Large Language Models Are Democracy Coders with Attitudes

Nils B. Weidmann, *University of Konstanz, Germany*
Mats Faulborn, *University of Konstanz, Germany*
David García, *University of Konstanz, Germany*

ABSTRACT Current political developments worldwide illustrate that research on democratic backsliding is as important as ever. A recent exchange in *Political Science & Politics* (February 2024) highlighted again that the measurement of democracy remains a challenge. With many democracy indicators consisting of subjective assessments rather than factual observations, trends in democracy over time could be due to human biases in the coding of these indicators rather than empirical facts. This article leverages two cutting-edge Large Language Models (LLMs) for the coding of democracy indicators from the V-Dem project. With access to huge amounts of information, these models may be able to rate the many "soft" characteristics of regimes at substantially lower costs. Whereas LLM-generated codings largely align with expert coders for many countries, we show that when these models deviate from human assessments, they do so in different but consistent ways. Some LLMs are too pessimistic and others consistently overestimate the democratic quality of these countries. Although the combination of the two LLM codings can alleviate this concern, we conclude that it is difficult to replace human coders with LLMs because the extent and direction of these attitudes is not known *a priori*.

The measurement of democracy has long been a contested subject of investigation in political science. A recent symposium in *PS: Political Science & Politics* (volume 57, issue 2) addressed this question by discussing whether the observation of a global trend of democratic backsliding could be due to subjective perceptions of human coders. Focusing on V-Dem, the largest collection of democracy data, the argument is that expert ratings in this dataset could be affected by psychological biases (Little and Meng 2024; Treisman 2024), leading to skewed assessments of democratic quality that are not supported by more factual observations of institutional characteristics (see, however, Knutsen et al. 2024 for a critique).

Rather than dismissing human coding in general, this article examines whether and how it could be supplemented with automatic coding. For political science, the advent of Artificial Intelligence (AI) and, in particular, Large Language Models (LLMs) has provided many new opportunities. The most promising and most frequent way in which these models are used is the generation of new data for empirical research, assisting (and sometimes even replacing) humans as creators or sources of these data. Much work in this area has shown that LLMs can reproduce voting decisions or responses collected in surveys (Mellon et al. 2024), thereby reducing the need to obtain large population samples. This article examines the use of LLMs for human coding in comparative research on democracy, which is a second way that AI can assist in the creation of typically human-sourced data for social science research. Although LLMs may be able to avoid some

Corresponding author: Nils B. Weidmann *is professor of political science at the University of Konstanz. He can be reached at nils.weidmann@uni-konstanz.de.*
Mats Faulborn *is a data scientist at scieneers GmbH. He can be reached at mats.faulborn@gmx.net.*
David García *is professor of social and behavioral data science at the University of Konstanz. He can be reached at david.garcia@uni-konstanz.de.*

© The Author(s), 2025. Published by Cambridge University Press on behalf of American Political Science Association. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

doi:10.1017/S1049096525101248

https://doi.org/10.1017/S1049096525101248 Published online by Cambridge University Press

of the biases that human coders typically display, this approach also could reduce dramatically the costs of coding. The exercise presented herein (i.e., the coding of one year's worth of indicators from the V-Dem dataset) was completed with a budget of less than €150; employing human coders for the same task likely would cost several tens of thousands of euros. However, the key question is whether LLMs produce codings of sufficient quality.

We show, perhaps not surprisingly, that LLMs can emulate human coders well and that they can do so "off the shelf," without any adaptation. However, we also show that LLMs are of little assistance with those countries that V-Dem coders find particularly difficult and where they disagree the most. It is even more concerning that although LLMs may not exhibit the cognitive biases typically attributed to humans, they have other extremely pronounced issues. Results show that one LLM in our study consistently underestimated democratic quality and another overestimated it in almost all cases. Therefore, LLMs apparently have particular political "attitudes" that strongly affect their coding. Because we do not know the direction and strength of these attitudes, it is difficult to trust assessments that are generated by particular LLMs alone or in combination.

...democracy coding is a prime example of what we consider to be a "difficult" human coding task.

CODING DEMOCRACY IS A DIFFICULT TASK

How does democracy coding compare to other human coding tasks, and why could LLMs be particularly helpful in this task? We start with a general categorization. In human coding, coders (typically experts) are tasked with the creation of standardized codings for particular instances and cases. For a simple categorization of human coding tasks, we can distinguish by (1) the extent to which the material required for the coding is provided as part of the task; and (2) the extent of interpretation that is required by the coder to perform the task. The first dimension refers to whether the coding is based on material that is readily provided to the coder. Some coding tasks involve particular instances of coding material with which the coders then work. For example, annotations of social media posts (Gilardi, Alizadeh, and Kubli 2023) and the coding of protest events from news reports (Weidmann and Geelmuyden Rød 2019, ch. 4) provide coders with material on which the coding is supposed to be based. Other coding tasks, such as the human coding of democracy indicators (Coppedge et al. 2024) and the fact checking of particular statements (Ni et al. 2024), often do not provide any material, and it is assumed that coders possess expert knowledge to perform the task.

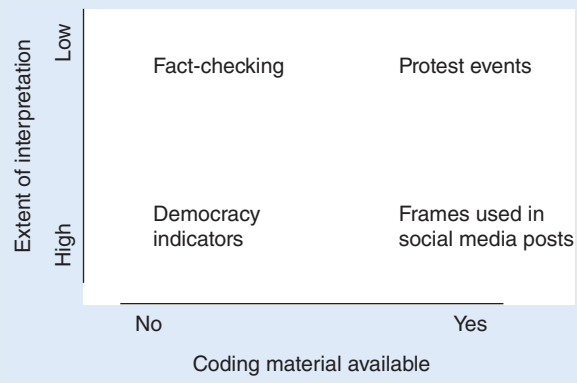
The second dimension by which we can distinguish different coding tasks is the extent of interpretation required by the coder. By "interpretation," we mean the process by which coders leverage their own expertise and intuition to complete a given task. Interpretation concerns several parts of the coding task. The particular question associated with the task must be interpreted —for example, the actors, institutions, and outcomes mentioned therein. Moreover, answer categories require interpretation before they can be applied. Some tasks involve very little interpretation. For example, coding the time and place of a protest event from a news report can be accomplished by identifying the corresponding information from the report, without requiring much interpretation by the coder (Weidmann and Geelmuyden Rød 2015). This is similar for fact-checking tasks (Ni et al. 2024), for which coders only need to establish whether a statement is factually true. In other tasks, such as the identification of a particular frame in social media posts, coders must interpret the information provided to determine whether it aligns with the particular label (Gilardi, Alizadeh, and Kubli 2023). Similarly, for coding many of the democracy indicators in the V-Dem project, coders must use their own expertise to determine whether, for example, opposition parties are "independent and autonomous of the ruling regime" (i.e., V-Dem indicator *v2psoppaut*) (Coppedge et al. 2024, 100). Therefore, to reduce the extent of interpretation by coders, V-Dem provides extensive clarifications on the background of each coding question and attempts to set anchoring points for the different answer categories. The two-dimensional categorization and examples for the different types of coding tasks are depicted in figure 1.

The difficulty of a coding task typically is higher if the coding material is not prespecified, which also is why these coding tasks should have a lower reliability. Tasks that require more interpretation typically are more subjective and concerns about reliability are higher. This means that coding difficulty is highest for tasks in the bottom left corner of figure 1. Democracy coding is one example of this type of task: it is virtually impossible to precisely specify which material coders should consult to rate political regimes. At the same time, many of the indicators used in this type of coding require extensive interpretation by coders. In summary, democracy coding is a prime example of what we consider to be a "difficult" human coding task.

Therefore, it would be important to analyze whether LLMs can help us to solve this difficult task. We already know that modern

Figure 1

Different Types of Human Coding Tasks



https://doi.org/10.1017/S1049096525101248 Published online by Cambridge University Press

AI models perform well (in some cases, even extremely well) for other examples of coding tasks. Overos et al. (2024) show that the coding of protest events from news reports works well and that LLMs are able to match the performance of humans on this task. Annotations of social media posts with labels indicating the political stance or a particular framing are more difficult; however, studies have shown that modern LLMs can perform this task with high accuracy compared to human annotators (Gilardi, Alizadeh, and Kubli 2023; Le Mens and Gallego 2025). Furthermore, LLMs have been employed successfully in fact checking, as illustrated in Ni et al. (2024).

In the following discussion, we replicated the difficult task of democracy coding with LLMs. Trained on a large amount of data, these models may possess a knowledge base that should be on par with human experts, thereby addressing the lack of prespecified coding material that we typically experience in these tasks. The human biases that can affect human interpretations in these coding tasks also should be reduced when using LLMs. Our aim was not to custom-tailor these models specifically for the task of rating political regimes; rather, we wanted to see how these models compare without prior adaptation in a so-called zero-shot setting. In addition, we were particularly interested in how LLMs perform in cases that are difficult for human coders because, for example, the available information about a particular country is limited.

CODING V-DEM WITH LLMS

Our analysis is based on the V-Dem dataset of political indicators that covers all countries worldwide with annual observations (Coppedge et al. 2024). We used the 2024 release (Version 14), which was published after the cutoff dates for the training data for the two LLMs in our study; it therefore rules out contamination of the LLMs with actual data. From Version 14, we replicated the coding of V-Dem variables for 2023, the most recent year in the data.

V-Dem's well-known aggregate democracy indices are based on a large set of constituent indicators produced by expert coders. The data for these indicators are collected in a series of survey questions that coders answer for the respective country and year, with responses typically recorded on an ordinal scale from "bad" (illiberal, authoritarian) to "good" (liberal, democratic). For example, to score the extent of media bias, experts answer the question, "Is there media bias against opposition parties or candidates?," on a 5-point scale from 0 (i.e., "The print and broadcast media cover only the official party or candidates, or have no political coverage, or there are no opposition parties or candidates to cover") to 4 (i.e., "The print and broadcast media cover all newsworthy parties and candidates more or less impartially and in proportion to their newsworthiness").

We selected all of the ordinal indicators coded by country experts (called Type-C variables in V-Dem) that are necessary for the coding of the high-level V-Dem democracy indicators (i.e., electoral, liberal, participatory, deliberative, and egalitarian). It is important to note that this excludes factual questions such as the first year of universal suffrage in a country. We are not interested in the LLM's ability to retrieve facts but rather the often "fuzzy" and subjective assessments that it provides and that constitute the main source of information on which V-Dem's democracy indices are based (see Marquardt et al. 2024 for a similar aim). Because we focused on only a single year (2023), our dataset included 53 indicators for 171 countries (see online appendix A for a full list). For each of these indicators, we used the final values provided in the V-Dem dataset, which were computed across the different coders that provided ratings for this particular indicator and country (Weidmann, Faulborn, and Garcia 2025).

Our prompts for the LLM were intentionally kept simple and used the exact wording of the questions and the responses provided in the V-Dem codebook. We only added a short introduction that provided the context (i.e., "You are an assistant who evaluates political systems in different countries and years. You will be asked to produce numeric scores derived from your knowledge of this country"; see details in online appendix B). Our experiment included two state-of-the-art LLMs: Llama-3.1 70B and GPT-4o (checkpoint GPT-4o-2024-08-06). We used GPT-4o because it currently is the main model of ChatGPT and because it is one of the best performing in the Chatbot Arena benchmark (Chiang et al. 2024). We complemented this with Llama-3.1 as one of the best-performing open-weights models at the time of this research; we chose the 70B size because the additional performance of the 405B version was not sufficient to justify the extra energy and hardware requirements (Grattafiori et al. 2024). We did no fine tuning and operated in a zero-shot setting, where we interacted with the models without providing any examples or other training material. To exclude sequence effects, for each question, we randomized the order in which the model coded the different countries.

Overall, coding the V-Dem indicators with LLMs worked well, despite the fact that we did not attempt to adjust or tune the models in any way. None of the models ever gave a response that was outside the range of the respective indicator—for example, by returning a score of 4 for an indicator with a range of 0–3. Also, the models refused to provide answers in only relatively few cases. LLama-3.1 failed to answer 129 of the total of 9,063 questions (i.e., 171 countries, 53 indicators), which corresponds to approximately 1.4%. GPT-4o refused to answer in only five cases (0.06%).

RESULTS

For our analysis, we used data from the main V-Dem release. To address different issues and biases that arise in the human coding process, V-Dem employs a sophisticated measurement model based on item-response theory (Marquardt 2020). This approach is designed to correct biases at the coder level, which could arise, for example, because some coders may be more critical and others are more lenient in their perception of a country. The model also is able to correct differential item functioning, which is the fact that different coders have different interpretations of the coding scales used for particular variables. The V-Dem measurement-model approach is described in detail in Pemstein et al. (2020).

V-Dem publishes a number of different transformations of the measurement model results for each indicator. Those that are relevant for our analysis were (1) the ordinal transformations (with suffix *ord*), which are the results transformed back to the original ordinal scale; and (2) the upper and lower bounds of the posterior distribution of the results (with suffixes *osp codelow* and *osp codehigh*). The former corresponds to an aggregated and corrected version of each indicator across all V-Dem coders and the latter helps us to gauge the level of disagreement among the coders. If upper and lower bounds are far apart, this means that coders had a higher level of disagreement even after coder-level

https://doi.org/10.1017/S1049096525101248 Published online by Cambridge University Press

and question-level biases were corrected. Thus, we used the distance between the upper and lower bounds as a measure of disagreement among coders.

To analyze patterns in how LLM codings compare to human coders, we aggregated the results by country. In other words, we used all of the 53 indicators for each country and computed how well the LLM codings correlated with the (aggregated) human ratings in the final V-Dem data and what their average deviation from these ratings was. The overall correlation between LLMs and human codings was positive and modestly strong. Without any adaptation for this particular coding task, Llama-3.1's ratings, on average, had a correlation of 0.5 with the V-Dem scores (ranging up to 0.81); GPT-4o even achieved an average correlation of 0.64 (up to 0.88).

However, simply examining the correlation between the LLM ratings and V-Dem's expert assessments is misleading. Although the correlation coefficients show whether human coders and LLMs identify similar tendencies, they cannot tell us whether they agree on the absolute rating of democratic qualities in different countries. After all, it is important to know where countries stand regarding their democratic quality in absolute terms—for example, whether members of the executive embezzle public funds only "occasionally" or "often" (i.e., V-Dem's *v2exembez* indicator). Therefore, the following discussion analyzes the difference between expert and LLM codings—that is, where LLMs place particular countries relative to human experts on the ordinal coding scale. Although this measure resembles what typically is called an "error," we refer to it as a "difference" because the true value of the outcome is not known.

Figure 2 plots the distribution of the mean difference per country, computed over all V-Dem questions. In these plots, negative values on the x-axis are those for which the model produced lower scores than the human coders; positive values correspond to the opposite. The blue line at 0 indicates overall correspondence between LLM and human coders. The histograms clearly show that the two LLMs in our experiment possess different "attitudes" toward the political situation in a country. Whereas GPT-4o largely provides lower ratings (on average by 0.28) than the experts for many countries (and therefore is a more pessimistic observer; see the left panel in figure 2), Llama-3.1 does the opposite and considers the political situation to be more liberal/democratic (on average by 0.5) than the experts and therefore is an optimistic observer (see the right panel in figure 2).

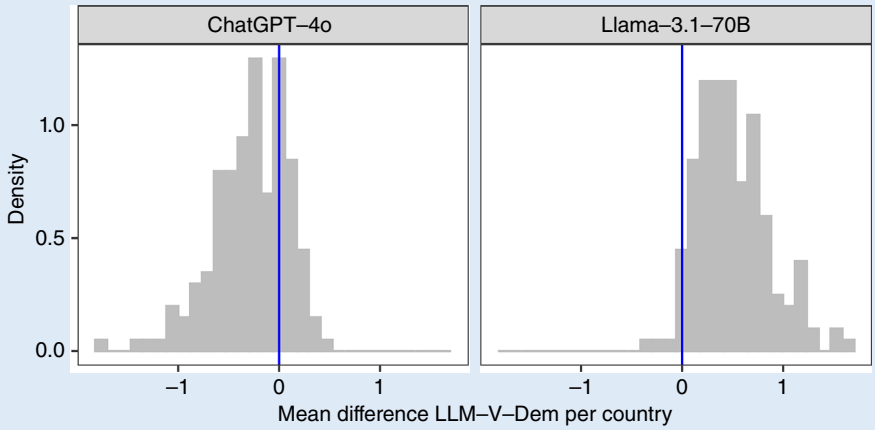
What is striking in figure 2 is not only the fact that the direction of how the LLMs deviate from human coders is different for the two models; it also is the consistency of this deviation across countries. The pessimistic LLM, GPT-4o, underestimated in approximately 80% of all countries, whereas the optimistic Llama-3.1 model overestimated in almost all of them (97%). In other words, when these models deviate from the expert assessments, they almost always deviate in a particular direction: GPT-4o gives a more cautious assessment whereas Llama-3.1 overestimates the liberal/democratic quality of a country.

...the two LLMs in our experiment possess different "attitudes": whereas GPT-4o largely provides lower ratings, Llama-3.1 does the opposite and considers the political situation to be more liberal/democratic.

We then analyzed for which countries the models were particularly likely to differ from the expert assessments. We were especially interested in how the LLMs fared in countries that the V-Dem coders found more difficult to code. To this end, we plotted the mean error per country against the level of coder disagreement, as measured by the average standard deviation among the different coders (i.e., the measure introduced previously). Figure 3 (see the left and center panels) shows the results for both LLMs individually. As shown in the plots, both models

Figure 2

Distribution of the Average Difference Across All V-Dem Questions per Country

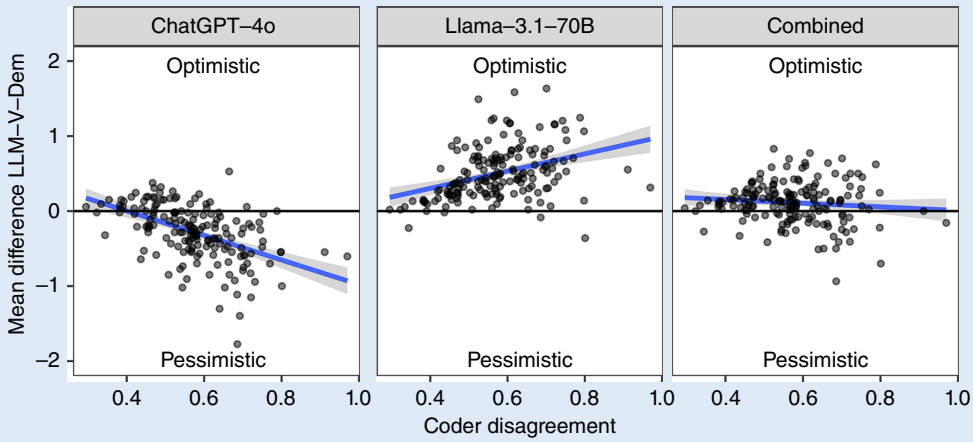


A difference of 0 (blue line) indicates perfect correspondence with the expert coders.

https://doi.org/10.1017/S1049096525101248 Published online by Cambridge University Press

Figure 3

Average Differences Between LLMs Ratings and Human Codings at the Level of Countries



Countries are ordered by difficulty of coding (i.e., disagreement among coders).

matched human codings almost perfectly for "easy" countries where coders agreed (see the left side of the x-axis). However, humans and LLMs differ more for those countries that the coders found difficult to code (see the right side of the x-axis). In this case, LLMs exhibited the political attitudes identified previously. Whereas GPT-4o rated these countries as less democratic than the expert coders, Llama-3.1 believed them to be more democratic.

Can we combine both LLMs to generate a better fit with human codings? The first two panels in figure 3 suggest that the political attitudes of the two LLMs apparently cancel out one another. When we combined their scores by taking the average (see figure 3, right panel), we could see that the resulting scores provided a much better match with human codings. The average difference between LLMs and human coders was now at a value of approximately 0.1, which was constant across the set of countries (i.e., the blue line). In other words, combining LLMs with different political attitudes can help to reduce differences between LLMs and human codings, especially for those countries that constitute more difficult cases. Online appendix C demonstrates that this pattern also holds for "unstable" cases—that is, those indicators and countries that changed from the previous year (2022) to the year that we examined (2023).

To determine where and how the LLMs differ from human codings, online appendix D shows where selected countries rank in terms of coder disagreement and LLM performance. It is also informative to study in more detail where LLMs err, if they do. To illustrate this, we chose two countries where LLMs perform poorly and show the codings that the LLMs produced for a subset of the V-Dem indicators. For the plot in figure 4, we selected eight indicators according to the level of observability (i.e., an ad hoc assessment of the authors). The four left indicators are based on information that should be relatively easy to obtain also for an LLM, such as whether particular political parties are banned. The four right indicators are more difficult to observe—for example, whether political power is distributed by socioeconomic position is unlikely to be discussed explicitly in the textual sources on which LLMs are trained.

Figure 4 plots the deviations of the LLM ratings from the human codings using the familiar metric from the previous plots: that is, values above zero indicate more optimistic assessments and values below zero indicate pessimism. The results for El Salvador (left panel) show again the result of figure 3: ChatGPT-4o approximated the human codings better than Llama-3.1, as indicated by the shorter lines in the top panel. For the latter LLM, the deviation from the human codings was not due to indicators that are more difficult to observe; rather, we also found larger deviations for the more easily observable indicators on the left. For Niger (right panel), the picture was different; Llama-3.1 showed very good performance (shorter lines) and ChatGPT-4o produced ratings that were too pessimistic. Again, there was no difference among indicators that were more easily observable and those that were not—ChatGPT-4o was too pessimistic across the board.

Our analysis leads to two important conclusions. First, LLMs can have particular and pronounced biases in how they assess "soft" political characteristics of regimes. We show that vis-à-vis human coders, certain LLMs are overly pessimistic observers of democratic quality and others are consistently optimistic. In other words, whereas some err on the side of caution when assessing democracy, others provide exaggerated ratings. For democracy scholars who are interested in absolute rather than relative country ratings, this is reason for concern. Clearly, single LLMs should not be used uncritically to produce democracy ratings because they may be affected by the respective model's attitude. Combining different models can be more useful, but this requires a prior assessment of the strength and direction of their attitudes.

CONCLUSION

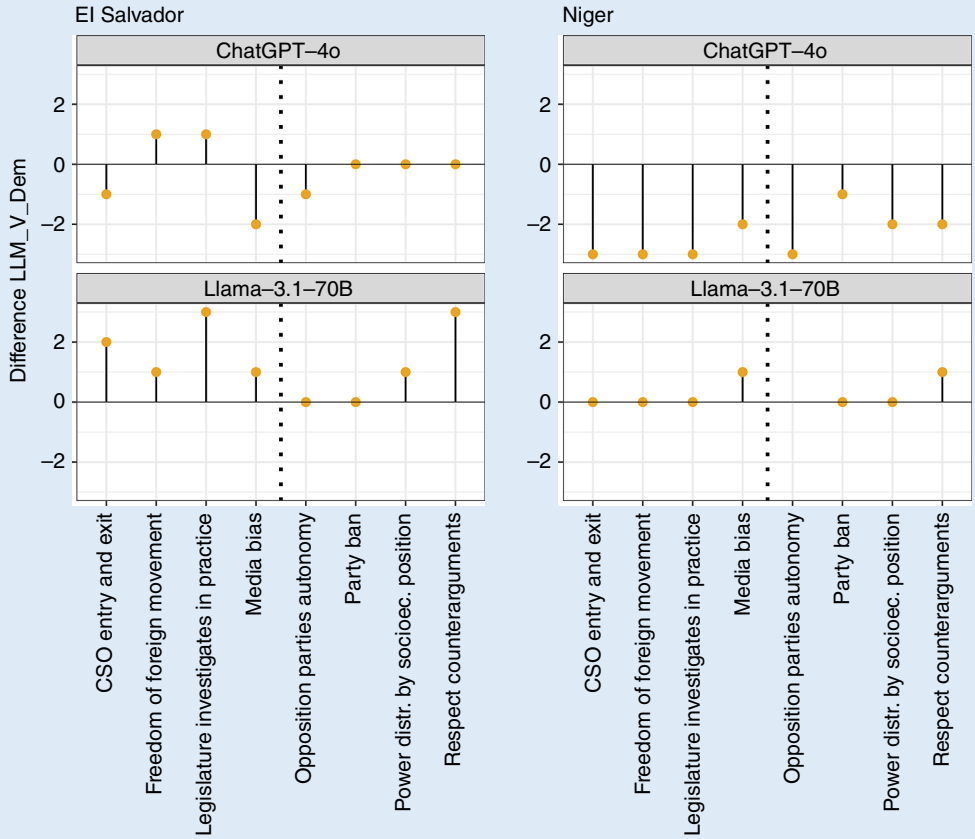
This article examines the use of LLMs for the generation of democracy scores. Coding democracy is a difficult and expensive task because the source material often is unspecified and the coding requires a high degree of interpretation by humans. This is why LLMs, with their large knowledge base and their automated reasoning, could be valuable assistants for these tasks.

To determine whether this is the case, we replicated the coding of the well-known V-Dem democracy indicators with two of the current cutting-edge LLMs: GPT-4o and Llama-3.1. These models require no adaptation for this task, and the automated coding is simple and can be done with a few lines of code. Results show that

https://doi.org/10.1017/S1049096525101248 Published online by Cambridge University Press

Figure 4

Difference Between LLM and Human Codings for Two Countries and Selected V-Dem Indicators



LLMs approximate human coding well. However, the models also struggle with countries that human coders find particularly difficult and where they disagree the most. For these countries, one of our models consistently underestimates democratic quality whereas the other model almost always overestimates it. In short, LLMs apparently have particular political attitudes—some can be pessimistic and others are overly optimistic about a political situation. A combination of a pessimistic model and an optimistic model, however, produces a much closer match between human and LLM scores.

Second, using combinations of different LLMs as political observers should work much better. Although ensembles of LLMs have been shown to perform well for tasks such as forecasting (Schoenegger et al. 2024), using them for the coding of democracy scores requires us to measure their political attitude beforehand so that they ideally complement one another to produce a more balanced outcome. Our results suggest that even a simple method of averaging codings from different LLMs improves results considerably. It remains to be seen whether even more pronounced gains could be made using a more sophisticated measurement-

...political ratings produced by a single model are unlikely to be sufficient.

These results lead to two main conclusions. First, political ratings produced by a single model are unlikely to be sufficient. We often do not have any indication whether a model has a particular political attitude and in which direction. This echoes the findings from other research that attempted to outsource expert coding to crowd coders and found that it does not work for tasks other than the simplest ones (Marquardt et al. 2024).

model approach, such as the one used in the V-Dem project (Marquardt and Pemstein 2018). This may be even easier to implement for human coders because LLMs typically produce scores for all of the countries included in a sample, not only the subset for which they have expertise. Furthermore, we may be able to improve LLM coding with custom-designed LLMs for this purpose—for example, by fine tuning. Nevertheless, a careful

https://doi.org/10.1017/S1049096525101248 Published online by Cambridge University Press

comparison with human coders such as the one presented in this article will remain essential as well as for ensemble approaches and custom-tuned models.

SUPPLEMENTARY MATERIAL

To view supplementary material for this article, please visit http://doi.org/10.1017/S1049096525101248.

DATA AVAILABILITY STATEMENT

Research documentation and data that support the findings of this study are openly available at the *PS: Political Science & Politics* Harvard Dataverse at https://doi.org/10.7910/DVN/I34X6P.

CONFLICTS OF INTEREST

The authors declare that there are no ethical issues or conflicts of interest in this research. ■

REFERENCES

Chiang, Wei-Lin, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. "Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference." Preprint. https://arxiv.org/abs/2403.04132.

Coppedge, Michael, et al. 2024. *V-Dem Dataset v.14 Codebook*. https://v-dem.net.

Gilardi, Fabrizio, Meysam Alizadeh, and Maël Kubli. 2023. "ChatGPT outperforms crowd workers for text-annotation tasks." *Proceedings of the National Academy of Sciences* 120 (30): e2305016120.

Grattafiori, Aaron, et al. 2024. "The Llama 3 Herd of Models." Preprint. https://arxiv.org/abs/2407.21783.

Knutsen, Carl Henrik, Kyle L. Marquardt, Brigitte Seim, Michael Coppedge, Amanda B. Edgell, Juraj Medzihorsky, Daniel Pemstein, Jan Teorell, John Gerring, and Staffan I. Lindberg. 2024. "Conceptual and Measurement Issues in Assessing Democratic Backsliding." *PS: Political Science & Politics* 57 (2): 162–77.

Le Mens, Gaël, and Aina Gallego. 2025. "Positioning Political Texts with Large Language Models by Asking and Averaging." *Political Analysis* 33 (3): 274–82.

Little, Andrew T., and Anne Meng. 2024. "Measuring Democratic Backsliding." *PS: Political Science & Politics* 57 (2): 149–61.

Marquardt, Kyle L. 2020. "How and How Much Does Expert Error Matter? Implications for Quantitative Peace Research." *Journal of Peace Research* 57 (6): 692–700.

Marquardt, Kyle L., and Daniel Pemstein. 2018. "IRT Models for Expert-Coded Panel Data." *Political Analysis* 26 (4): 431–56.

Marquardt, Kyle L., Daniel Pemstein, Constanza Sanhueza Petrarca, Brigitte Seim, Steven Lloyd Wilson, Michael Bernhard, Michael Coppedge, and Staffan I. Lindberg. 2024. "Experts, Coders and Crowds: An Analysis of Substitutability." *International Political Science Review*. https://doi.org/10.1177/01925121241293459.

Mellon, Jonathan, Jack Bailey, Ralph Scott, James Breckwoldt, Marta Miori, and Phillip Schmedeman. 2024. "Do AIs Know What the Most Important Issue Is? Using Language Models to Code Open-Text Social Survey Responses at Scale." *Research & Politics* 11 (1): 20531680241231468.

Ni, Jingwei, Minjing Shi, Dominik Stammbach, Mrinmaya Sachan, Elliott Ash, and Markus Leippold. 2024. "AFaCTA: Assisting the Annotation of Factual Claim Detection with Reliable LLM Annotators." In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, ed. Lun-Wei Ku, Andre Martins, and Vivek Srikumar, 1890–912. Bangkok, Thailand.

Overos, Henry David, Roman Hlatky, Ojashwi Pathak, Harriet Goers, Jordan Gouws-Dewar, Katy Smith, Keith Padraic Chew, Johanna K. Birnir, and Amy H. Liu. 2024. "Coding with the Machines: Machine-Assisted Coding of Rare Event Data." *PNAS Nexus* 3 (5): 165.

Pemstein, Daniel, Kyle M. Marquardt, Eitan Tzelgov, Yi-ting Wang, Juraj Medzihorsky, Joshua Krusell, Farhad Miri, and Johannes von Römer. 2020. "The V-Dem Measurement Model: Latent Variable Analysis for Cross-National and Cross-Temporal Expert-Coded Data." V-Dem Working Paper 21, 5th edition. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3595962.

Schoenegger, Philipp, Indre Tuminauskaite, Peter S. Park, Rafael Valdece Sousa Bastos, and Philip E. Tetlock. 2024. "Wisdom of the Silicon Crowd: LLM Ensemble Prediction Capabilities Rival Human Crowd Accuracy." *Science Advances* 10 (45): 1528.

Treisman, Daniel. 2024. "Psychological Biases and Democratic Anxiety: A Comment on Little and Meng (2023)." *PS: Political Science & Politics* 57 (2): 194–97.

Weidmann, Nils B., and Espen Geelmuyden Rød. 2015. "Making Uncertainty Explicit: Separating Reports and Events in the Coding of Violence and Contention." *Journal of Peace Research* 52 (1): 125–28.

Weidmann, Nils B., and Espen Geelmuyden Rød. 2019. *The Internet and Political Protest in Autocracies*. New York: Oxford University Press.

Weidmann, Nils B., Mats Faulborn, and David Garcia. 2025. "Replication Data for 'Large Language Models Are Democracy Coders with Attitudes.'" *PS: Political Science & Politics*. DOI:10.7910/DVN/I34X6P.

https://doi.org/10.1017/S1049096525101248 Published online by Cambridge University Press