Summary of Changes based on Reviewer comments:

Reviewer o2Kg:

Weaknesses mentioned:

1. The relationship between the three probing components (F-scale, FavScore, and role model analysis) could be articulated more systematically. Currently, they appear somewhat parallel, and a deeper discussion of how they complement or reinforce one another would strengthen the conceptual coherence of the paper.

   Response: We added an explanatory sentence after the research questions in Section 3 Methodology (Line 195) and emphasized the connection in the Conclusion.

   (Note also, that the connection is mentioned in the last paragraph of the Related Work section.)

2. The number of models evaluated is relatively limited, which may constrain the generality of the findings. While this is understandable due to budget constraints, it still slightly weakens the empirical coverage.

   Response: We have updated Section 4.1 to include justifications for our model selection.

   Furthermore, we addressed this point in the response to the reviewer as follows:

   While a broader sample would be ideal, our 8-model selection provides strong empirical coverage. Consistent language-based patterns across this set suggest generalizability. The scale of our study—over 150,000 API calls, including 15,000+ queries per model for FavScore—made it impractical to include more models.

3. The writing quality could be improved; several sections would benefit from clearer exposition to aid reader comprehension.

   Response: We have revised Sections 3.1 and 3.2 to improve clarity and include key methodological details, such as the choice of Likert scales. We have also added a clearer exposition of the soundness of our experimental design for RQ2 (see Appendix E.1) and provided more explanation for our use of the Wasserstein distance in Section 3.2.
   Additionally, we have made a concerted effort to improve the overall writing quality throughout the paper to enhance readability and comprehension.

4. The use of LLM-as-a-judge in the third component introduces potential bias, which remains a limitation even if cost-effective and realistic; this may affect the reliability of the role model classification results.

   Response: We addressed this comment in the author's response for the last submission as follows:

   “To mitigate bias risks in using an LLM-as-judge, we grounded judgments in external evidence: each prompt included regime data from V-Dem tied to the leader’s country and era, with explicit instructions to use that data (see App. C.4.3). A human audit of 100 random classifications confirmed ~93% agreement with historical records, supporting the reliability of our pipeline.”

Comments

1. In Section 3.1, only the mean score across the 30 F-scale statements is reported. Would it be informative to also report the standard deviation to reflect consistency or disagreement in model responses?

   Response: We now report the standard deviation across runs accompanied by a short analysis in Appendix G.

2. Why is a 6-point Likert scale used in Section 3.1, while a 4-point scale is used in Section 3.2? Some justification for this design choice would help clarify the methodology.

   Response: We adopted these choices from the original work and have revised Section 3.1 (Paragraph 1) and Section 3.2 (Paragraph 3) to include a justification.

3. In Section 3.2, I'm not fully clear on the purpose of using the Wasserstein Distance—what exactly is it meant to demonstrate? While I've read Section 5.2, I'm still somewhat confused about the conceptual significance of measuring the distance between the authoritarian and democratic distributions.

   Response: We revised the last paragraph of Section 3.2 to justify the use and purpose of the Wasserstein Distance.

4. Regarding the use of LLM-as-a-judge in Section 3.3, I wonder if models might perform more reliably when evaluating figures from large or globally prominent regimes, but less so for smaller or less represented countries. Could this introduce systematic bias or variance in the classification outcomes?

   Response: We discuss this now in Section 5.3 (Lines 485-489).

5. A key methodological detail: each component involves LLM generation, yet the temperature is fixed at 0. Have the authors considered whether sampling with non-zero temperature might lead to qualitatively different outputs or introduce greater variance in the results?

   Response: T=0 was our choice to favor stable cross-model comparison and noise reduction. Exploring stochastic outputs is a valuable direction for future work.

6. The choice of the term "role model" in Section 3.3 seems to have a strong influence on model behavior. Have the authors considered testing near-synonyms to evaluate whether the wording itself affects the type of names returned?

   Response: To test the influence of "role model," we ran a follow-up experiment with 5 prompt variants across 20 countries, which we report in App. B.3 and refer to in Subsection 4.2.

7. From what I understand, the current generation setup involves first selecting an answer and then providing a rationale—i.e., the model makes a decision and justifies it afterward. This is one valid prompting strategy, but other common approaches include generating the rationale first (to encourage chain-of-thought reasoning before committing) or omitting the rationale entirely. Have the authors considered whether these alternatives might lead to different outputs or affect the measured biases?

   Response: This is a misunderstanding, as the prompts actually place the reasoning field before the chosen answer. Because of a limited budget, we chose this approach, as it encourages CoT generation in line with best practice (https://arxiv.org/abs/2201.11903).

Reviewer 3gZ5:

Weaknesses mentioned:

1. The lacks a comprehensive discussion in the related works section of what aspect of political values and/or bias is explored in existing research. There are many papers related to political debate, persona, etc. of LLMs, and the paper also mentions in its abstract that "prior work has primarily examined socio-demographic and left–right political dimensions". It is important to survey them, and state how this paper distinguishes from them and why it is a need.

   Response: We rewrote Section 2 Paragraph “Measuring (Political) Bias in LLMs”, surveying prior work. Furthermore, we explain in the Introduction (Lines 56-93) how our work is novel and distinct from prior work.

2. It is unclear whether the experiment setup for RQ2 is sound. It involves eliciting answers from LLMs for questions such as "Do you approve or disapprove of {leader}’s handling of the economy?", "Do you think {leader} can get things done?", etc. Some of the questions are related to the "effectiveness" of a leader. it is still possible that non-elective leaders in an authoritarian society launch economic policies that suit the needs of nation. Agreeing that this leader is effective does not necessary indicates that the models has good feelings about the authoritarian system. If LLM states that the authoritarian leader X is successful in boosting economy (which is supported by data), is it a fact or opinion?

   Response: Our questions are grounded in established surveys. We reviewed over 300 items and carefully selected a subset that is as opinion-based and framing-neutral as possible. In our previous version, we already included an explanation of the steps taken to make the setup sound (Section 3.2). We have now included a subsection in App. E.1 with further explanations and clear examples to which we refer from Section 3.2.

Comments:

1. The papers conducts experiments on models developed by companies from different countries. Is model's political value correlate with its country of origin? If so, how this trend may be related to model's training data?

   Response: Investigating the influence of model origin was beyond the scope of this study, and our sample size is too limited to support statistically meaningful conclusions in that regard.

   We addressed this point further in the answer to the review:

   “Our results represent an initial suggestion that a model's country of origin is a far weaker predictor of its political alignment than the language of the prompt. We found that both US-developed models (e.g., Llama 4 Maverick) and China-developed models (e.g., DeepSeek V3) can exhibit a similar, significant shift toward favoring authoritarian figures when prompted in Mandarin.“

Reviewer EKTN:

Weaknesses mentioned:

1. It is not clear how the prompts have been tested as to whether they are stable. Have the authors, for example, reworded the prompts and tested whether the results are stable? See, e.g., https://arxiv.org/abs/2505.13546 and https://aclanthology.org/2022.findings-emnlp.445.pdf

   Response: We ran a follow-up paraphrasing experiment with 5 prompt variants across 20 countries, which we report in App. B.3 and refer to in Subsection 4.2. Furthermore, we carefully formulated prompts to minimize framing effects and acquiescence bias (App. B and C).

2. For the experiments, to make final statements, the paper needs statistical significant difference tests (p-value) to show whether the difference between languages is actually (statistically) significantly different and whether this trend can be observed across all experiments (i.e., across the three research questions)

   Response: Significance tests for all three experiments are now included. See Section 5.1, Table 4, Appendix G, H, I.

3. For the test of role models, it would be interesting to have a discussion of whether the number of people are more limited in certain regions of the world. I.e., as we know the LLMs have a Western bias, see, e.g., https://aclanthology.org/2023.c3nlp-1.12/ and https://arxiv.org/abs/2203.07785, could it be that for countries in the west there are more popular figures to pick from for the model? How does these factors play into the answers of the model and how do they interplay/correlate with the results of the other experiments? In other words, if the model proposes Gaddafi as a Libyan role model, how many other Libyan known figures are known to the LLMs?

   Response: We discuss this now in Section 5.3 (Lines 485-489).

   Furthermore, here is our answer to the review:

   “We agree that sparse coverage of non-Western public figures is a confounder: if an LLM “knows” only one internationally salient Libyan, it will inevitably offer Gaddafi. Yet this scarcity can be part of the bias we investigate. A helpful system asked for a “role model” should either retrieve a genuinely praiseworthy national figure, even if less globally famous, or admit there are no solid options. The fact that models routinely present long-standing autocrats as exemplary citizens reveals a structural knowledge imbalance coupled with a missing normative filter, both of which can reinforce authoritarian narratives.”

4. The authors should extend the difference between “democracy–authoritarianism” vs “left-right” political bias, the latter seems to be a subset of the former and left-right political bias is a well-discussed topic, see, e.g., [1] More human than human: measuring ChatGPT political bias https://link.springer.com/article/10.1007/s11127-023-01097-2 [2] The political preferences of LLMs https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0306621 [3] The political ideology of conversational AI: Converging evidence on ChatGPT’s pro-environmental, left-libertarian orientation https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4316084 [4] More human than human: measuring ChatGPT political bias https://link.springer.com/article/10.1007/s11127-023-01097-2 [5] On the Relationship between Truth and Political Bias in Language Models https://arxiv.org/html/2409.05283 And even as a policy topic: [6]The Politics of AI: An Evaluation of Political Preferences in Large Language Models from a European Perspective https://cps.org.uk/wp-content/uploads/2024/10/CPS_THE_POLITICS_OF_AI-1.pdf

   Response: We rewrote Section 2 Paragraph “Measuring (Political) Bias in LLMs”, surveying prior work. Furthermore, we explain in the Introduction (Lines 56-93) how our work is distinct from prior work.

5. More broadly, the paper lacks in the theoretical framework. As it is grounded in political sciences, it would be great to describe in greater detail the motivation for the paper, how the research questions were chosen, and further extend the chosen languages.

Response: We have added more explanations of theoretical grounding, such as the choice of Likert scales (Section 3.1 and Section 3.2), an additional explanation of the soundness of our experiment for RQ2 (App. E.1) and a justification of the use of the Wasserstein distance (Section 3.2). We further explained the choice and connection between research questions in more detail (Section 3 Methodology).

6. Given the lack of the theoretical grounding, the novelty of the paper is limited. As the methodology is not novel, nor introduces novel aspects in the experimental section or dataset creation, it lacks substance to justify being a full paper. The most important aspects, the comparison across languages is limited to only two languages, although widely used, it would have been interesting to extend the study to larger set of languages.

Response: We respectfully note that our study is the first to systematically explore LLM bias along the democratic-authoritarian axis, representing a novel contribution that addresses a gap in existing literature. The critique that our work lacks novelty stems from a fundamental misunderstanding of political science, claiming the left-right axis is a subset of the democracy-authoritarian axis (see weakness 4). These are, in fact, orthogonal concepts. Our core contribution is precisely to introduce and systematically evaluate this unexplored procedural dimension of bias, for which we developed novel instruments like FavScore.

Comments:

1. A 6-point/4-point Likert scale seems to be an odd choice. How did you decide on this? It is typically a 5 or 7-point scale.

Response: We adopted these choices from the original work and have revised Section 3.1 (Paragraph 1) and Section 3.2 (Paragraph 3) to include a justification.

2. Please label the bars in Figure 2 so it is possible to compare the bars based on the numerical values. E.g.,iIs the value for English for Llama and Grok the same?

Response: We omit individual bar labels to avoid visual clutter, as the scale is shown on the axis and full results are provided in App. G.

Meta-Reviewer:

1. Provide a justification of how well the prompts for RQ2 are suited to the aims, and potentially reframe the RQ to be more specific about the dimensions of "evaluation" and why they are important to the overall goals.

   Response: Our questions are grounded in established surveys. We reviewed over 300 items and carefully selected a subset that is as opinion-based and framing-neutral as possible. In our previous version, we already included an explanation of the steps taken to make the setup sound (Section 3.2). We have now included a subsection in App. E.1 with further explanations and clear examples to which we refer from Section 3.2. We further now justify the research questions in Section 3 Methodology.

2. Incorporate the discussion from the rebuttal period of how F-scale, FavScore and role model analysis complement or reinforce one another, and strengthen their independent contributions

   Response: We added an explanatory sentence after the research questions in Section 3 Methodology (Line 195) and emphasized the connection in the Conclusion.

   (Note also, that the connection is mentioned in the last paragraph of the Related Work section.)

3. Justify the selection of LLMs for this experiment

   Response: We have updated Section 4.1 to include justifications for our model selection.

4. Strengthen the discussion of related work by adding prior work on bias issues that are indirectly related to the paper's concerns (e.g., left-right political bias; sociodemographic bias) and in key parts of the paper emphasize how the present work fits into the overall enterprise of investigation of political bias in LLMs

   Response: We rewrote Section 2 Paragraph “Measuring (Political) Bias in LLMs”, surveying prior work

5. Address the issue of potential bias introduced by using LLM-as-a-judge, for example, when applied to globally prominent countries versus small, less prominent countries, again, incorporating the additional data presented in the rebuttal period

   Response: We have made our LLM-as-a-Judge setup explanation clearer section 3.3 and Appendix C. We have also added all additional experimental data from the rebuttal period into the paper.

6. Justify the two different scales used in different parts of the experiment

   Response: We adopted these choices from the original work and have revised Section 3.1 (Paragraph 1) and Section 3.2 (Paragraph 3) to include a justification.

7. Motivate the use of Wasserstein distance

   Response: We revised the last paragraph of Section 3.2 to justify the use and purpose of the Wasserstein Distance.

8. Report the variance in results resulting from synonym substitution in prompts

   Response: Reported the ablation experiment in App. B.3