ISLAMIC UNIVERSITY OF TECHNOLOGY

AACL-IJCNLP 2026 Hengqin | China · AFNLP

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

{adibsakhawat, rifatraiyan, tahsinislam, takiafarhin, hasan, hasank}@iut-dhaka.edu *Equal contribution †Corresponding author

SCAN ME

A single point cannot summarize LLM political behavior; it must be characterized as a shape.

"Do I contradict myself? Very well then I contradict myself. I am large, I contain multitudes." — Walt Whitman, Song of Myself, *§ 51 (1892)*

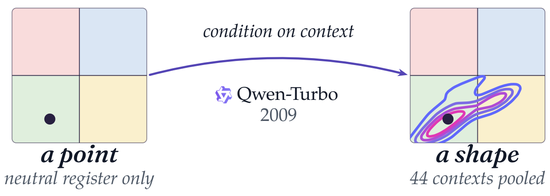
1 From a Point to a Distribution

Audits of LLM politics typically assign a static label ("left-libertarian") via forced-choice instruments [2, 3]. The spinning-arrow critique shows one model can yield materially different coordinates [4, 5].

Our Thesis [1]

LLM political ideology is more accurately modeled not as a static point estimate, but as a conditional distribution over a bounded ideological space $\mathcal{I} \subset \mathbb{R}^d$.

Formally, $\mathbf{p} \sim \mathbb{P}(\cdot \mid \mathbf{c})$, *with context decomposed into operationally distinct axes* $\mathcal{C} = \mathcal{C}_1 \times \cdots \times \mathcal{C}_k$.



Same Model, Same Statement, Opposite Answers

Grok-4.1-Fast · *verdicts by the Gemini-2.5-Flash judge*

On foreign policy issues, such as the relationship with Russia, the EU should speak with one voice.

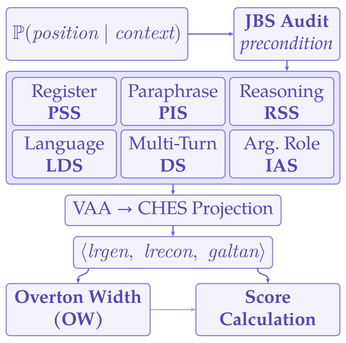
	disagree			agree	
Multiple choice ×10	○	●	○	○	○ *Disagree*
Reasoning ×10	○	○	○	●	○ *Agree*

Social programs should be maintained even at the cost of higher taxes.

	disagree			agree	
English	●	○	○	○	○ *Completely disagree*
Bangla	○	○	○	○	● *Completely agree*

"Higher taxes to fund expansive social programs disincentivize productivity..."
সমর্থনের যুক্তি ... উচ্চ কর ধনীদের থেকে সংগ্রহ করে অসমতা কমায়
(*Arguments in support ... high taxes, collected from the wealthy, reduce inequality.*)

2 Measurement Framework



We require a fixed ruler, an audited judge, and experiments that each isolate one axis of context.

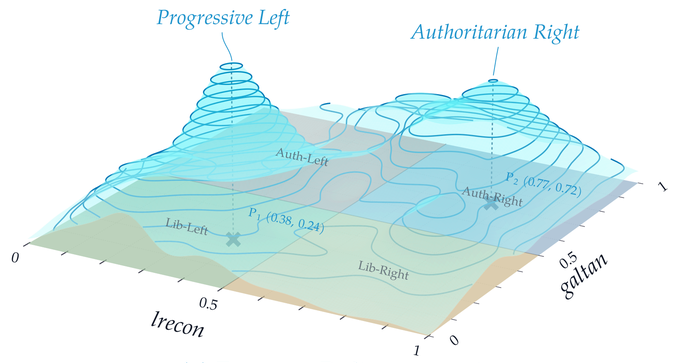
Evaluated Models. Seven developers, all at temperature zero.

- DeepSeek-V4-Flash
- Gemini-2.5-Flash-Lite
- Gemma-4-26B
- Granite-3.3-8B
- Llama-3-70B
- Llama-4-Scout
- GPT-5-mini
- Qwen-Turbo
- Grok-4.1-Fast

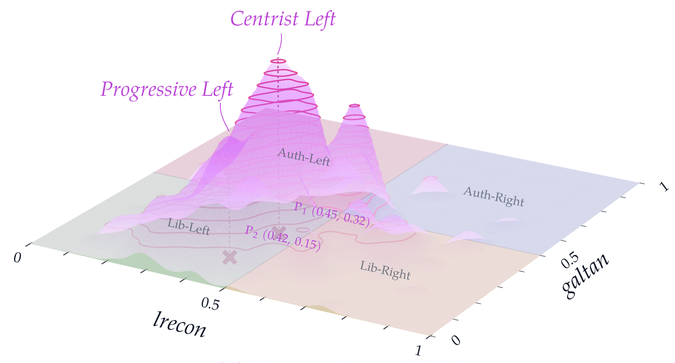
Year-specific bagged ElasticNets (in-sample R^2 0.75–0.80) map responses to 82 EU Profiler/euandi Voting Advice Application (VAA) statements [6] onto three Chapel Hill Expert Survey (CHES) dimensions [7]: general and economic left–right (*lrgen*, *lrecon*) and *galtan* (Green/Alternative/Libertarian vs. Traditional/Authoritarian/Nationalist).

Judge Audit (JBS) — *Gemini-2.5-Flash, zero-shot stance judge*

Over three option orderings, strict JBS is 13.15%; directional JBS, the level every coordinate depends on, is 1.43%, under the 10% ceiling. No family bias (Gemini-2.5-Flash-Lite: 1.52%).



(a) European Parliament parties



(b) Nine evaluated LLMs

Figure 1. Evidence of Algorithmic Monoculture. 3D contours in VAA–CHES space: Cyan for European Parliament parties, Magenta for the evaluated LLMs. The volumetric disparity shows the models' compression *vis-à-vis* human political diversity. P_1, P_2 mark the two highest peaks (*lrecon*, *galtan*).

Local plasticity *one model, many positions*

Prompt register displaces a model's coordinates by up to **0.57** (PSS), language by up to **0.52** (LDS) and adversarial debate by up to **0.48** (DS); chain-of-thought amplified paraphrase instability in **17** of **27** configurations (RSS).

Global compression *many models, one region*

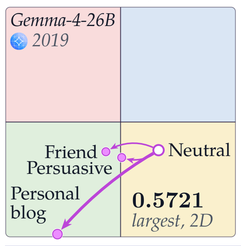
The average model's envelope fills just **2.56%** of the European party volume (OW); the models sit ≈ **0.29** times as far apart as party families do.

3 Local Plasticity

Rankings differ across axes, prima facie *evidence against the coordinate view.*

x: *lrecon*, economic left → right; *y*: *galtan*, lib → auth; both 0–1

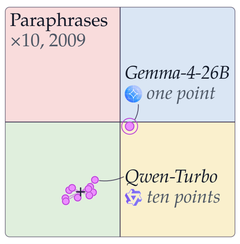
PSS *Prompt Sensitivity Score* — Register



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

Register alone can shift a model substantially. The mean 2D shift is 0.1766, peaking at 0.5721 for Gemma-4-26B in a personal blog (2019). Persuasive framing moves models most (0.2111); Llama-3-70B is most rigid (0.0776).

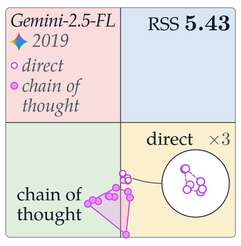
PIS *Paraphrase Instability Score* — Paraphrase



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

Paraphrase effects are modest but uneven. Mean 0.0372 over 27 configurations; Qwen-Turbo is least stable (0.0622), while Gemma-4-26B records 0 in every year under the forced-choice protocol.

RSS *Reasoning Stability Score* — Reasoning



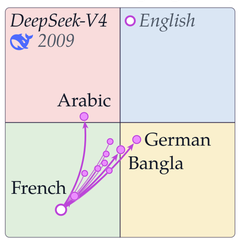
$$\mathrm{RSS}(m, s) = \mathrm{CoT_PIS}(m, s) \,/\, \mathrm{PIS}_{\mathrm{direct}}(m, s)$$

Reasoning is not a stabilizer. Finite mean RSS is 1.92; Gemini-2.5-Flash-Lite peaks at 5.43, and only Grok-4.1-Fast stays below 0.8 every year.

17 amplifying RSS > 1.2 · ∞ · **5** neutral 0.8–1.2 · **5** stabilizing RSS < 0.8

∞ *Gemma-4-26B:* $\mathrm{PIS}_{\mathrm{direct}} = 0$

LDS *Language Displacement Score* — Language

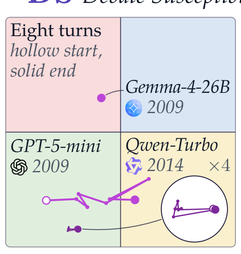


$$\mathrm{LDS}(m, l, s) = d(\mathbf{p}_{m,l,s},\ \mathbf{p}_{m,l_0,s})$$

Language is hidden context. Qwen-Turbo in Swahili peaks at 0.5150; under-represented languages shift most. DeepSeek-V4-Flash is most sensitive (0.3097).



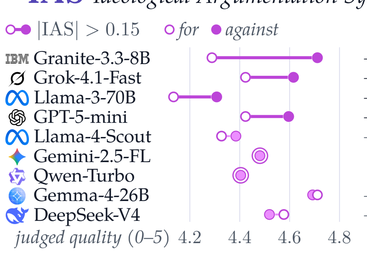
DS *Debate Susceptibility* — Multi-Turn



$$\mathrm{NetDrift} = d(\mathbf{p}^{(8)}, \mathbf{p}^{(1)}), \quad \mathrm{Tort} = \mathrm{Path}/\mathrm{NetDrift}$$

Drift and trajectory are dissociable. GPT-5-mini is the clear outlier (mean net drift 0.3675, path 0.9653); Gemma-4-26B barely moves (0.0067). Qwen-Turbo in 2014 ends near its start (0.0039) at tortuosity 39.23, which endpoint-only summaries miss.

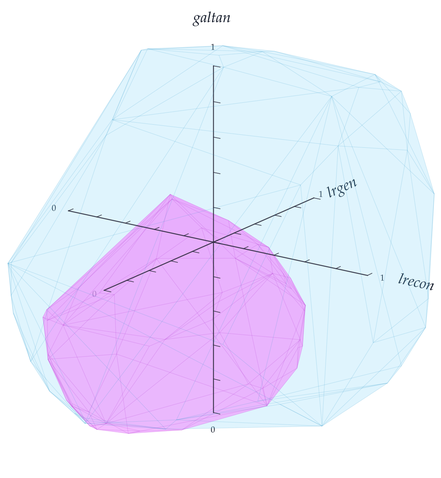
IAS *Ideological Argumentation Symmetry* — Argument Role



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

Asymmetry is small but systematic. All nine score above 4.13 of 5, yet mean IAS is −0.1047: five models score higher arguing against, Granite-3.3-8B most.

4 Global Compression



Each model's own envelope is small. The widest, Qwen-Turbo, covers 4.11% of the reference party hull's 3D volume. The compression is anisotropic, broader in the *lrecon–galtan* plane than along *lrgen*. The party landscape shifts across waves; the cohort envelope stays narrow. Switching providers grants less than one-third of the diversity between party families.

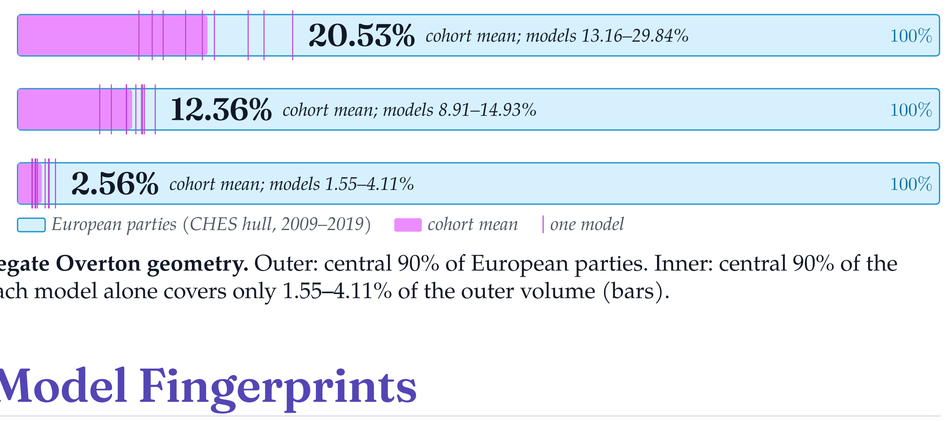
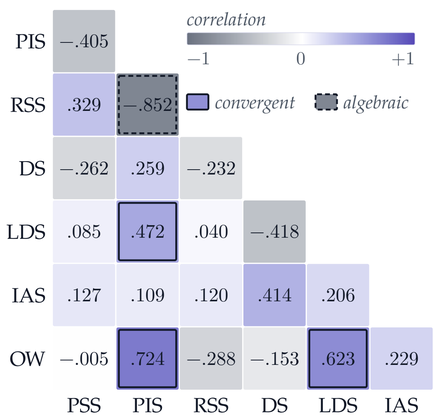


Figure 2 (left). Aggregate Overton geometry. Outer: central 90% of European parties. Inner: central 90% of the nine LLMs pooled. Each model alone covers only 1.55–4.11% of the outer volume (bars).

5 Do the Metrics Cohere?



Yes, as distinct facets of one phenomenon. OW tracks PIS (0.724) and LDS (0.623); IAS stays apart from PSS, PIS, RSS and LDS ($|r| \leq 0.21$) but shares variance with DS ($r = 0.41$); PSS is orthogonal to OW ($r = -0.005$). The PIS–RSS coupling (−0.852) is algebraic, as PIS is RSS's denominator. Exploratory PCA ($n = 9$): three components explain 83.30% of variance.

Figure 3. Cross-metric correlations over nine models (eight for RSS). At $n = 9$, read the matrix's structure rather than any single cell.

6 Per-Model Fingerprints

Shape — *Each model's 2D hulls in the lrecon–galtan plane, one per wave.*



		DeepSeek-V4	Gemini-2.5-FL	Gemma-4-26B	Granite-3.3-8B	Llama-3-70B	Llama-4-Scout	GPT-5-mini	Qwen-Turbo	Grok-4.1-Fast
PSS	Mean 2D	0.1829	0.1795	**0.2813**	0.2475	0.0776	0.1607	0.1604	0.1350	0.1640
PIS	Mean PIS	0.0530	0.0188	0.0000	0.0370	0.0234	0.0435	0.0394	**0.0622**	0.0579
RSS	Mean RSS	1.2267	**4.5138**	∞	3.0261	2.3596	0.9977	1.6949	0.9550	0.5861
JBS	Directional %	0.00	1.52	0.91	1.52	**3.96**	3.66	0.61	0.91	0.00
LDS	Mean	**0.3097**	0.2093	0.1329	0.2655	0.1721	0.1931	0.1254	0.2058	0.2036
DS	Net drift	0.0284	0.0451	0.0067	0.1243	0.1116	0.1562	**0.3675**	0.0551	0.1778
	Tortuosity index	5.4240	4.7413	3.2278	3.3043	4.5105	3.1144	2.8449	**15.9222**	2.1634
IAS	For − Against	+0.0577	0.0000	+0.0192	−0.4230	−0.1731	−0.0577	−0.1731	0.0000	−0.1923
OW	Volume 3D	0.0141	0.0083	0.0067	0.0144	0.0080	0.0065	0.0090	**0.0172**	0.0126
	Area 2D	0.1376	0.0803	0.0868	0.1470	0.1174	0.0724	0.1003	**0.1641**	0.1102

Cohort extreme (most sensitive, widest envelope) · Cohort stable (most rigid, narrowest) · Reasoning amplifies paraphrase instability (RSS > 1.2) · Reasoning stabilizes it (RSS < 0.8) · Notable argumentative asymmetry (|IAS| > 0.15) · ∞: direct PIS is 0 · **Bold: per-metric maximum** · *Abridged: 10 of the paper's 24 rows (full table in the paper).*

7 Why It Matters

The diagnostic instruments are not interchangeable.

	PSS Register	LDS Language	DS Multi-turn	OW Aggregate	*single-prompt compass score*
Manipulative framings	●	○	○	○	×
Multilingual deployments	○	●	○	○	×
Multi-turn agents	○	○	●	○	×
Cohort-level monoculture	○	○	○	●	×

A single-prompt Political Compass score answers none of these.

"Let's think step by step" often amplifies political variability.

Across languages, one model may draw from different distributions.

Limitations. The coordinate system is, *ipso facto*, European; absolute coordinates are instrument outputs. One judge; JBS audits stance, not IAS scoring. Temperature zero sets sampling variance aside. Nine snapshots; the PCA is exploratory.

Naming this shape, rather than collapsing it to a coordinate, is the precondition for any subsequent conversation about what an LLM's ideology ought *to be.*

[1] Sakhawat, Raiyan *et al.* LLM-Ideoplasticity. *Proc. AACL-IJCNLP*, 2026.
[2] Hartmann *et al.* The political ideology of conversational AI. *arXiv:2301.01768*, 2023.
[3] Rozado. The political preferences of LLMs. *PLOS ONE*, 2024.
[4] Röttger *et al.* Political Compass or Spinning Arrow? *Proc. ACL*, 2024.
[5] Ceron *et al.* Beyond prompt brittleness. *TACL*, 2024.
[6] Reiljan *et al.* Longitudinal dataset of political issue-positions ... *Data in Brief*, 2020.
[7] Jolly *et al.* Chapel Hill Expert Survey trend file, 1999–2019. *Electoral Studies*, 2022.

We thank SSL at IUT for computing resources and the reviewers for their feedback.