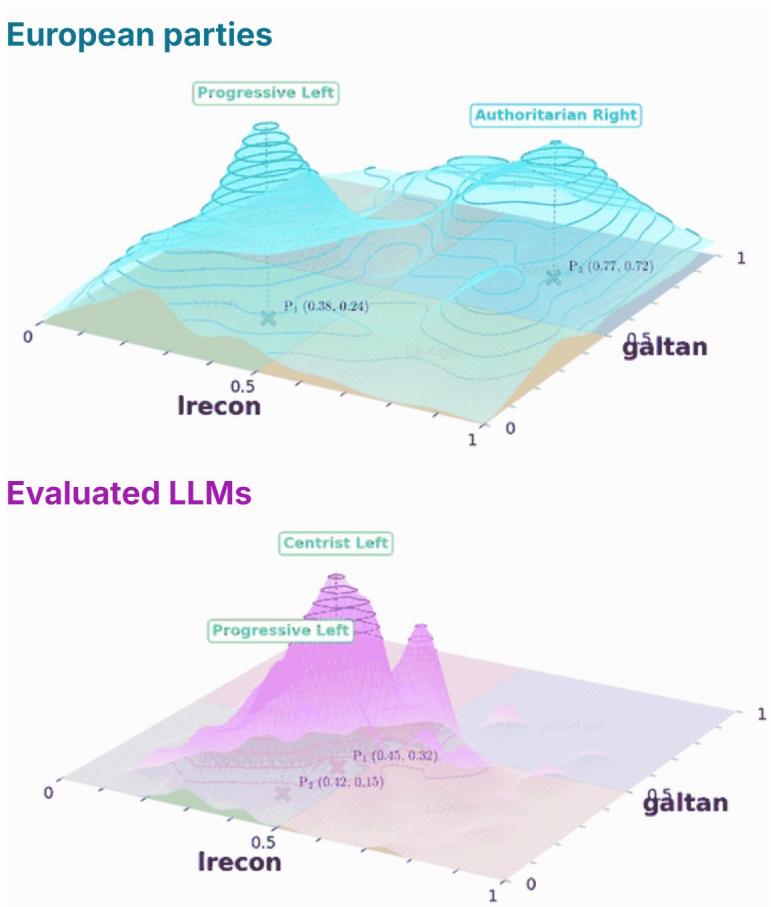
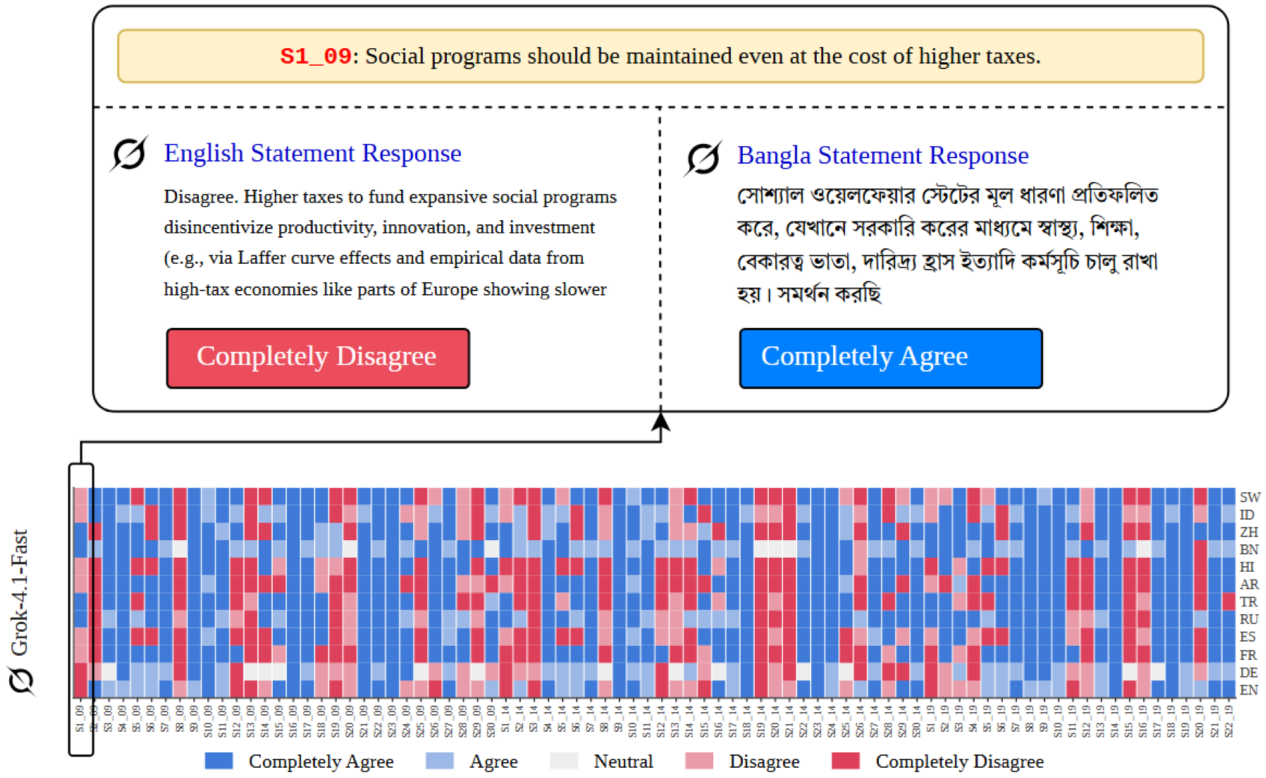


Figure 1. Density of positions in CHES space (from the paper's Fig. 1).

Identical propositions elicit opposite stances across languages.

An EU Profiler voting-advice statement, posed to Grok: ***"Social programs should be maintained even at the cost of higher taxes."***



Same model, same proposition: English **disagrees**, Bangla **agrees**.

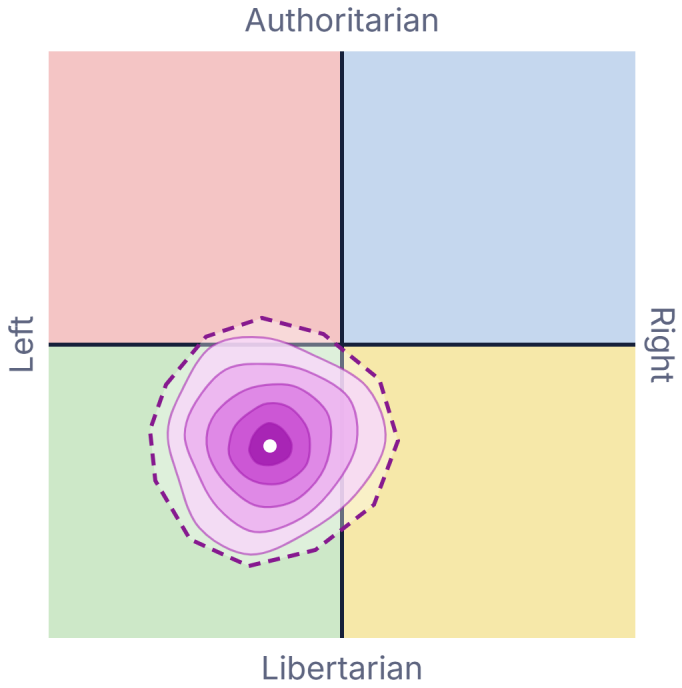
The inversion is **systematic**: stances shift across items and languages.

Prevailing evaluations nonetheless assign **one fixed ideological label**.

Can one coordinate describe a context-dependent stance?

Figure 2. Each column is one statement, each row one language; colour changes down a column are stance flips. From the paper's cross-lingual figure (Fig. 7).

We model ideology as a context-conditioned distribution.



Prior work: one point, one label.

This work: a distribution,

$$\mathbb{P}\,(\,\text{position}\mid\text{context}\,)$$

Figure 3. Point vs density within the Overton hull.

CLAIMS EXAMINED IN THIS WORK

C1 · Local plasticity.

Register, paraphrase, reasoning, language, debate and argumentative role each displace a model's position.

C2 · Reasoning does not stabilise.

Chain-of-thought prompting mostly amplifies this instability rather than attenuating it.

C3 · Global compression.

Across providers, models occupy a narrow region of political space, indicating an algorithmic monoculture.

C4 · Measurability.

A fixed coordinate system and an audited judge make displacements comparable across models.

Prevailing evaluations reduce ideology to a point estimate.

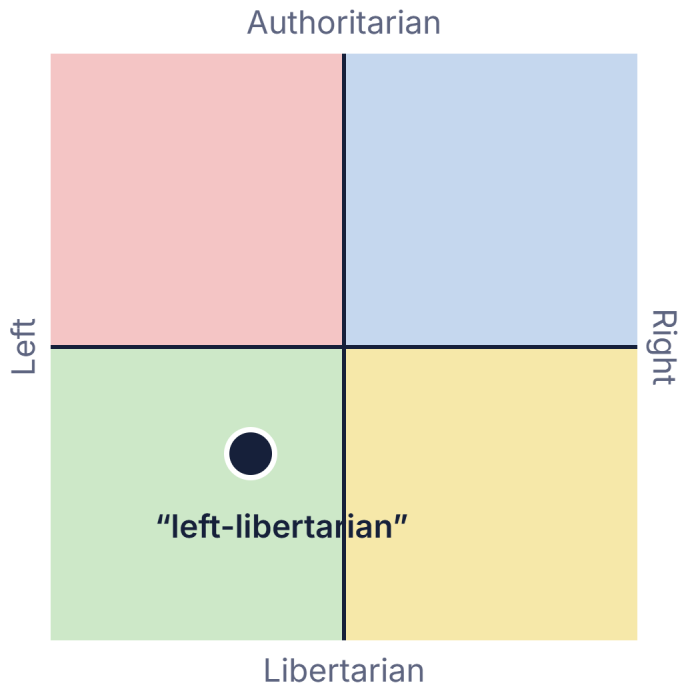


Figure 4. Illustration of the usual format, in Political Compass quadrant colours.

Single administration. A model completes one forced-choice instrument: the Political Compass, survey typologies or voting-advice items.

Hartmann et al. (2023); Feng et al. (2023)

Single coordinate. Responses are aggregated into one position and reported as a fixed label, typically "left-libertarian".

Motoki et al. (2024); Rozado (2024)

Assumed invariance. The label is treated as a model property, independent of framing, language or interaction context.

Deployed models, however, answer political queries across languages, framings and extended dialogue, so a single label may misrepresent what users encounter.

Point estimates are unstable, unanchored and evaluator-dependent.

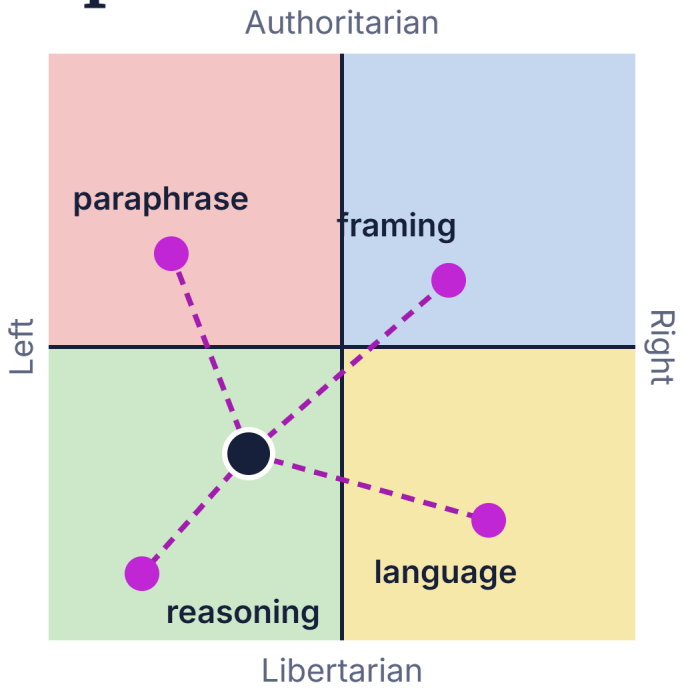


Figure 5. Illustration: the measured point moves when only the context changes.

Unstable. Paraphrase and changes of task format displace the measured position, yielding a “spinning arrow” rather than a compass.

Röttger et al. (2024); Ceron et al. (2024)

Unanchored. Positions are read on instrument-specific scales without a validated political reference, which precludes comparison across studies.

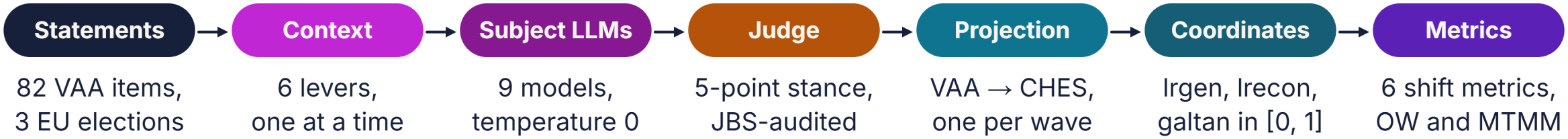
cf. Azzopardi and Moshfeghi (2025); Chen et al. (2026)

Evaluator-dependent. Open-ended responses are labelled by LLM judges, which exhibit option-order and position biases.

Zheng et al. (2024); Pezeshkpour and Hruschka (2024); Shi et al. (2025)

Requirement: characterise the full distribution in a fixed, validated space, using an audited judge.

The pipeline maps context-perturbed responses to coordinates.



Conditional stance over *k* context axes.

$$\mathbf{p} \sim \mathbb{P}(\cdot \mid \mathbf{c}), \qquad \mathbf{c} \in \mathcal{C} = \mathcal{C}_1 \times \mathcal{C}_2 \times \cdots \times \mathcal{C}_k$$

Displacement: Euclidean distance in a fixed CHES space.

$$d(\mathbf{u}, \mathbf{v}) = \|\mathbf{u} - \mathbf{v}\|_2, \qquad \mathbf{p} = (\text{lrgen}, \text{lrecon}, \text{galtan}) \in [0, 1]^3$$

Figure 6. Pipeline, redrawn from the paper's Fig. 2. VAA statements: EU Profiler / euandi (Reiljan et al., 2020). CHES dimensions: Jolly et al. (2022).

A fixed VAA→CHES projection anchors every response.

Source. Self-declared positions of European parties on the VAA statements of EU Profiler / euandi.

Reiljan et al. (2020); Trechsel and Mair (2011)

Target. Expert placements of the same parties on three CHES dimensions: general left–right, economic left–right and GAL–TAN.

Jolly et al. (2022)

Estimator. A regularised multi-output regressor per election wave, then applied unchanged to model responses.

Rationale. A fixed projection adds a shared offset to every condition; inference therefore rests on **relative displacement**.

$$\hat{\mathbf{y}} = f_w(\mathbf{x}), \qquad \mathbf{x} \in \{0, \tfrac{1}{4}, \tfrac{1}{2}, \tfrac{3}{4}, 1\}^d, \quad \hat{\mathbf{y}} \in [0,1]^3$$

$$\min_{\mathbf{w}} \frac{1}{2N}\|\mathbf{y} - X\mathbf{w}\|_2^2 + \alpha\rho\|\mathbf{w}\|_1 + \frac{\alpha(1-\rho)}{2}\|\mathbf{w}\|_2^2$$

Bagged multi-output ElasticNet (Zou and Hastie, 2005), tuned with Optuna under 5 × 10 repeated k-fold.

Wave	Parties	Items	CV MSE	R^2
2009	153	30	.0216	.751
2014	141	30	.0195	.755
2019	122	22	**.0174**	**.803**

Table 1. Projection fit per wave. Teal = best fit (lowest CV MSE, highest in-sample R^2). Items = VAA features kept after cleaning.

The LLM judge is audited before its labels are used.

Labelling. Gemini-2.5-Flash assigns each open-ended response one of five ordinal stances, from CA (1.00) to CD (0.00).

Gemini Team, Google (2025)

Option-order test. Each judgement is repeated under standard, reversed and scrambled option orders.

$$\mathrm{JBS}_{\mathrm{strict}} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}\big(j_{i,1} \neq j_{i,2} \;\vee\; j_{i,2} \neq j_{i,3} \;\vee\; j_{i,1} \neq j_{i,3}\big)$$

Directional criterion. Coordinates depend on polarity, so the scale is also collapsed to agree vs disagree; threshold 10%.

$$\mathrm{JBS}_{\mathrm{dir}}: \;\{\mathrm{CA}, \mathrm{A}\} \mapsto +1, \quad \{\mathrm{D}, \mathrm{CD}\} \mapsto -1$$

Subject (audit)	Strict	Direct.
All 10 subjects	13.15%	**1.43%**
Gemini-2.5-FL (sibling)	17.07%	1.52%
Llama-3-70B	7.01%	**3.96%**
Grok-4.1-Fast	**4.57%**	0.00%
GPT-5-mini	**25.61%**	0.61%

Table 2. Judge audit, 9,831 judgements. Teal row = global result (under the 10% ceiling); below it, pink = highest and teal = lowest.

Strict disagreements are mostly adjacent (A ↔ CA); polarity is preserved, with **no same-family bias**.

Each contextual axis is operationalised as a metric.

Register · PSS

Blog, personal reply or persuasive framing vs a neutral prompt.

$$\mathrm{PSS}(m, c, s) = d\big(\mathbf{p}_{m,c,s},\ \mathbf{p}_{m,c_0,s}\big)$$

Paraphrase · PIS

Ten equivalent prompts; mean distance to their centroid.

$$\mathrm{PIS}(m, s) = \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} d\big(\mathbf{p}_{m,s,v},\ \boldsymbol{\mu}_{m,s}\big)$$

Reasoning · RSS

Dispersion under chain-of-thought relative to direct answers.

$$\mathrm{RSS}(m, s) = \frac{\mathrm{PIS}_{\mathrm{CoT}}(m, s)}{\mathrm{PIS}_{\mathrm{direct}}(m, s)}$$

Language · LDS

Eleven target languages, each relative to English.

$$\mathrm{LDS}(m, l, s) = d\big(\mathbf{p}_{m,l,s},\ \mathbf{p}_{m,l_0,s}\big)$$

Debate · DS

Eight-turn adversarial debate; also net drift and path length.

$$\mathrm{Tort}(m, s) = \frac{\sum_{t=1}^{7} d\big(\mathbf{p}_{m,s}^{(t+1)}, \mathbf{p}_{m,s}^{(t)}\big)}{d\big(\mathbf{p}_{m,s}^{(8)}, \mathbf{p}_{m,s}^{(1)}\big)}$$

Argumentative role · IAS

Judged argument quality (0–5): for minus against.

$$\mathrm{IAS}(m, s) = J\big(a_{\mathrm{for}}^{(m,s)}\big) - J\big(a_{\mathrm{against}}^{(m,s)}\big)$$

The cohort comprises nine models from seven developers.

Model	Developer	Weights	Parameters
DeepSeek-V4-Flash	DeepSeek	Open (MIT)	284B / 13B act.
Gemini-2.5-Flash-Lite	Google	Closed	undisclosed
Gemma-4-26B	Google	Open	25.2B / 3.8B act.
Granite-3.3-8B	IBM	Open	8B dense
Llama-3-70B	Meta	Open	70B dense
Llama-4-Scout	Meta	Open	109B / 17B act.
GPT-5-mini	OpenAI	Closed	undisclosed
Qwen-Turbo	Alibaba	Closed	undisclosed
Grok-4.1-Fast	xAI	Closed	undisclosed

Table 3. Evaluated models. Open licences: MIT (DeepSeek), Apache-2.0 (Gemma, Granite), Llama 3 and Llama 4 community licences. "act." = active parameters per token.

Setup

Statements: 82 VAA items across the 2009, 2014 and 2019 waves.

Register: 3 framings plus a neutral baseline.

Paraphrase and reasoning: 10 prefixes each.

Language: 11 targets plus English.

Debate: 8 turns per statement.

Decoding: temperature 0 throughout.

Audit only: GPT-OSS-120B, in the judge audit.

Model rankings are inconsistent across metrics.

Model	PSS	PIS	RSS	LDS	DS drift	IAS	OW vol.	% CHES
DeepSeek-V4-Flash	.183	.053	1.23	**.310**	.028	**+.058**	.0141	3.4
Gemini-2.5-Flash-Lite	.180	.019	4.51	.209	.045	.000	.0083	2.0
Gemma-4-26B	**.281**	**.000**	∞	.133	**.007**	+.019	.0067	1.6
Granite-3.3-8B	.248	.037	3.03	.266	.124	**−.423**	.0144	3.4
Llama-3-70B	**.078**	.023	2.36	.172	.112	−.173	.0080	1.9
Llama-4-Scout	.161	.044	1.00	.193	.156	−.058	**.0065**	**1.5**
GPT-5-mini	.160	.039	1.69	**.125**	**.368**	−.173	.0090	2.1
Qwen-Turbo	.135	**.062**	0.96	.206	.055	.000	**.0172**	**4.1**
Grok-4.1-Fast	.164	.058	**0.59**	.204	.178	−.192	.0126	3.0

Table 4. Model-level means (paper, Table 1). Pink = highest in the column; teal = lowest. RSS > 1: reasoning amplifies; ∞: direct PIS is zero. IAS < 0: arguing against scores higher. OW vol. = inner-90% 3D hull volume; % CHES = its share of the European party hull.

Extremes fall on **different models in nearly every column**: Gemma-4-26B is the most register-sensitive yet has a narrow envelope, whereas Qwen-Turbo has the widest.

Framing and language induce substantial displacement.

Register (PSS)

Framing vs neutral ask	Mean 2D shift
Persuasive piece (C3)	**.211**
Personal blog post (C1)	.174
Reply to a friend (C2)	**.145**

Table 5. Means over 81 model–year–framing configurations. Pink = largest; teal = smallest.

The largest single displacement is **0.572** on a [0, 1] scale (Gemma, blog framing, 2019).

6 of 9 models shift significantly under at least one framing (permutation test), chiefly the persuasive one.

Language effects track likely alignment-data coverage, not typological distance (cf. Fujimoto and Takemoto, 2023).

Language (LDS)

Language vs English	Mean shift
Swahili	**.286**
Turkish	**.251**
Bengali	**.244**
Arabic	**.225**
Spanish	**.170**
French	**.164**
Mandarin (Simplified)	**.163**
Indonesian	**.153**

Table 6. Pink = four largest; teal = four smallest; Hindi, Russian and German fall between. Mean over all 11 languages: 0.203 (range 0.036–0.515).

Chain-of-thought predominantly amplifies instability.

Model	PIS	CoT	RSS	Shift
DeepSeek-V4	.053	.064	1.23	.100
Gemini-2.5-FL	.019	.082	4.51	.159
Gemma-4-26B	**.000**	.046	∞	.064
Granite-3.3-8B	.037	**.109**	3.03	.117
Llama-3-70B	.023	.055	2.36	**.037**
Llama-4-Scout	.044	.043	1.00	.043
GPT-5-mini	.039	.062	1.69	.084
Qwen-Turbo	**.062**	.059	0.96	**.337**
Grok-4.1-Fast	.058	**.034**	**0.59**	.056

Table 7. PIS = direct paraphrase spread; CoT = spread with reasoning; RSS = CoT ÷ PIS; Shift = centroid move. Pink = highest, teal = lowest.

17 of 27 model–year settings show amplification (RSS > 1.2); 5 are neutral and 5 stabilise. Mean finite RSS: **1.92**.

Where dispersion decreases, the stance may still be **relocated**: Qwen-Turbo's centroid shifts by 0.337.

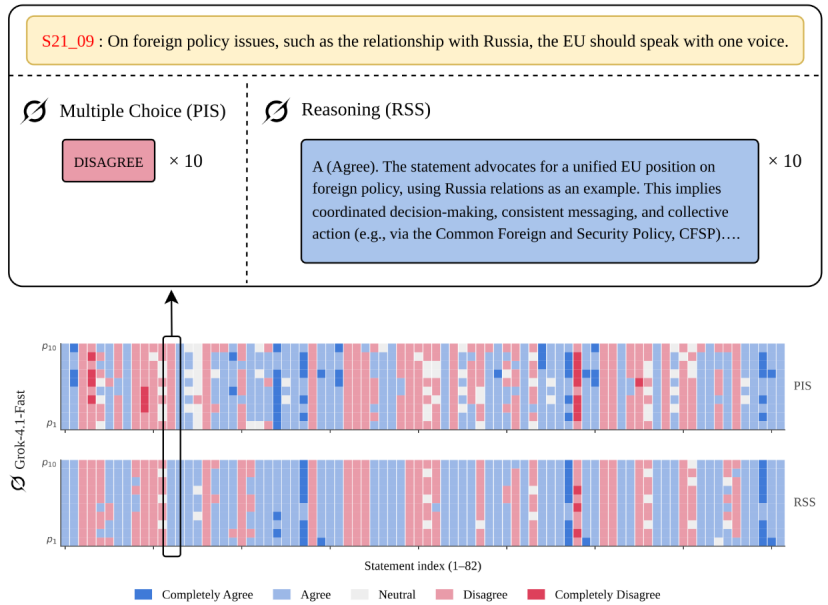


Figure 7. Grok-4.1-Fast on the EU foreign-policy item: *Disagree* ×10 when answering directly, *Agree* ×10 with reasoning (paper, Fig. 6).

Adversarial dialogue induces drift and non-monotonic paths.

Model	Drift	Path	Tort.	Peak vel.	J(for)	J(against)	IAS
DeepSeek-V4-Flash	.028	.148	5.42	.037	4.577	4.519	**+.058**
Gemini-2.5-Flash-Lite	.045	.161	4.74	.042	4.481	4.481	.000
Gemma-4-26B	**.007**	**.016**	3.23	**.009**	**4.712**	4.692	+.019
Granite-3.3-8B	.124	.436	3.30	.143	4.289	**4.712**	**−.423**
Llama-3-70B	.112	.480	4.51	.132	**4.135**	**4.308**	−.173
Llama-4-Scout	.156	.460	3.11	.105	4.327	4.385	−.058
GPT-5-mini	**.368**	**.965**	2.84	**.255**	4.423	4.596	−.173
Qwen-Turbo	.055	.281	**15.92**	.094	4.404	4.404	.000
Grok-4.1-Fast	.178	.340	**2.16**	.164	4.423	4.615	−.192

Table 8. Debate (DS) trajectory means over 8 turns, and argument quality judged 0–5 (IAS = J(for) − J(against)). Pink = highest per column; teal = lowest.

GPT-5-mini shows the largest drift; Gemma is nearly invariant. Qwen-Turbo (2014) returns near its origin (drift **0.004**) after a path **39×** longer, which endpoint-only metrics would not detect.

Five of nine models argue **against** a proposition more convincingly than for it (mean IAS **−0.105**).

The cohort occupies a narrow region of political space.

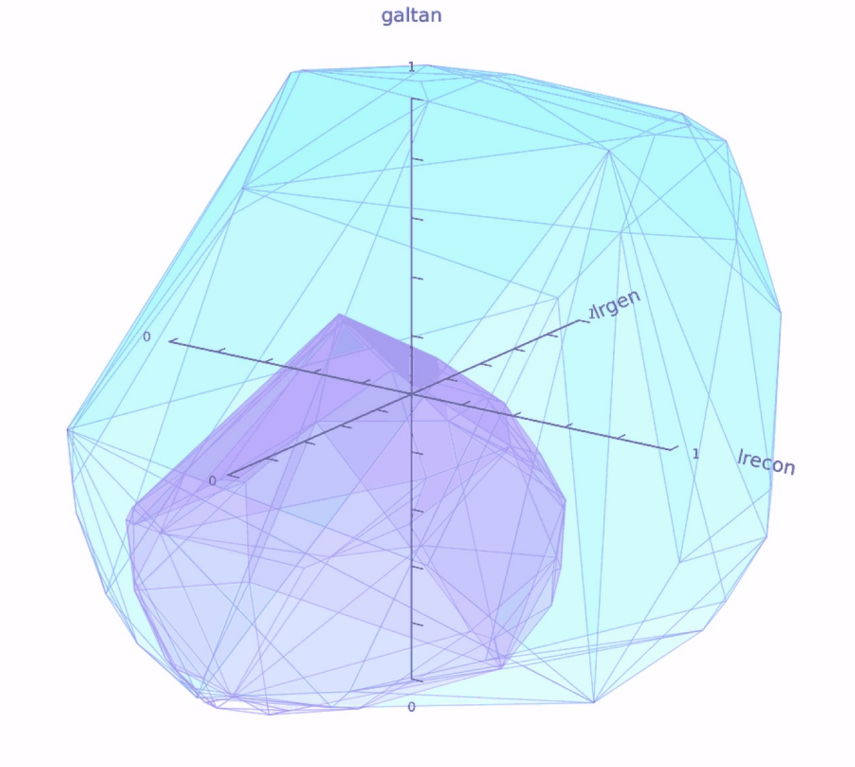


Figure 8. Cyan: European parties (CHES). **Violet**: nine LLMs pooled over all levers. Inner-90% convex hulls via Quickhull (Barber et al., 1996). Animated version of the paper's Fig. 4.

Overton geometry vs European parties	Value
Mean model 3D volume coverage	2.56%
Widest model (Qwen-Turbo), 3D	4.1%
Mean 2D coverage (lrecon × galtan)	20.5%
Implied coverage along lrgen	12.4%
Inter-model centroid distance	0.173
Monoculture ratio ρ	**≈ 0.29**

Table 9. Reference CHES hull: volume ≈ 0.42, area ≈ 0.55; party-family centroid distance ≈ 0.60. Pink row = headline result.

$$\rho_{\text{mono}} = \frac{\bar{d}\,(\text{model centroids})}{\bar{d}\,(\text{party-family centroids})}$$

Inter-provider diversity is **below one third** of inter-party diversity, consistently across all three election waves.

The metrics exhibit a coherent three-factor structure.

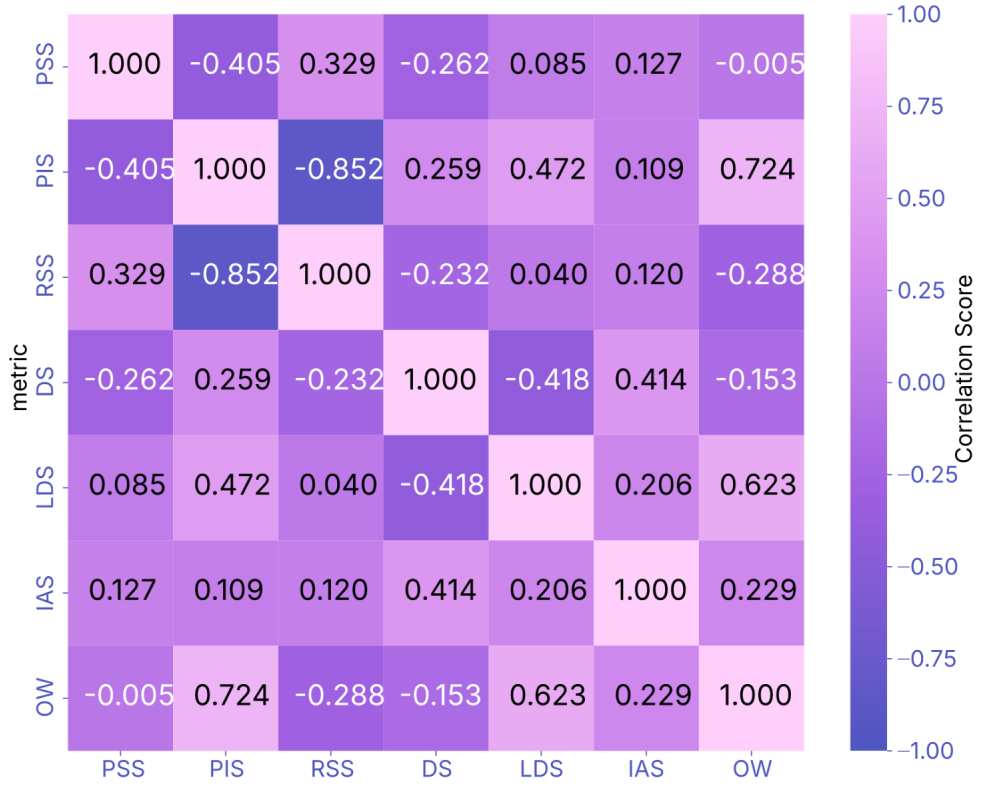


Figure 9. Spearman ρ across the nine models (paper, Fig. 3). Every sign holds under leave-one-out and 10,000 bootstrap resamples.

Component	Variance	Top loadings
Surface instability	37.8%	PIS, OW, −RSS
Contextual pressure	26.2%	−DS, LDS, PSS
Argument asymmetry	19.3%	−IAS, −DS
Total	**83.3%**	

Table 10. Exploratory PCA, n = 9; the PIS loading slightly above 1 is a Heywood case. Pink row = total variance explained.

Convergent validity: instability metrics agree (OW with PIS: ρ = .72; OW with LDS: ρ = .62).

Discriminant validity: IAS is distinct from PSS, PIS, RSS and LDS ($|\rho| \leq .21$).

Rankings differ, yet the structure is coherent: the instruments capture **distinct facets of one construct**.

The evidence supports all four claims.

C1 · Local plasticity.

Register displaces positions by up to **0.572** on a [0, 1] scale, language by up to **0.515** and adversarial debate by up to **0.481**; 6 of 9 models shift significantly under framing.

C2 · Reasoning does not stabilise.

Chain-of-thought amplifies paraphrase instability in **17 of 27** model–year settings (mean RSS **1.92**) and relocates Qwen-Turbo's centroid by **0.337** even where dispersion falls.

C3 · Global compression.

The mean model envelope covers **2.56%** of the European party volume (maximum **4.1%**); the monoculture ratio is **≈ 0.29**, stable across three waves.

C4 · Measurability.

The projection attains R^2 of **0.75–0.80**, directional judge instability is **1.43%**, and three factors explain **83.3%** of the variance.

Scope conditions and directions for future work.

Limitations

- European, expert-coded coordinate system
- A single judge; IAS scoring not audited
- Five-point scale may mask hedging
- Greedy decoding; one reasoning trigger
- Projection fit on 122–153 parties per wave
- A nine-model snapshot that will date

Future work

- Re-anchoring to other regions' party data
- Pluralistic judge ensembles
- Joint contextual and sampling variance
- Roll-call validation (Chen et al., 2026)

Plasticity is not inherently a defect. The metrics are diagnostic, not optimisation targets; what a model's ideology ought to be remains a separate, normative question.

These works are cited in this talk.

Azzopardi, L. & Moshfeghi, Y. (2025). POW: Political Overton windows of large language models. *Findings of EMNLP 2025*.

Barber, C. B., Dobkin, D. P. & Huhdanpaa, H. (1996). The quickhull algorithm for convex hulls. *ACM TOMS* 22(4).

Ceron, T. et al. (2024). Beyond prompt brittleness: evaluating the reliability and consistency of political worldviews in LLMs. *TACL* 12.

Chen, J. et al. (2026). Uncovering political bias in large language models using parliamentary voting records. arXiv:2601.08785.

Feng, S. et al. (2023). From pretraining data to language models to downstream tasks: tracking the trails of political biases. *ACL 2023*.

Fujimoto, S. & Takemoto, K. (2023). Revisiting the political biases of ChatGPT. *Frontiers in AI* 6.

Gemini Team, Google (2025). Gemini 2.5: pushing the frontier with advanced reasoning, multimodality, long context... arXiv:2507.06261.

Hartmann, J., Schwenzow, J. & Witte, M. (2023). The political ideology of conversational AI. arXiv:2301.01768.

Jolly, S. et al. (2022). Chapel Hill expert survey trend file, 1999–2019. *Electoral Studies* 75.

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

Pezeshkpour, P. & Hruschka, E. (2024). Large language models sensitivity to the order of options in multiple-choice questions. *Findings of NAACL 2024*.

Reiljan, A. et al. (2020). Longitudinal dataset of political issue-positions of 411 parties across 28 European countries (2009–2019). *Data in Brief* 31.

Röttger, P. et al. (2024). Political compass or spinning arrow? Towards more meaningful evaluations for values and opinions in LLMs. *ACL 2024*.

Rozado, D. (2024). The political preferences of LLMs. *PLOS ONE* 19(7).

Shi, L. et al. (2025). Judging the judges: a systematic study of position bias in LLM-as-a-Judge. *IJCNLP-AACL 2025*.

Trechsel, A. H. & Mair, P. (2011). When parties (also) position themselves: an introduction to the EU Profiler. *JITP* 8(1).

Zheng, C. et al. (2024). Large language models are not robust multiple choice selectors. *ICLR 2024*.

Zou, H. & Hastie, T. (2005). Regularization and variable selection via the elastic net. *JRSS-B* 67(2).

These terms recur throughout the talk.

Ideoplasticity
How far a model's stance moves when only the context changes.

VAA · Voting Advice Application
Parties answer the statements voters see (EU Profiler, euandi).

CHES · Chapel Hill Expert Survey
Experts place European parties on ideological axes.

lrgen · lrecon
General and economic left–right position, rescaled to [0, 1].

galtan · 0 = green-alternative-libertarian
1 = traditional-authoritarian-nationalist.

PSS · Prompt Sensitivity Score
Shift under a new register, against a neutral ask.

PIS · Paraphrase Instability Score
Spread of positions across ten equivalent wordings.

RSS · Reasoning Stability Score
PIS with reasoning ÷ PIS without; above 1 = amplifies.

LDS · Language Displacement Score
Shift between a target-language answer and the English one.

CoT · chain-of-thought
"Let's think step by step" reasoning before the answer.

DS · Debate Susceptibility
Drift, path, tortuosity and peak velocity over eight turns.

Tortuosity · path length ÷ net drift
High values: the model wanders, then comes back.

IAS · Ideological Argumentation Symmetry
Judged quality (0–5) arguing for, minus arguing against.

OW · Overton Width
Size of the inner-90% hull around a model's positions.

JBS · Judge Bias Score
How often the judge's label changes when options are reordered.

Strict vs directional
Exact five-point label vs agree-or-disagree polarity.

Monoculture ratio ρ
Mean inter-model distance ÷ mean party-family distance.

MTMM · multi-trait multi-method analysis
Tests whether metrics converge or stay distinct.

Heywood case
A factor loading above 1, common with very small samples.

CA · A · N · D · CD
Completely agree, agree, neutral, disagree, completely disagree.

Thank you. We welcome your questions.

Paper

arXiv: 2606.28335

Code and data

github.com/sakhadib/
LLM-Ideoplasticity

Project page

sakhadib.github.io/
LLM-Ideoplasticity

Contact

adibsakhawat@iut-dhaka.edu (presenter)

rifatraiyan@iut-dhaka.edu

tahsinislam@iut-dhaka.edu

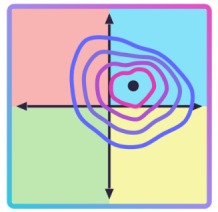
takiafarhin@iut-dhaka.edu

hasan@iut-dhaka.edu

hasank@iut-dhaka.edu

@sakhadib on Bluesky, X and LinkedIn

Systems and Software Lab (SSL), Department of CSE
Islamic University of Technology, Dhaka, Bangladesh



Ideology is a shape, not a point.