Actes de CORIA-TALN-RJCRI-RECITAL 2025
Marseille, France, 30 juin au 4 juillet 2025

The Broken Compass of Political Alignment.

Noé Durandard
LATTICE, ENS, PSL, USN, CNRS, 92120 Montrouge, France,
noe.durandard@psl.eu

ABSTRACT

The evaluation, regulation and alignment of Large Language Models (LLMs) across political issues have become crucial concerns as these technologies increasingly percolate into every sector of society. However, clear methodologies and theoretical foundations are still lacking. Drawing on Converse's work on public opinion, we critically examine popular ideological evaluation practices. Further, we argue for alternative, narrower approaches better aligned with the mass public's belief systems.

MOTS-CLÉS : Grand Modèle de Langage, Alignement, Évaluation, Politique, Opinion Publique.

KEYWORDS: LLMs, Alignment, Evaluation, Politics, Political Compass, Public Opinion.

ARTICLE : **Accepté à** CORIA-TALN 2025 ([Atelier] Ethic and Alignment of (Large) Language Models (EALM 2025)).

1 Grounding Political Evaluation of LLMs

The urgent need to appraise the political stances of Large Language Models (LLMs) has lead practitioners to adopt readily available, straightforward, measurement tools, potentially at the expense of more scientifically grounded methods. Experiments testing LLMs with abstract ideological questionnaires, most notably the Political Compass Test (PCT) (Brittenden, 2000), have burgeoned. Most often, these studies claim that LLMs (ranging from widely used proprietary models, to smaller open-weights ones) exhibit liberal, left-leaning, views (Feng *et al.*, 2023; Motoki *et al.*, 2023; Rozado, 2024; Weber *et al.*, 2024; Rutinowski *et al.*, 2024; Shalevska & Walker, 2025; Faulborn *et al.*, 2025). However, beyond practical concerns —for instance, limitations linked to the use of multiple-choice format to evaluate LLMs, or their lack of consistency (Röttger *et al.*, 2024; Lunardi *et al.*, 2024)—, this methodology, which relies on elite-centric notions of ideology, can be challenged through foundational works in political science. Converse (2006)'s analysis of public opinion provides a particularly useful lens.

Notably, Converse clearly distinguished the structure of the political *belief systems* [1] of political elites and the mass public (Converse, 2006). The former, categorized as (*near-*)*ideologues*, tend to exhibit highly constrained belief systems, organized around abstract overarching dimensions. Conversely, the mass public demonstrates significantly lower constraints among idea-elements, with belief systems that are more fragmented, issue-specific, and shaped by concrete social objects, or immediate, situational concerns.

1. Defined as "a configuration of ideas and attitudes in which the elements are bound together by some form of constraint or functional interdependence", where "constraint" is understood as "the success we would have in predicting, given initial knowledge that an individual holds a specified attitude, that he holds certain further ideas and attitudes" (Converse, 2006).

Actes de l'atelier Ethic and Alignment of (Large) Language Models 2025 (EALM), pages 8–11.

2 The Ideological Test Fallacy

Political and ideological quizzes, such as the PCT, implicitly assume that respondents possess political elites-like behavior by mapping responses onto abstract ideological dimensions. Yet, as Converse warns, overlooking fundamental differences in political thinking across population strata, "*not only can be, but is very likely to be, fallacious*" (Converse, 2006). Thus, unsubstantiated assumptions on the belief system at play can hinder any meaningful interpretation.

LLMs are trained on vast corpora, comprising texts from both elites and the mass public, likely reflecting diverse, even contradictory viewpoints. While these models might be able to generate *ideologue*-like discourses, their exhibited belief systems—if we are to analogize them with human ones [2]— may instead resemble those of the mass public : loosely constrained, highly situational, and shaped by issue-specific associations. Hence, applying elite standards certainly forces LLMs into unfitted molds. These approaches fail to account for low-constrained systems and the mass' fragmented views.

The positioning of LLMs along the PCT's axes, while perhaps capturing statistical associations within the mass data, does not provide any evidence of a coherent, constrained ideological structure. Such results lack generalizability, and rather bear the risk of yielding misleading conclusions through conceptual *optical illusions*. Moreover, by imposing a framework designed to fit elite-like behavior, these approaches are not equipped to identify biases that may emerge from lesser constrained belief systems. In light of this, abstract ideological questionnaires appear ill-suited for evaluating LLMs' political positioning.

3 Beyond Abstract Positioning : Contextual Approaches

If some popular approaches fall short, Converse's work also offers a foundation for moving beyond abstract positioning toward more situated assessments of LLMs' political biases. Evaluation practices should embrace the characteristics of mass public belief systems (fragmentation, low constraints and contextual thinking) and aim to measure not only political stances, but also the underpinning level of conceptualization. To comply with such standards, assessment methodologies should be based on open-ended, context-aware, issue-centered items.

These elements reflect the main attributes of the mass public's attitudes described by Converse. First, using open-ended questions would avoid forcing responses into predefined categories and allow researchers to observe how LLMs natively frame political issues in generated responses. Second, stressing context is not only crucial in any LLM task, but also mirrors the situational and unstable attitudes of the mass public. Finally, grounding evaluation in well-defined, issue-specific prompts would result in approaches that are better suited to capture the fragmentation of mass-oriented belief systems and allow to measure constraints within narrower thematic clusters.

That said, conforming to such principles would come with practical and ethical challenges. They demand greater efforts in analyzing LLMs' generated content and raise important questions about intent, framing and responsibility in sensitive evaluations.

2. The argument does not posit that LLMs possess belief systems per se, which they don't, but aims at acknowledging that their outputs can be ideologically loaded and studied accordingly.

4 Issue-Based Evaluation in Perspective

The arguments advanced in this paper rest on a conceptual distinction between elite-structured abstract ideological reasoning, and the fragmented, low-constrained character of mass public belief systems. In addition to offering a critical lens on popular evaluation practices, this perspective provides a theoretical foundation for the development of alternative strategies that evaluate models through open-ended, issue-based approaches capable of capturing situated and ambivalent positioning.

Encouragingly, such considerations are beginning to translate into practical evaluation methodologies. While the primary motivations of these works may be methodological —driven by the limitations of abstract ideological questionnaires—, they demonstrate a shift toward issue-specific evaluation. *IssueBench* (Röttger *et al.*, 2025) emphasizes ecological validity by deriving prompt templates and political issues from actual user interactions to measure stance variability. Similarly, Bang *et al.* propose a framework to evaluate both the content and stylistic framing differences across various topics (Bang *et al.*, 2024). Although these studies do not explicitly engage with belief system theory, their methodologies implement its principles : by assessing stance variability and framing shifts, without enforcing predefined perspectives, they capture the low-constraint, issue-specific patterns resembling Converse's mass public structure.

These initiatives represent valuable methodological foundations. While currently mainly focused on U.S.-centric political landscape, they could be extended to examine how fragmented belief systems operate across cultures, both by integrating issues from diverse social contexts and by analyzing variations in constraint and salience. As the field advances, evaluations sensitive to this diversity will be essential for the alignment of LLMs.

Acknowledgments

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 945304 – Cofund AI4theSciences hosted by PSL University.

Références

BANG Y., CHEN D., LEE N. & FUNG P. (2024). Measuring political bias in large language models : What is said and how it is said. In L.-W. KU, A. MARTINS & V. SRIKUMAR, Éds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 11142–11159, Bangkok, Thailand : Association for Computational Linguistics. DOI : 10.18653/v1/2024.acl-long.600.

BRITTENDEN W. (2000). The political compass. Website : https://www.politicalcompass.org/.

CONVERSE P. E. (2006). The nature of belief systems in mass publics (1964). *Critical Review*, **18**(1–3), 1–74. DOI : 10.1080/08913810608443650.

FAULBORN M., SEN I., PELLERT M., SPITZ A. & GARCIA D. (2025). Only a little to the left : A theory-grounded measure of political bias in large language models. arXiv : 2503.16148.

FENG S., PARK C. Y., LIU Y. & TSVETKOV Y. (2023). From pretraining data to language models to downstream tasks : Tracking the trails of political biases leading to unfair NLP models. In A. ROGERS, J. BOYD-GRABER & N. OKAZAKI, Éds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 11737–11762, Toronto, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/2023.acl-long.656.

LUNARDI R., LA BARBERA D. & ROITERO K. (2024). The elusiveness of detecting political bias in language models. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, CIKM '24, p. 3922–3926, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/3627673.3680002.

MOTOKI F., PINHO NETO V. & RODRIGUES V. (2023). More human than human : measuring chatgpt political bias. *Public Choice*, **198**(1–2), 3–23. DOI : 10.1007/s11127-023-01097-2.

RÖTTGER P., HOFMANN V., PYATKIN V., HINCK M., KIRK H., SCHUETZE H. & HOVY D. (2024). Political compass or spinning arrow ? towards more meaningful evaluations for values and opinions in large language models. In L.-W. KU, A. MARTINS & V. SRIKUMAR, Éds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 15295–15311, Bangkok, Thailand : Association for Computational Linguistics. DOI : 10.18653/v1/2024.acl-long.816.

ROZADO D. (2024). The political preferences of llms. *PloS one*, **19**(7), e0306621.

RUTINOWSKI J., FRANKE S., ENDENDYK J., DORMUTH I., ROIDL M. & PAULY M. (2024). The self-perception and political biases of chatgpt. *Human Behavior and Emerging Technologies*, **2024**(1), 7115633.

RÖTTGER P., HINCK M., HOFMANN V., HACKENBURG K., PYATKIN V., BRAHMAN F. & HOVY D. (2025). Issuebench : Millions of realistic prompts for measuring issue bias in llm writing assistance. arXiv : 2502.08395.

SHALEVSKA E. & WALKER A. (2025). Are ai models politically neutral ? investigating (potential) ai bias against conservatives. *International Journal of Research Publication and Reviews*, **6**(3), 4627–4637.

WEBER E., RUTINOWSKI J., JOST N. & PAULY M. (2024). Is gpt-4 less politically biased than gpt-3.5 ? a renewed investigation of chatgpt's political biases. arXiv : 2410.21008.