From Text to Events: Turning Freedom House Reports into Evidence of Democratization and Democratic Backsliding*

Matthew Wilson, Kelsey Martin-Morales, and Gregory Nelson

March 2026

Abstract

Political scientists continue to debate how best to measure democracy and, by extension, whether democratic backsliding is occurring. A central limitation is a mismatch between theoretical claims – which are inherently about specific institutional actions, mechanisms, and processes – and the data available to evaluate them. Political events are the appropriate unit of analysis for bridging this divide: they denote discrete, bounded actions by identifiable actors with consequences for democratic institutions, but they are difficult to produce at scale. This paper presents a transparent, LLM-based pipeline that transforms Freedom House's annual country reports (1990–2024) into structured event data, providing an evidentiary layer that supports both qualitative and quantitative assessment. Our four-stage process – digitization, text extraction, event extraction, and annotation – produces nearly 200,000 annotated events across 228 countries, each tied to textual evidence. Comparing the outputs against human coders and existing datasets shows that the pipeline produces outputs with high construct validity: sentence selections align with human coders at rates comparable to human-human agreement, 99 percent of benchmark events traceable to Freedom House sources appear in our output, and aggregate trends correspond to established indices of democracy. We use the data to compare democratic backsliding in Hungary and Poland, accounting for uncertainty in how individual events are scored. The primary contribution is in developing and advocating for *auditable interpretation* as an appropriate standard for LLM-assisted measurement, which focuses on producing outputs that are traceable to verbatim source material and assessable by downstream users. Event-level data generated through such pipelines enables analyses of institutional sequencing and mechanisms of democratic change where both indices and case studies sometimes fall short.

Keywords

Democracy; Democratic backsliding; Qualitative data; Freedom House; Large Language Models

*The authors would like to thank Aram Jung for their invaluable assistance with case study validation. This work received financial support from the Institute for Humane Studies under grant number IHS018727 and by a University of South Carolina College of Arts and Sciences Faculty Research Initiative under grant number 25R-4001.

https://doi.org/10.33774/apsa-2026-hgrxk **ORCID:** https://orcid.org/0000-0003-1955-281X Content not peer-reviewed by APSA. **License:** All Rights Reserved

Introduction

A 2024 exchange in *PS: Political Science & Politics* debated the validity of expert assessments of democracy for measuring democratic backsliding and which indicators are appropriate for gauging it. Little and Meng (2024) argue that perceptions of declining democracy can be driven by subjective assessments that reflect time-varying coder bias and that more 'objective' indicators like incumbent turnover and multiparty competition show little evidence of backsliding. A central point of contention with these claims is that democracy is a complex, high-level concept that encompasses multiple features and is hard to represent with just one or a few items (Knutsen et al., 2024). Even ostensibly 'factual' indicators reflect subjective choices about what to focus on and how to handle ambiguous cases, and they can even be more biased because of what they do not capture (Knutsen et al., 2024).

To some extent, the debate over which measures can be trusted and how to best quantify democracy and democratic backsliding distracts from the reality on which most scholars agree. As early as 2015, scholars have begun documenting the varied ways that recent leaders and elites have undermined democracy (Diamond, 2015). Scholars are generally united in consensus that the quality of democracy has been in decline in a number of countries worldwide, even if specific measures do not reflect it (Bermeo, 2016; Diamond, 2015; Haggard & Kaufman, 2021; Levitsky & Ziblatt, 2018; Waldner & Lust, 2018). Such works emphasize that democratic declines have been more subtle and gradual than in the past and rely on comparative historical analyses to show the processes at work. Although quantitative measures can effectively communicate real-world trends, they do not always do justice to qualitative interpretations of the events that underlie them (Gerring, 2012).

Answering the question of whether democratic backsliding is real, measurable, and reversible is thus hampered by a mismatch between the theoretical claims scholars make and the data available to evaluate them. Those claims are inherently about sequences of actions that connote distinct processes and mechanisms – about which institutions are targeted, in what order, and whether structural changes occur. Creating assessments of democratic quality and compressing information about it into indices obscures the concrete actions that drive those assessments and characterize specific processes. What we need to complement existing assessments are data that identify the

https://doi.org/10.33774/apsa-2026-hgrxk **ORCID:** https://orcid.org/0000-0003-1955-281X Content not peer-reviewed by APSA. **License:** All Rights Reserved

discrete actions that supposedly justify composite judgments, making the factual basis auditable rather than opaque.

The appropriate unit of analysis for addressing this is the political event – a discrete action by an identifiable actor with traceable consequences for democratic institutions. Events occupy the analytical space between purely quantitative measures – which reduce complex realities to scalar values – and purely qualitative assessments – which offer narrative richness but resist systematic comparison. By anchoring analysis in what actors did rather than interpretations of why or what it means, event data help guard against the post-hoc attributive reasoning that Little and Meng (2024) identify as a source of bias, while preserving enough specificity to make the evidence underpinning democratic assessments auditable. Manually producing a dataset of this kind is extremely labor-intensive, however, and requires innovative solutions.

In what follows, we outline our approach to using qualitative narratives to produce event data. This paper presents a pipeline for automating the extraction and classification of political events from Freedom House country reports at scale, using large language models – specifically GPT-4o. The primary question motivating the project is whether LLMs can transform large amounts of qualitative information into structured, auditable event data that preserves the specificity of narrative accounts while enabling systematic cross-national comparison. Through carefully constructed prompts developed through extensive iteration, we extract and classify nearly 200,000 discrete political events across more than 140 countries from 1990 to 2024, each tied to verbatim textual evidence and annotated by institutional theme and potential democratic impact. The resulting data capture both illiberal and pro-democratic actions, enabling analysis of the fuller dynamic. Because we begin with the full population of politically relevant actions as Freedom House reports them, the approach supports analyzing patterns that emerge from the record without prior selection based on theoretical expectations about what backsliding should look like.

We evaluate the construct validity of the outputs by examining evidence at each stage of the pipeline and comparing it against different standards. We show that the pipeline's sentence selections fall within the range of human-to-human agreement, suggesting the model responds to the same textual features that human analysts would treat as meaningful. We also show that the events

https://doi.org/10.33774/apsa-2026-hgrxk **ORCID:** https://orcid.org/0000-0003-1955-281X Content not peer-reviewed by APSA. **License:** All Rights Reserved

generated by our pipeline have near-complete correspondence with human-coded events that are traceable to the same Freedom House sources, while yielding substantially greater coverage and finer disaggregation of compound actions. Moreover, we show that democratic sentiment scores derived from event annotations track known trajectories of democratization and backsliding and align with patterns in V-Dem and Freedom House indices while preserving the institutional heterogeneity that aggregate scores compress. Additionally, outputs demonstrate the reliability that construct validity requires as a precondition: the vast majority of sentences are consistently selected across independent runs, and nearly all events show identical or highly similar descriptions across iterations. This supports computational reproducibility conditional on model access and version, which is weaker than full replicability but constitutes a significant methodological contribution. The pipeline generates outputs that adequately capture discrete political actions with traceable consequences for democratic institutions and that resemble what human coders could produce.

The results of our validity checks demonstrate *auditable interpretation*: the pipeline's choices are reasonable approximations of human judgment, traceable to verbatim source material, and transparent enough for downstream users to assess fitness for their specific purposes. This shifts the evaluative question from "Does the LLM match expert judgment?" to "Does the pipeline produce a valid and defensible measure of the construct it claims to capture?" Many of the challenges we identify – edge cases in language, variation in coder judgment, difficulties distinguishing discrete actions from background conditions – are not unique to LLMs but are general measurement problems that affect human coding of qualitative material as well. The key difference is transparency: the pipeline makes interpretive choices explicit, provides verbatim evidence for every extracted event, and documents all prompt templates and processing decisions, making the accumulated evidence for construct validity available for scrutiny. We present this project as a demonstration of ground covered and lessons learned rather than a final, definitive dataset – the rapid pace of LLM development precludes that claim – but the validation results establish that transparent, well-documented LLM pipelines can produce research-grade event data at scale, albeit with some unavoidable variation between runs.

Despite its current flaws and the many ways in our pipeline could be improved, the scope of this project underscores the value of pursuing the approach. Scholars want to get at democratization and

https://doi.org/10.33774/apsa-2026-hgrxk **ORCID:** https://orcid.org/0000-0003-1955-281X Content not peer-reviewed by APSA. **License:** All Rights Reserved

democratic backsliding in ways that go beyond single-digit annual assessments, and rich information from qualitative narratives like the Freedom House country reports has long existed but gone underutilized. Improvements on a pipeline for producing event data from qualitative sources enables scholars to establish consensus on 'objective' events that connote distinct processes and then use them to interpret meaning and consequences. To that end, LLM prompting supports the large-scale development of an auditable record that can be used to advance comparative analysis in new ways.

The paper proceeds as follows. We first review the use of LLMs for measurement in political science, emphasizing their capabilities, known limitations, and emerging best practices for mitigating bias and ensuring reliability. We then describe our four-stage pipeline: digitization of PDF and HTML reports, text extraction to identify event-relevant sentences, event extraction to create structured records with verbatim evidence, and annotation to classify events by institutional theme and democratic impact. After detailing the pipeline and discussing its stability across iterations, we present comparisons that show alignment with human coding and correspondence with benchmark datasets. We then demonstrate the pipeline's utility through a case study of democratic backsliding in Poland and Hungary, showing how event-level data reveals institutional pathways and timing that aggregate indices obscure. We conclude by discussing implications for understanding democratic backsliding, honest limitations of the approach, and potential extensions and refinements. The goal is not to suggest that automation should replace human oversight in data collection, but rather to show that transparent, well-validated LLM pipelines can produce research-grade event data at scale – a significant methodological advance that can help reconnect quantitative cross-national research with the rich qualitative evidence on which democratic assessments ultimately rest.

The Challenge of Measuring Democratic Backsliding

The exchange prompted by Little and Meng (2024) reflects genuine uncertainty in the field about what kind of evidence is sufficient to establish that democracy has deteriorated. Their critique encompasses two related but distinct concerns. The first is that expert judgments are shaped by information and frameworks that may bias assessments of democratic quality at any given point in time. The second is that these judgments are aggregated into composite scores from which it is

https://doi.org/10.33774/apsa-2026-hgrxk **ORCID:** https://orcid.org/0000-0003-1955-281X Content not peer-reviewed by APSA. **License:** All Rights Reserved

difficult to recover what drove them. The argument by Little and Meng (2024) brings to the fore doubts about what democracy indices tell us because it is not always clear what drives them. This issue is not unique to democracy measurement, however. Virtually all indicators reflect deliberate choices about what to measure, how to handle ambiguous cases, and which aspects of a complex concept deserve priority.

Scholars' ongoing efforts to characterize and explain democratic erosion nevertheless exemplify the issue. The comparative politics literature on democratic backsliding converges on the view that contemporary democratic erosion is incremental (Bermeo, 2016; Diamond, 2015), legally ambiguous (Levitsky & Ziblatt, 2018), and can take different institutional pathways (Haggard & Kaufman, 2021; Waldner & Lust, 2018). Thus, while scholars in this area may all agree that democratic backsliding is real, they also underscore the difficulty of conceptualizing and measuring it. If reducing democratic erosion to a composite number risks obscuring individual actions, reducing it to whether a single action occurred is equally blind to incremental change and broader context (Knutsen et al., 2024). Both sides acknowledge the challenge of matching the available data to the theoretical claims being made. Skepticism of expert-derived indices is often grounded in qualitative knowledge. Systematic description can therefore serve as a bridging strategy – creating a record against which numerical estimates can be evaluated.

The comparative historical analyses that outline and describe different forms of democratic backsliding are, at their core, claims about the actions that political actors take to undermine democratic processes. The unit of measurement capable of demonstrating it is at the event-level. To be accurate, data about them must be fine-grained enough, and varied enough, to capture the incremental and diverse nature of contemporary backsliding. To be believable, they must be traceable to source material that scholars accept as true. To be empirically useful, they must be classifiable enough to compare and generalize across cases. These criteria define event-level datasets often used in the literature on conflict and protest, which provide structured information about discrete actions that enables systematic analyses of patterns (Chenoweth & Lewis, 2013; LaFree & Dugan, 2007; Raleigh et al., 2010). Despite subtle differences in the contexts and manners in which terrorist and conflict events occur, such data restrict the focus to what actors *did* and what *occurred* rather than what coders believed about their intentions or significance, prioritizing observation over

https://doi.org/10.33774/apsa-2026-hgrxk **ORCID:** https://orcid.org/0000-0003-1955-281X Content not peer-reviewed by APSA. **License:** All Rights Reserved

interpretation.

When scholars attempt cross-national comparison, they face a choice between sacrificing specificity for breadth – whether by using aggregate scores or narrowing focus to specific indicators that are tractable but potentially miss what matters most – or sacrificing comparability for depth, relying on case studies that illuminate mechanisms but cannot easily be placed in dialogue with one another. Event data offer a middle path between the two. When events are defined in the same way across cases, extracted from the same source material, and annotated using the same classification scheme, the records are comparable in ways that case study narratives are not. Their attractiveness has motivated a turn toward collecting more fine-grained, context-rich information about political actions to identify illiberal behavior that may not fit neatly into standard measures. One of the biggest contributions to date comes from the *Democratic Erosion Event Dataset* (DEED) (Baron et al., 2024), which uses human coding to document specific political events associated with democratic erosion for more than 140 countries since 2000. By drawing on qualitative materials to create structured event data, it helps to identify what happened and why, providing context that would otherwise be obscured or overlooked.

The DEED dataset is an impressive endeavor but has two theoretically important limitations that we hope to build on. The first is scope. By design, DEED initially focused on events associated with democratic erosion and largely did not include actions that are essential for evaluating changes improving democratic quality. The second limitation is scale. Manually coding political events from qualitative reports is enormously labor-intensive, as evidenced by the large consortium of faculty, students, and research assistants required to produce and maintain the dataset (Baron et al., 2024). This imposes practical limits on coverage, update frequency, and variability of the output. Different coders involved in the projects applied different judgments about cases, and the reasoning behind those judgments is not easily recoverable from the final coded record.

Recent advances in large language models (LLMs) offer a means of potentially addressing both limitations. LLMs can process large volumes of qualitative text, identify passages that describe discrete political actions, extract structured event records from those passages, and annotate each record according to a consistent classification scheme at a scale that human coding cannot match

https://doi.org/10.33774/apsa-2026-hgrxk **ORCID:** https://orcid.org/0000-0003-1955-281X Content not peer-reviewed by APSA. **License:** All Rights Reserved

(Barrie et al., 2025). Critically, they can do this for a broader spectrum of politically relevant events, enabling analyses of institutional dynamics in both directions. The case for automation is not simply that it is faster or cheaper, though it is both. It is that automation, properly designed, can help produce event data that are more consistent, comprehensive, and transparent than human coding at scale, even if it not perfectly replicable. Consistency follows from applying the same prompt and classification logic to every document. Comprehensiveness reflects the volume and scale that this approach makes tractable, and transparency from tying each output to verbatim source text and/or an explanation of coding decisions. By isolating discrete actions for annotation rather than assessing them against broader knowledge of a country's trajectory, the approach also helps reduce the post-hoc rationalization that Little and Meng (2024) identify as a source of bias in expert-derived assessments.

Using Large Language Models for Measurement

LLMs and GPT models – a specific family of LLMs that rely on autoregressive next-token prediction as their core learning objective – represent a move from task-specific supervised learning to foundation models capable of flexible responses, text generation, and classification across domains. In political science, scholars are increasingly exploring how LLMs can be used to generate structured data from unstructured text. Much of this work evaluates their abilities to assign structured labels from predefined categories – for example, classifying open-ended survey responses (Mellon et al., 2024), identifying political party or tone from messages (Heseltine & Clemm von Hohenberg, 2024; Törnberg, 2025), or benchmarking performance across multiple annotation tasks (Alizadeh et al., 2024; Gilardi et al., 2023). Other studies extend this logic beyond classification, using LLMs to map texts onto broader constructs – such as reproducing expert democracy assessments (Weidmann et al., 2026), estimating ideological position (Le Mens & Gallego, 2025), or simulating political behavior (Chen et al., 2026; Potter et al., 2024).

The findings from published work show that LLMs can approximate and sometimes exceed human coders. Across multiple applications, model outputs correlate strongly with expert coding and display high levels of coherence (Gilardi et al., 2023; Heseltine & Clemm von Hohenberg,

https://doi.org/10.33774/apsa-2026-hgrxk **ORCID:** https://orcid.org/0000-0003-1955-281X Content not peer-reviewed by APSA. **License:** All Rights Reserved

2024; Weidmann et al., 2026). They can emulate human annotation and approach human-human agreement (Mellon et al., 2024). In particular, GPT-4 has demonstrated high accuracy across a range of classification tasks (Heseltine & Clemm von Hohenberg, 2024) and outperformed trained experts, crowd coders, and conventional supervised classifiers such as BERT and Naïve Bayes on accuracy metrics (Törnberg, 2025). Gilardi et al. (2023) find that GPT exceeds crowd-coding not only in accuracy but also in inter-coder agreement, suggesting that model outputs can be more internally consistent than those produced by distributed human coders. Le Mens and Gallego (2025) also find that GPT-4 is highly accurate for estimating ideological position. LLMs are thus capable not only of classification but of transforming text into quantifiable measures by mapping them onto structured dimensions (Chen et al., 2026; Le Mens & Gallego, 2025; Weidmann et al., 2026).

The limitations of LLMs as coders are well-documented and must be taken seriously. Models exhibit position bias, disproportionately weighting content at the beginning and end of long documents while under-attending to material in the middle (Wu, 2025). They are sensitive to prompt wording: small changes in phrasing or instruction ordering can shift outputs in ways that are difficult to predict (Halterman & Keith, 2025). They can hallucinate – generating confident but unverifiable content – particularly when asked to reason beyond what is present in the provided text (Bowman, 2023; Zhao et al., 2025). And because they are trained on internet-derived corpora, their outputs can reflect the ideological and demographic skews present in that data (Bender et al., 2021; Weidmann et al., 2026). Each of these shortcomings has direct implications for our pipeline design, and we describe the specific mitigations we adopted below.

It is worth noting, however, that these limitations are not categorically different from those that attend human coding of qualitative material. Rater agreement on complex political coding tasks is often poor, and systematic bias, inconsistent application of rules, and undocumented discretionary judgments are endemic to large-scale human coding projects (Mikhaylov et al., 2012; Treisman, 2024). O'Brochta (2025) observes that scholars routinely treat human-coded datasets as authoritative products without examining the process that produced them. Using human assistants to code data rarely requires one to address the degree to which different circumstances could have yielded different codings from the same source material – effectively, its *replicability*. Skepticism about the suitability of LLMs for data collection should therefore acknowledge analogous issues in how human

https://doi.org/10.33774/apsa-2026-hgrxk **ORCID:** https://orcid.org/0000-0003-1955-281X Content not peer-reviewed by APSA. **License:** All Rights Reserved

data are produced and the standards used to assess them. The question is not whether LLMs introduce measurement error – they do, as do human coders – but whether the error is transparent, bounded, and validatable. Our approach is designed with that standard in mind, aiming to produce valid outputs that mimic human judgments while being as transparent as possible and documenting its limitations.

Our Approach

This study uses GPT-4o to extract and classify political events from Freedom House country reports at scale. GPT-4o is a large language model – a transformer-based neural network trained on massive text corpora that can perform classification, extraction, and annotation tasks across domains with little or no task-specific fine-tuning (Brown et al., 2020). Our decision to use GPT-4o was based on its performance as a leading proprietary model that was shown to outperform open-source alternatives at the time of development. This prioritizes performance over replicability, a tradeoff we return to below (Barrie et al., 2025). Like all LLMs, it generates outputs probabilistically rather than deterministically: responses are sampled from a learned distribution over likely next tokens, which means that without deliberate design constraints, the same input can produce different outputs across runs (Brown et al., 2020; Radford et al., 2019). Using it as a measurement tool therefore requires treating prompt construction, decoding parameters, and output validation with the same methodological care one would apply to a survey instrument or a human coding protocol (Halterman & Keith, 2025; Le Mens & Gallego, 2025; Törnberg, 2025).

Although prior studies demonstrate that LLMs can substitute for human coders in classification and scaling tasks, these works largely focus on binary classification from short texts and single-stage annotation. Our project goes beyond strict classification by developing a multi-stage pipeline for extracting discrete political events from narrative regime assessments that emulates the output of human coders (Liu et al., 2025). Like the Democratic Erosion Event Dataset (Baron et al., 2024), we begin with rich qualitative material and use it to extract discrete political events – but at a larger scale than typically possible through manual coding. Our approach draws on capabilities at which LLMs should excel based on their probabilistic assessments of words: identifying relevant

https://doi.org/10.33774/apsa-2026-hgrxk **ORCID:** https://orcid.org/0000-0003-1955-281X Content not peer-reviewed by APSA. **License:** All Rights Reserved

information, parsing descriptions of what happened, and inferring thematic or qualitative attributes such as topics or indicators of democratic practice.

Four design principles guided our construction of the pipeline, each responding to a known limitation of LLM-assisted coding. First, we rely on structured prompting with explicit instructions to guide model behavior (Halterman & Keith, 2025). We treat each prompt as methodologically analogous to a codebook entry: instructions are stated precisely, critical rules are repeated at the top and bottom of each prompt, and formatting requirements are enforced through structured output schemas. To mitigate position bias in long documents, we chunk reports into overlapping segments so that no content appears only once in the middle of a long sequence (Wu, 2025). Instructions explicitly direct the model to draw only on information present in the provided text, reducing reliance on priors from training data.

Second, we use deterministic decoding. To minimize output variability across runs, we set the model temperature to zero, which suppresses sampling stochasticity in the token generation process (Alizadeh et al., 2024; Gilardi et al., 2023). In practice, this does not fully eliminate residual non-determinism for proprietary models. Barrie et al. (2025) demonstrate that between-run correlations for GPT-4o remain below 1.0 even at temperature = 0, with a degree of variance across tasks in ways that are difficult to predict. Our stability checks should therefore be understood as providing evidence of robustness reproducibility in Brodeur et al.'s terms: the same patterns emerge across runs despite some variation in specific outputs, rather than confirming exact replicability (Brodeur et al., 2024).

Third, we include verbatim evidence requirements. Every extracted event must be anchored to an exact quotation from the source text. This constraint serves two purposes. It prevents hallucination by bounding the model output to what is actually present in the document: the model cannot fabricate an event without also fabricating a quotation, which is easier to detect and filter (Bowman, 2023). It also ensures that the pipeline's interpretive choices are auditable at the sentence level – every event record can be traced back to the exact passage that generated it, allowing downstream users to evaluate whether a given extraction is defensible.

Fourth, we use a decomposed task structure. Rather than asking the model to perform

https://doi.org/10.33774/apsa-2026-hgrxk **ORCID:** https://orcid.org/0000-0003-1955-281X Content not peer-reviewed by APSA. **License:** All Rights Reserved

extraction, classification, and annotation in a single pass, we break the pipeline into discrete stages, each with a focused and bounded task. This follows the principle that accuracy degrades as task complexity increases (Heseltine & Clemm von Hohenberg, 2024), and that separating extraction from classification allows each step to be validated independently. It also produces a structured audit trail: if an annotation is questionable, the error can be localized to the specific stage at which it occurred rather than being lost in the workflow.

Data Source

We focus exclusively on the qualitative country narratives published annually by Freedom House, a non-profit organization founded in 1941 that is widely recognized for its focus on democracy and human rights. Since 1972, Freedom House produces annual global assessments assigning numerical ratings of political rights and civil liberties and classifying each country as *Free*, *Partly Free*, or *Not Free*. Beginning in 1978, these scores were accompanied by country-level narrative reports written by teams of in-house and external experts summarizing each year's major political developments (Giannone, 2010; Karatnycky, 2002). We use reports from 1990 onward (covering events from 1989) because the earliest reports were very short and because the end of the Cold War represented a widely recognized shift in global politics that serves as a natural starting point (Kalyvas & Balcells, 2010; Woldense & Kroeger, 2024).

Working from a single body of text supports a replicable and transparent approach to event extraction that can later be extended to additional corpora, such as reports from news and non-government organizations. Although the numerical scores produced by Freedom House are regularly used in cross-national research (Coppedge et al., 2011; Munck & Verkuilen, 2002), the textual narratives on which those scores are based remain underutilized. They describe not only the outcomes that shape regime trajectories but also the events, mechanisms, and processes underlying those outcomes. By transforming these qualitative narratives into structured event data, we aim to provide greater specificity than is possible from the numerical scores alone.

Freedom House is well-suited for LLM-based event extraction for several reasons. First, its assessments are well known and generally trusted among political analysts and researchers as a

https://doi.org/10.33774/apsa-2026-hgrxk **ORCID:** https://orcid.org/0000-0003-1955-281X Content not peer-reviewed by APSA. **License:** All Rights Reserved

credible source of democratic evaluation. Second, the reports span a long period of time and are exhaustive, covering nearly every country in the world from 1978 to the present. Third, Freedom House offers conceptual consistency, insofar as the organization maintained a clear and stable mission over five decades even as the reports themselves have changed. The reports also do not have strong retrospective bias, since each year's report details the events that occurred in the prior year.

The reports cover 228 countries over the period 1990-2024 (published for 1989-2023) and constitute over 20,000 pages of material. Using the NLTK sentence tokenizer in Python to split text into sentences based on punctuation yields 562,022 total sentences (see Figure A.1). As might be expected, the number of sentences used to describe political activity in a country increased over time, particularly after 2015 when reports moved online. For some countries and years, Freedom House did not write a full report because little had changed; instead, it referred readers to the prior year's report.[1] There are also a small number of countries without reports that were not included in that year's assessment. Table C.1 in Appendix C lists the year range and number of country reports that we processed.

Our data source has substantial implications for the output that we obtain, as it is guided by at least three distinct layers of interpretation that each introduce different biases. The first is editorial selection: Freedom House narratives are curated assessments that prioritize events judged relevant to democratic rights and civil liberties.[2] The second is interpretive framing: analysts embed events within broader narratives using normatively laden language that reflects Freedom House's analytical framework for assessing democratic quality. The third is algorithmic extraction: our LLM processes only the published text, making extraction decisions conditioned on Freedom House's framing and on the model's own learned priors about political significance – priors derived from training data rather than political science theory. Because GPT-4o was trained on internet text, it may also have prior knowledge of Freedom House reports or the events on which they are based. It is for this reason that we require the model to base its decisions on a verbatim quote.

[1] Countries with a very small number of sentences are 'abridged'.

[2] Freedom House has been criticized for having differential coverage that can reflect U.S. foreign policy interests and a normative framework that may evaluate institutions in non-Western contexts through a particular lens, though such critiques are directed primarily at the opacity of the aggregation process that converts qualitative assessments into numerical scores (Giannone, 2010).

https://doi.org/10.33774/apsa-2026-hgrxk **ORCID:** https://orcid.org/0000-0003-1955-281X Content not peer-reviewed by APSA. **License:** All Rights Reserved

The output should therefore be understood as systematized Freedom House interpretations of events rather than direct observations of empirical reality. The events we identify are model predictions from Freedom House descriptions that resemble human recognition of events and annotations of those reports. This is not a limitation unique to LLM-based approaches – manually coded datasets based on news sources face analogous issues of media selection bias and journalistic framing (O'Brochta, 2025). What distinguishes our approach is the explicitness with which we acknowledge these layers and the transparency with which we document the choices made at each stage. At the same time, these layers of interpretation define the inferential scope of the data in a way that is analytically useful. Our pipeline does not resolve the critiques noted above, but it does address the opacity problem directly: by making Freedom House's own judgments about democratic relevance explicit, traceable, and systematically organized, it allows downstream users to interrogate the assessments that underlie its numeric scores. By aiming for descriptive data that map back onto the reports with high fidelity, we are developing a resource for evaluating and interrogating those interpretations. This is different from a complete account of what might exist outside of Freedom House reports.

To anchor our understanding of whether model outputs were face-valid, we worked as a team of four – a faculty member, post-graduate student, graduate student, and high-school volunteer – to thoroughly evaluate outputs at every step in the process for eight countries selected to represent varied regional and democratic trajectories based on trends in the Freedom House data. These included cases with little change over time at different levels of 'democraticness' (Austria, Mexico, Zimbabwe), cases with gradual democratic decline (Bangladesh, Belarus, Nicaragua, Serbia), and instances of sudden democratic change (Tunisia after 2010), with overlap across evaluators. We scrutinized the reports for these countries to develop deep knowledge of the case and the events that Freedom House described in each. This provided a basis for identifying candidate sentences and evaluating edge cases, judging the correctness of extracted information, and comparing the outputs against alternative standards (i.e., DEED data and V-Dem and Freedom House scores).

https://doi.org/10.33774/apsa-2026-hgrxk **ORCID:** https://orcid.org/0000-0003-1955-281X Content not peer-reviewed by APSA. **License:** All Rights Reserved

The Pipeline Process

The process of converting the qualitative narratives from Freedom House (FH) reports into a structured set of information is designed to transform information that has historically been too large or unwieldy for systematic use into analytically tractable units of political behavior (Grimmer et al., 2022). The entire workflow, illustrated in Figure 1 and accomplished using Python, moves through four key stages: Digitization (pre-processing), where the raw PDF files are converted and cleaned; Text Extraction (1), where the LLM identifies potentially salient information in the cleaned text; Event Extraction (2), where the identified sentences are converted into discrete, structured event records; and Annotation (3), where the events are tagged with thematic and sentiment labels to assist with sorting and analysis.[3] This approach provides a viable and cost-efficient alternative to manual coding while incorporating strict validation steps to manage the risks of high variance and fragility associated with LLM outputs (Barrie et al., 2025; Bisbee et al., 2024). Our validation steps are layered and sequential rather than relying on any single point of verification, resembling quality assurance protocols for determining failures in production runs. After prompt development and stability testing, we executed the full pipeline in a single run over approximately one week in November 2025, thereby limiting the risk of temporal drift.[4] Below, we briefly outline the major decisions and steps (also enumerated in Appendix D), as well as consistency checks.

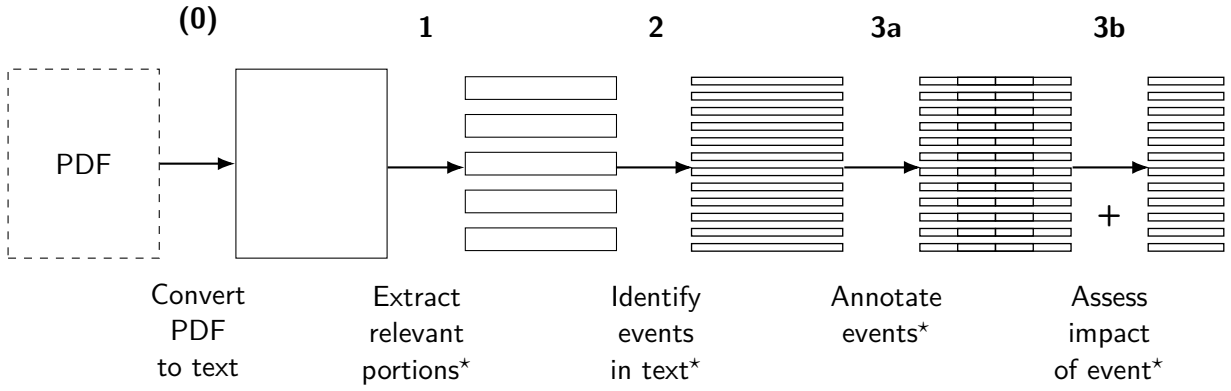


Figure 1: Steps in the process (numbered). Stars indicate GPT prompts.

[3] In Figure 1, we enumerated the digitization step with (0) to distinguish it from the remaining steps, all of which involved LLM prompting.

[4] We did not version lock, mainly because we were developing prompts over a long period of time and did not anticipate the same version to be used in future applications. Not version locking reduces the reproducibility of our event records, but downstream inferences are unlikely to be sensitive to model version changes (Bisbee et al., 2024).

https://doi.org/10.33774/apsa-2026-hgrxk ORCID: https://orcid.org/0000-0003-1955-281X Content not peer-reviewed by APSA. License: All Rights Reserved

Digitization (pre-processing)

The initial step – which preceded the use of an LLM – involved digitizing the textual data and turning it into standardized country-year observations. Reports for 1990-2016 were originally published as PDF files, while later reports (2017 through 2024) are available as HTML web pages. Due to shifting editorial and formatting conventions over time (see Figure E.1), we processed each report year separately and tailored cleaning routines and verification steps to specific years and issues. As we manually checked and cleaned the digitized data, we maintained detailed documentation, recording each fix and the logic behind it. The result is a relatively clean and structured corpus in which the original narrative reports are organized into standardized country-year blocks of text. Although the workflow successfully produced a corpus suitable for event extraction and quantitative text analysis, the data still contain some errors (e.g., OCR errors and formatting errors such as missing punctuation or incorrect spacing). In working with the output, however, we found that improperly spaced wording did not affect the model output and that improperly split sentences did not substantially affect the descriptive outputs for a subsample of test cases.

Text Extraction (1)

The Freedom House country reports contain a lot of information that is not relevant to the year in question (e.g., history and background description), as well as general descriptions of the type of government and state of affairs in the country. This first step, therefore, breaks text into sets of sentences that might contain relevant information, akin to highlighting text in a book. This was necessary to lower the cost associated with per-token LLM usage and mitigate potential bias towards shorter text. The script provides structured context windows of ± two sentences (blocks of sentences that overlap by two) and employs the "Persona Pattern" (White et al., 2023), an output customization technique in which the model is instructed to act as a specific role and generate outputs that such a persona would produce – in this case, a political scientist reading the Freedom House report published in `{report year}`. The model is required to flag each sentence as 'KEEP' or 'SKIP' based on whether it describes events according to predefined criteria, including legislation, protests, and institutional changes. We added an explicit rule that automatically tags a sentence as

https://doi.org/10.33774/apsa-2026-hgrxk **ORCID:** https://orcid.org/0000-0003-1955-281X Content not peer-reviewed by APSA. **License:** All Rights Reserved

'KEEP' if it mentions the target year (report year − 1).

To prevent hallucinations, we employ a conservative fallback strategy: when the model returns fewer tags than expected, missing entries default to exclusion (`KEEP=False`) and are logged as failures. We reprocessed remaining failed batches and manually reviewed them where necessary, ensuring that every segment of text is either successfully processed or explicitly flagged. This stage extracted approximately 226,329 sentences – about 40 percent of the initial corpus – which serve as the input to the event extraction stage. Figures A.2 and A.3 show the proportion of highlights extracted from country-year reports and the overall distribution.

To assess the extent to which GPT's selection of sentences mimics the decisions human coders, we manually performed the same task for our eight cases. To accomplish this, two human coders independently reviewed each case and highlighted sentences describing political events.[5] We then calculated overlap among all three coders using F-scores as the primary validation metric (Grimmer et al., 2022), following Mikhaylov et al. (2012)'s standard of whether a coding instrument yields the same data within a tolerable margin of error. Human-to-human agreement averaged F = 0.71; it revealed substantial variability, however, ranging 0.40-0.91. GPT-to-human agreement fell well within the range of human performance, averaging F = 0.66 and ranging 0.51-0.75. Figure A.4 breaks out the F-score for each of the two human coders and GPT-4o into its components and compares the share of extra items selected (precision) against the share of true items captured (recall). In every case, GPT-4o retrieved a substantially larger set of sentences than either human coder (see Figure A.5), penalizing its precision while achieving consistently high recall.[6] Human recall was more variable, with substantial disagreement on Belarus and Nicaragua. For some countries, GPT-4o's performance exceeded human-human agreement, consistent with similar findings (Rytting et al., 2023). After reviewing sentences highlighted by humans but missed by GPT-4o, coders found that many had been erroneously selected in the first place, suggesting the LLM output falls within the range of humanly-plausible interpretations.

[5]Country reports were output as Microsoft Word documents; participants used the highlight tool to indicate important text. The highlight extraction prompt retrieved candidate sentences from the same reports.

[6]The GPT-4o-only fringe largely consists of ambiguous or low-salience phrasings rather than clear political events rejected by humans, reflecting the fact that the extraction prompt is tuned for recall rather than precision; since downstream discards are irreversible, we cast a wide net at this stage.

https://doi.org/10.33774/apsa-2026-hgrxk **ORCID:** https://orcid.org/0000-0003-1955-281X Content not peer-reviewed by APSA. **License:** All Rights Reserved

We also evaluated whether highlight extraction is stable across multiple runs by rerunning the text extraction prompt five times each for our eight test cases and examining the number of highlights that were kept each time (Figure A.6). The difference in the number of highlights selected for each country year ranged between a minimum of 0 and maximum of 30 (Belarus, report year 1992), with a median of 3 highlights. Agreement ranges between 95 and 98 percent in pairwise comparisons (Figure A.7), and between 87 and 90 percent of sentences were selected across all runs (Figure A.8). This confirms both that the model contains some inherent unpredictability but also that it is largely consistent in which text it returns as potentially containing a political event.[7]

Event Extraction (2)

The second step used GPT-4o to turn retained sentences into structured event records, further filtering irrelevant information. The script processes extracted highlights in adjacent chunks of ten sentences, running multiple batches simultaneously to minimize runtime. The model is instructed to function as an event coder, identifying discrete, verifiable actions involving broadly defined political actors (state elites, opposition, and collective citizen behavior). Each event is constrained to a structured JSON format – a "Template Pattern" (White et al., 2023) – with three fields: a concise `Shorthand` phrase, a fuller `Description`, and a verbatim source `Quote`. After the main run, an auxiliary script recovered events from rejected highlights containing month references without explicit years (e.g., "In January..."). The event-extraction stage produced 198,165 discrete political events from 165,262 unique highlights (approximately 73 percent of highlighted sentences and 29 percent of total Freedom House text). Figures A.9 through A.11 show the proportion of event entries extracted from highlights for each country-year and the overall distribution.

At this stage, we inspected sentences that passed the first stage but were not used to construct events, evaluating whether rejected sentences should have been rejected. We also manually reviewed extracted events for reasonableness – not whether descriptions perfectly matched how we would have documented an action, but whether they were acceptable representations of it. As a check for

[7]Fleiss' kappa averaged 0.496 (range: 0.468-0.528), indicating moderate agreement across runs. Pairwise Jaccard similarity averaged 0.94, with 87-90 percent of sentences consistently selected (5 times), demonstrating high reproducibility at this stage.

https://doi.org/10.33774/apsa-2026-hgrxk **ORCID:** https://orcid.org/0000-0003-1955-281X Content not peer-reviewed by APSA. **License:** All Rights Reserved

replicability, we reran the event extraction prompt multiple times using Nicaragua as a test case. Out of 5 runs, GPT-4o identified 1322 events 3 times; 1 run produced 1320 events, and another 1328. Across all candidate events, 71 percent (4680/6614) had identical shorthand descriptions and another 15 percent clustered together based on high cosine similarity ($\geq$0.65), for a total of 86 percent. Adding together clusters with multiples of five (see Figure A.12) indicates $\sim$90 percent agreement, suggesting that all five runs picked up on the same types of events.

Annotation (3)

While the previous stages isolated candidate sentences and extracted discrete event records, those records consisted of natural language descriptions requiring further organization for comparative analysis. The annotation stage maps these qualitative descriptions onto fixed, pre-defined categories. Unlike event extraction, which required the model to act inductively to identify what happened, this stage is deductive: mapping observed behaviors onto specific concepts. This mitigates concept drift across the 35-year corpus and supports statistical aggregation. Because annotation involves substantial interpretation, however, we did not expect high classification accuracy. The tags are not meant to be definitive, but to make it easier to sort and group event descriptions into manageable subsets that can be reviewed and checked against the underlying text.

We conducted several supervised classification passes using GPT-4o, enriching each event record along three dimensions: its *form*, its *theme*, and its *democratic sentiment* (outlined in Table 1). For the event-type and theme classifications (passes 1 and 2), the model returns 1–5 verbatim evidence spans and selects from predefined taxonomies with few-shot prompting. We used the V-Dem codebook (Coppedge et al., 2025) to create theme categories, which grounds the taxonomy in one of the most widely used democracy measurement frameworks in the discipline rather than reflecting arbitrary design choices.[8] For democratic sentiment (pass 3), the model selects an ordinal label and returns a concise justification (<20 words) using the "Reflection Pattern" (White et al., 2023). All passes use predefined taxonomies, include an 'other' category, and run at temperature = 0.

[8]We initially explored inductive approaches, including dictionary-based methods and retrieval-augmented classification. However, we found that these produced taxonomies with too many categories whose natural aggregation points mirrored existing institutional schemes.

https://doi.org/10.33774/apsa-2026-hgrxk **ORCID:** https://orcid.org/0000-0003-1955-281X Content not peer-reviewed by APSA. **License:** All Rights Reserved

Automated string-matching confirms that returned evidence spans are exact matches of the original text, ensuring each classification decision is traceable to its source. The result is an annotated dataset where each event is classified by type, theme, and democratic sentiment, with verbatim evidence supporting each decision.

After running the annotation prompts, we manually reviewed classifications for each of the eight test cases to determine whether the model was reasonably tagging and classifying events in terms of the type of description and themes that might apply. We also evaluated whether the sentiment scores are consistent across multiple runs by rerunning the annotation pass for Nicaragua five times, finding a ∼91-percent match. For all but one of the remaining entries, votes straddled two neighboring values.[9] There is no discernible relationship between the value selected by the model and variation in the output.

Table 1: Summary of classification passes.

Pass	Classification	Task and Taxonomy
1	Event type (with examples)	**Distinguish discrete actions from background statements.** *Taxonomy (0–11):* 0 Specific action; 1 Habitual/Generic; 2 Status/Law; 3 Passive/Agentless; ... 6 Background/History; 7 Intent/Proposal; ... 11 Past action (< target year).
2a	Theme (with examples)	**Identify the relevant institution.** *Taxonomy:* `BUREAU` (Bureaucracy), `CIVSOC` (Civil Society), `ELECT` (Elections), `EXEC` (Executive), `INTL` (International), `JUD` (Judiciary), `LEGIS` (Legislature), `MEDIA` (Media), `PARTY` (Parties), `SECUR` (Security), `WOMMIN` (Women and Minorities).
2b	Sub-theme	**Capture additional data for high-interest themes.** Triggered *only* for events tagged `CIVSOC` or `WOMMIN`. Uses an external taxonomy to assign specific codes (e.g., "Labor and Union Activities," "Indigenous Groups")
3	Democratic sentiment	**Assess the impact of the event on democracy.** *Label corresponding to ordinal score:* Entirely Bad (−2), Mostly Bad (−1), Neutral/Mixed (0), Mostly Good (+1), Entirely Good (+2).

[9]Approximately 5 percent (70) produced the same score 4 out of 5 times with one being ±1, while 3 percent (54) produced the same score 3 times and a neighboring score 2 times.

https://doi.org/10.33774/apsa-2026-hgrxk **ORCID:** https://orcid.org/0000-0003-1955-281X Content not peer-reviewed by APSA. **License:** All Rights Reserved

Summary

The pipeline converted 562,022 sentences from Freedom House reports for 228 countries across 1989–2023 into 198,165 discrete, annotated event records. Additional de-duplication passes documented in Appendix F reduced this to 183,909 records. Excluding the many runs required to optimize prompts and parameters, running the final pipeline took approximately 24 hours implemented over one week and cost roughly $1,500 in API charges to produce.[10] We manually confirmed that all quotes are directly attributable to the original Freedom House text. Full documentation of system prompts, configuration parameters, and output files is provided in Appendices G, H, and I, all of which will be made publicly available.

Validation

The dominant validation paradigm in LLM-assisted text coding treats expert human annotations as a gold standard and measures model performance against it. But for interpretive tasks – where coding requires judgment about what counts as an event, how to bound it, or what it means – there is not an obvious ground truth to approximate. We found early on that we as human coders frequently disagreed on what information was relevant, what actions were being described, and how events should be interpreted.[11] Human coders make defensible but contestable choices, and elevating any one set of those choices to gold-standard status smuggles in important theoretical assumptions. We therefore argue that a more appropriate standard is *auditable interpretation*: rather than asking how closely LLMs replicate human judgment, researchers should ask whether the system's interpretive choices are reasonable, be transparent about the criteria and prompts that guided them, document limitations, and make outputs assessable for downstream use. This shifts the validation goal from accuracy against a specific benchmark to fitness for use.

We nonetheless use two human-produced standards as benchmarks – not as ground truth, but as

[10] The full cost of developing the approach, testing prompts, and evaluating output consistency was much larger.

[11] One clear example comes from the 1990 report for Afghanistan: "After the troop pull-out Najibullah declared a state of emergency, launched a crackdown against suspected resistance supporters and quadrupled the wages of officers and soldiers." Whereas all coders might agree that this constituted illiberal behavior, only GPT saw it as three distinct actions that each had different implications for democracy.

https://doi.org/10.33774/apsa-2026-hgrxk **ORCID:** https://orcid.org/0000-0003-1955-281X Content not peer-reviewed by APSA. **License:** All Rights Reserved

a basis for structured comparison. First, we compare our extracted events to those created by Baron et al. (2024) to identify how well the pipeline outputs match with descriptions created by a team of human coders that shared a similar goal. Second, we examine aggregated sentiment scores against V-Dem and Freedom House values to see how well the LLM judgments about the consequences of events for democracy follow expert-derived assessments. These validation checks demonstrate both the interpretive challenge of creating structured data from unstructured information and the alignment of LLM outputs with existing data and approaches that use human coding.

Comparison to DEED

The Democratic Erosion Event Dataset (DEED) (Baron et al., 2024), a human-coded dataset of illiberal actions associated with democratic backsliding, provides a benchmark for determining whether our pipeline flags the right kinds of political actions and describes them appropriately. For each of our eight test cases, we subset the published DEED events (Version 7) and manually classified each record as a full match, partial match, no match, or out of scope.[12] In addition to manual review, we used cosine similarity scores from a pretrained sentence embedding model (all-MiniLM-L6-v2) to retrieve the closest matching sentence in the original Freedom House reports, which we checked to see if our output included.

The events produced by GPT-4o adhere very closely to the 1544 events that DEED coded for our 8 test cases, as described by Table 2. There is a clear match in the LLM-constructed set with 75 percent of the observations in the DEED version; an additional 3 percent of cases found partial matches (mentions that describe actions in a different way), and 3 percent that our coders initially thought were not relevant or non-matching but that had high cosine similarity (they were in our outputs). Importantly, DEED used a variety of sources to identify backsliding events. Version 7 of the DEED data indicate whether they consulted Freedom House in an identifier (*Event_Number*) and a column called *FH_year*. For 142 observations, this column contains 'NA', suggesting that they used a different source. Among these, we still found matching events for 84 (59 percent). There were also 91 observations that we strongly suspect of coming from other sources, which is

[12] Bangladesh and Zimbabwe are not included in Version 7; we used Version 6 for these cases.

https://doi.org/10.33774/apsa-2026-hgrxk **ORCID:** https://orcid.org/0000-0003-1955-281X Content not peer-reviewed by APSA. **License:** All Rights Reserved

supported by the very low cosine similarity scores and our inability to trace the descriptions back to the Freedom House reports. We consider an additional 53 entries as either being outside of our focus (37), problematic (11, discussed below), or described in the Freedom House report in a way that we would have excluded it (5).

Table 2: Comparison of outputs to DEED events for eight countries.

Match	*N*	**Avg. Sim.**	**% (total)**	**% (counted)**	**% (NA)**
Counted, in data					
Full match	1161	0.769	75	92	—
Partial match	46	0.544	3	4	—
Non-matching, sim-matched	31	0.680	2	2	—
Non-event, sim-matched	8	0.743	1	1	—
Missing	12	0.671	1	1	—
NA (not counted)					
Matching, NA	84	—	5	—	59
Non-matching, NA	58	—	4	—	41
Not counted, not in data					
Other source suspected	91	0.483	6	—	—
Non-event	37	0.682	2	—	—
Time problem	11	0.755	1	—	—
Vaguely worded	5	0.644	0	—	—

Among all the descriptions in the DEED data for our eight cases, we found an approximately 86-percent match among our outputs. This includes 81 percent of descriptions that we anticipated could have come from Freedom House, plus 5 percent that likely came from alternative sources. For the 1246 entries that we would have expected to capture and that clearly have a reference in the Freedom House Country Reports, our outputs show a *99-percent match*. This evidences strong adherence between the event descriptions that our pipeline produced and those constructed by a human team.

Comparing our outputs with DEED is also remarkable because it reveals a number of issues that are either potential problems also confronted by our approach or that our approach addressed. First, some of the entries in DEED have a granularity problem; it frequently aggregates multiple actions into a single event. In other places (e.g., cataloging the murders of individual journalists in Mexico), events in DEED are highly itemized. DEED also does not contain direct quotes that are easily mapped back to the source, making it nearly impossible to identify where the information (or interpretation about it) came from. The project initially focused exclusively on documenting democratic erosion, leaving out actions that might balance out assessments about the health of

https://doi.org/10.33774/apsa-2026-hgrxk **ORCID:** https://orcid.org/0000-0003-1955-281X Content not peer-reviewed by APSA. **License:** All Rights Reserved

democracy in a country. Moreover, it used sources from later years to document events for a given year, which introduces the potential for retrospective bias. It also suffers from inconsistencies that might be expected in both human-coded and automated data: spelling and punctuation errors in event descriptions, duplicate entries, and incorrect temporal references. In fact, Version 7 of the DEED data seems to be missing the entirety of the year 2006 for Nicaragua. The purpose of noting these issues is not to disparage the laudable effort to advance our understanding of democratic backsliding by scouring qualitative sources. Rather, it demonstrates that many of the issues that might arise from an automated approach also appear in the human analogue, despite its reliance on extensive person-power and validation.

The observation that our automated pipeline finds such close correspondence with a human-coded benchmark is important. Despite the fact that both projects were produced independently and used different approaches, instructions, and assistant pools to identify events, we find direct and substantial overlap between what human coders identified as relevant events in FH narratives and what our automated pipeline extracted from the same source material. Our pipeline yielded substantially more discrete event records than DEED across the eight cases (8,956 versus 1,554). Although some of the disparity could be due to over-extraction of background statements, we anticipate that it reflects both the broader event scope of the extraction task and the fact that FH reports frequently describe multiple actions that DEED does not enumerate as separate records. The close approximation matters not only as evidence of scalability for a difficult and worthwhile task, but also because key coding choices in DEED are sometimes unspecified or difficult to reproduce, whereas our pipeline is fully scripted and can be more easily audited and extended to additional material. This approach may therefore offer one way to build on and expand the admirable product that DEED represents.

Comparison to V-Dem and Freedom House Indices

Here, we focus on GPT-4o's assessment of an event's potential impact on democracy in a country. While GPT-4o is fundamentally a predictive language model that does not independently reason, its training on extensive corpora of expert analysis enables it to produce responses that approximate

https://doi.org/10.33774/apsa-2026-hgrxk **ORCID:** https://orcid.org/0000-0003-1955-281X Content not peer-reviewed by APSA. **License:** All Rights Reserved

expert judgment patterns. When asked to justify a democratic sentiment classification, the model can draw on patterns learned from political science texts, policy analyses, and expert commentary to select the closest textual response (not a number) and generate rationales that mirror the analytical frameworks that human experts would apply. We leverage this trained approximation of expert reasoning not because the model "understands" democracy but because its probability estimates, conditioned on appropriate prompts and taxonomies, should produce justifications that reflect scholarly consensus on what makes events democratically significant.

We evaluated the model's assessments of democratic sentiment by comparing the extent to which summed ordinal values for a country-year align with aggregated democracy indices such as the Electoral Democracy Index from V-Dem (Coppedge et al., 2025) or Freedom House's numerical ratings for Political Rights and Civil Liberties. To make them comparable, we added together, normalized, and inverted the two Freedom House scores so that higher values indicated a more democratic country and that both V-Dem and Freedom House values are bounded between zero and one. We similarly added together the ordinal scores across events for each country-year – assigning a -2 to 'entirely bad' events, a -1 to 'mostly bad events', and so on – and normalized values to lie on the same scale. As Figure 2 shows, there is considerable alignment with the two indices for our eight test cases. Values fluctuate in ways that match the direction of the trajectories of each case, showing the steady decline of democracy in Nicaragua under Daniel Ortega after 2007 and the initial optimism for democracy in Tunisia after 2010.

The summed democratic sentiment scores show virtually no correlation with either index overall; rather, the pattern varies considerably across countries.[13] Among our eight cases, correlations between summed sentiment scores and V-Dem's Electoral Democracy Index (EDI) range between ρ=0.091 (Mexico) and ρ=0.749 (Serbia). Correlations between summed sentiment scores and Freedom House values range between ρ=0.328 (Mexico) and ρ=0.773 (Nicaragua), while correlations between the EDI and Freedom House range from ρ=0.650 (Serbia) to ρ=0.948 (Tunisia). Notably, the summed sentiment scores are more strongly correlated with either of the two measures than they are with each other for Nicaragua and Serbia (Table B.1). The correlation between annual

[13]Summed sentiment scores correlate with V-Dem's Electoral Democracy Index (EDI) at ρ=-0.081 and with Freedom House values at ρ=-0.113, while the EDI and Freedom House correlate at ρ= 0.939.

https://doi.org/10.33774/apsa-2026-hgrxk **ORCID:** https://orcid.org/0000-0003-1955-281X Content not peer-reviewed by APSA. **License:** All Rights Reserved



Figure 2: Alignment of sentiment with Freedom House and the EDI for eight countries.

https://doi.org/10.33774/apsa-2026-hgrxk **ORCID:** https://orcid.org/0000-0003-1955-281X Content not peer-reviewed by APSA. **License:** All Rights Reserved

changes in the sentiment scores and changes in Freedom House values is also stronger than the correlation between changes in the EDI and Freedom House values for Tunisia (Table B.2).

In large part, the moderate correlations reflect the fact that our outputs characterize the impacts of specific events rather than overall trends in democratic quality. Additionally, because the model disaggregated compound actions into distinct events, episodes with multiple positive or negative developments receive stronger weighting, amplifying swings in the summed score. This is apparent in Belarus and Zimbabwe, for example. Changes in the number of positive and negative events also create fluctuations that are more muted in the indices. We therefore tried three smoothing techniques to reduce the volatility of the raw series: a rolling three-year average, a locally estimated scatterplot smoother (LOESS), and a smoothing spline.[14]

Smoothing consistently improves alignment relative to the unsmoothed baseline. Mean country-level correlations between sentiment levels and Freedom House rise from ρ=0.584 (unsmoothed) to ρ=0.654 (LOESS), and between sentiment levels and V-Dem from ρ=0.479 (unsmoothed) to ρ=0.576 (spline). When using spline smoothing, the most flexible of the three, five of the eight countries in our sample show sentiment correlations with at least one index that surpass the Freedom House–V-Dem benchmark. Figures A.13 and A.14 illustrate differences in the country-level correlations for our eight cases, and summary information about country correlations using the different smoothing techniques is provided in Table B.3. Applying a smoothing technique to the full sample of countries also improves correlations between the sentiment scores and the two indices, as summarized by Tables B.4 and B.5.[15] Country-level correlations between the sentiment scores and either of the two indices is stronger for 55 of 157 countries (35 percent), which increases to nearly 46 percent when using a rolling average.

The correspondence between summed sentiment scores and the indices surrounding known periods of democratization and democratic backsliding is remarkable, as shown by changes in Bangladesh and the closeness of Freedom House scores and the model's assessments for Serbia.

[14]The LOESS span is 0.4; for countries with fewer than ten observations, the span was widened adaptively to ensure a minimum of four local data points in each regression window. The smoothing spline was fit with five degrees of freedom.

[15]In all three cases, countries with insufficient observations to fit the smoother were excluded: the spline required at least five non-missing observations (given df = 5), the LOESS at least four, and the rolling average at least three (given k = 3).

https://doi.org/10.33774/apsa-2026-hgrxk **ORCID:** https://orcid.org/0000-0003-1955-281X Content not peer-reviewed by APSA. **License:** All Rights Reserved

The fact that changes in the values mimic each other at all is also important, considering that one is based on approximated judgments about distinct, qualitative *actions* and the others are highly aggregated indices based on expert answers to specific questions. We see the closeness of the values in these eight cases as very promising for helping to align qualitative depictions of political activity with numbers that are supposed to summarize them. The alignment between our aggregate sentiment scores and the two indices in Figure 2 should be understood not merely as an empirical correlation but as evidence of construct validity: the annotated events produce trends that are visually consistent with expert assessments and enable one to use the register of events to understand changes in those values – particularly in Freedom House scores. This provides convergent evidence from an independent measurement approach. Even so, the sentiment scores are merely ordinal annotations meant to help with filtering and ordering events and not well-founded numerical scores; we discourage users from using them to for quantitative analysis.

Revisiting Democratic Backsliding in Poland and Hungary

Comparing Poland and Hungary illustrates the ability to use event-level data to show systematic, comparable evidence of institutional change across cases using a common evidentiary standard. Both Poland and Hungary were once considered success stories after transitioning from communist rule in 1989, joining NATO in 1999 and the EU in 2004 (Bernhard, 2021). However, economic recession beginning in 2007, and growing dissatisfaction over immigration, created opportunities for outsider parties in both countries to frame the situation as the product of a 'bad deal' (Bernhard, 2021; Krekó & Enyedi, 2018; Muliavka et al., 2025). In Hungary, Fidesz won a two-thirds supermajority in 2010 that enabled it to rewrite the constitution, pack the Constitutional Court, redraw electoral districts, and consolidate control of the media environment (Bánkuti et al., 2012; Jenne & Mudde, 2012; Krekó & Enyedi, 2018). In Poland, the Law and Justice Party (PiS) won a parliamentary majority in 2015 and attempted similar maneuvers – refusing to seat elected judges, legislating procedural constraints on the Constitutional Tribunal, and attempting to force Supreme Court retirements – but faced sustained resistance from judges, civil society, and the international community (Bojarski, 2021; Muliavka et al., 2025; Sadurski, 2019). The PiS lost its Senate majority in 2019 and its Sejm

https://doi.org/10.33774/apsa-2026-hgrxk **ORCID:** https://orcid.org/0000-0003-1955-281X Content not peer-reviewed by APSA. **License:** All Rights Reserved

majority in 2023.

Figure 3 shows the number of positively and negatively rated actions in Freedom House reports for both countries between 1989 and 2023.[16] We also classified the DEED descriptions as plausibly positive or negative and plotted them for comparison.[17] The figure illustrates several differences between cases. First, attacks on democracy were largely absent in Hungary until 2010 and in Poland until 2015. Second, negative actions were more sustained in Hungary, while Poland later saw a surge of positively-rated events. The DEED descriptions also suggest that negative actions became more prominent in Hungary after 2010 and in Poland after 2015, but they do not show steady accumulation in Hungary, nor do they show improvements in Poland.

Among the events in our sample, negative events associated with the judiciary are more prevalent in Poland; in Hungary, negative legislative events predominate, alongside more negative events associated with minority groups and civil society (see Figure A.15). The qualitative research on backsliding in Hungary and Poland underscores that Fidesz's supermajority in the legislature enabled the ruling party to gain control over its constitutional court through its legislative actions – amending and rewriting the constitution –, whereas in Poland its legislative ability to permanently alter courts was more limited. In the first five years of ruling party control (2010-2014), Fidesz curtailed the Constitutional Court's authority over budgets, amended the constitution to prevent future judicial review of constitutional amendments, and stripped the court of authority – all accomplished through legislative action. The PiS, lacking a supermajority, was limited to attempting to stack courts, forcibly retiring justices, and disciplining dissenting judges between 2015 and 2019. Figure 4 confirms this pattern: Poland's backsliding largely involved harassing the courts but using fewer legislative maneuvers, while Hungary's legislative activities were more frequent and severe. Fidesz also used its legislative supermajority to target minority rights and civil society more aggressively than PiS, passing laws that discriminated against Roma and LGBTQ+ communities and restricted NGOs receiving foreign funding. Using information about 'Actors' and 'Targets' in the DEED dataset to identify which events involved the judiciary and legislature, as illustrated in Figure A.16, does not provide such strong evidence of these trends, however.

[16] Neutral/mixed events are omitted.
[17] DEED descriptions were classified using Claude Sonnet 4.6 on 23 March and then manually checked.

https://doi.org/10.33774/apsa-2026-hgrxk **ORCID:** https://orcid.org/0000-0003-1955-281X Content not peer-reviewed by APSA. **License:** All Rights Reserved

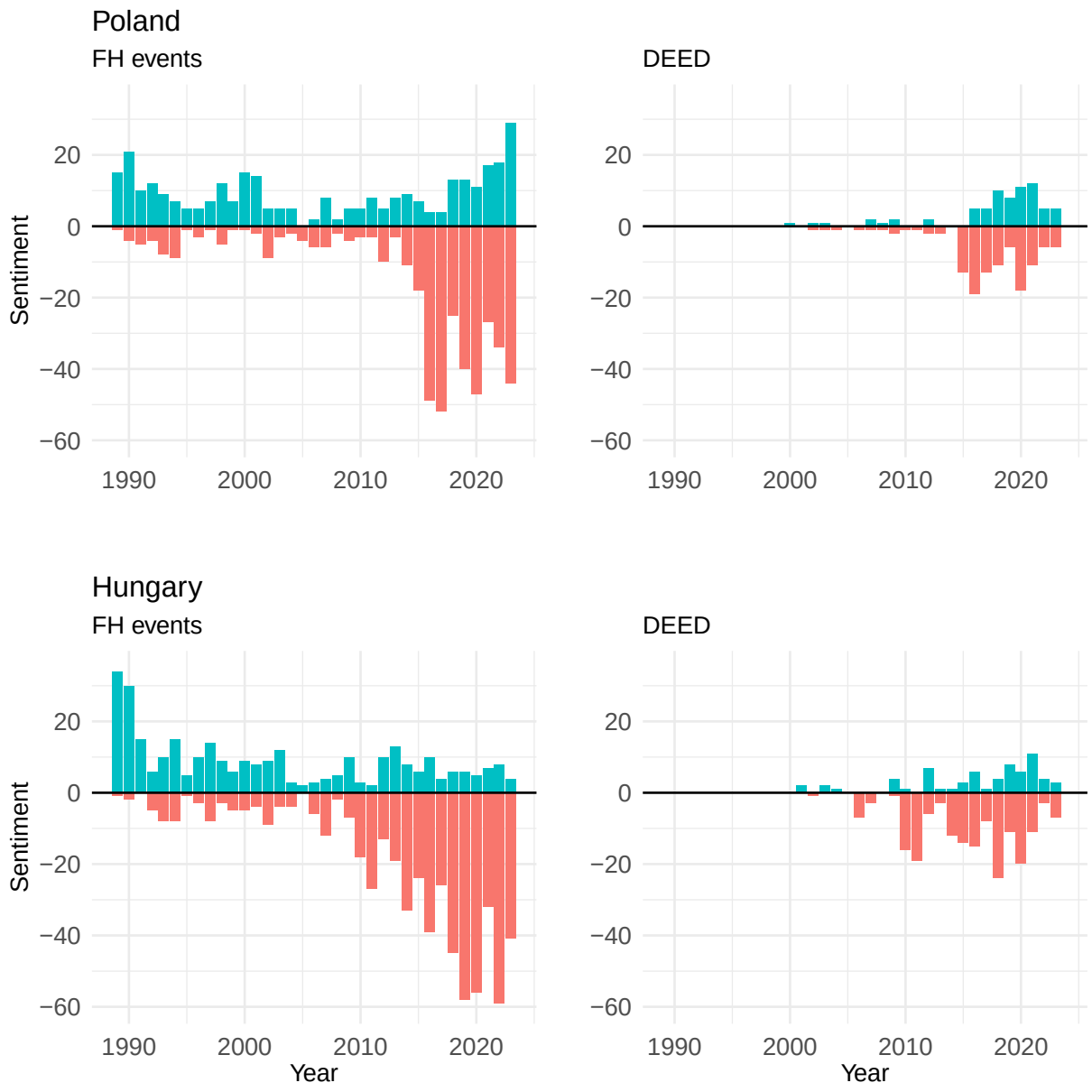


Figure 3: Positive and negative events in Poland and Hungary.
Note: using author interpretation to classify DEED events as positive or negative.

https://doi.org/10.33774/apsa-2026-hgrxk **ORCID:** https://orcid.org/0000-0003-1955-281X Content not peer-reviewed by APSA. **License:** All Rights Reserved

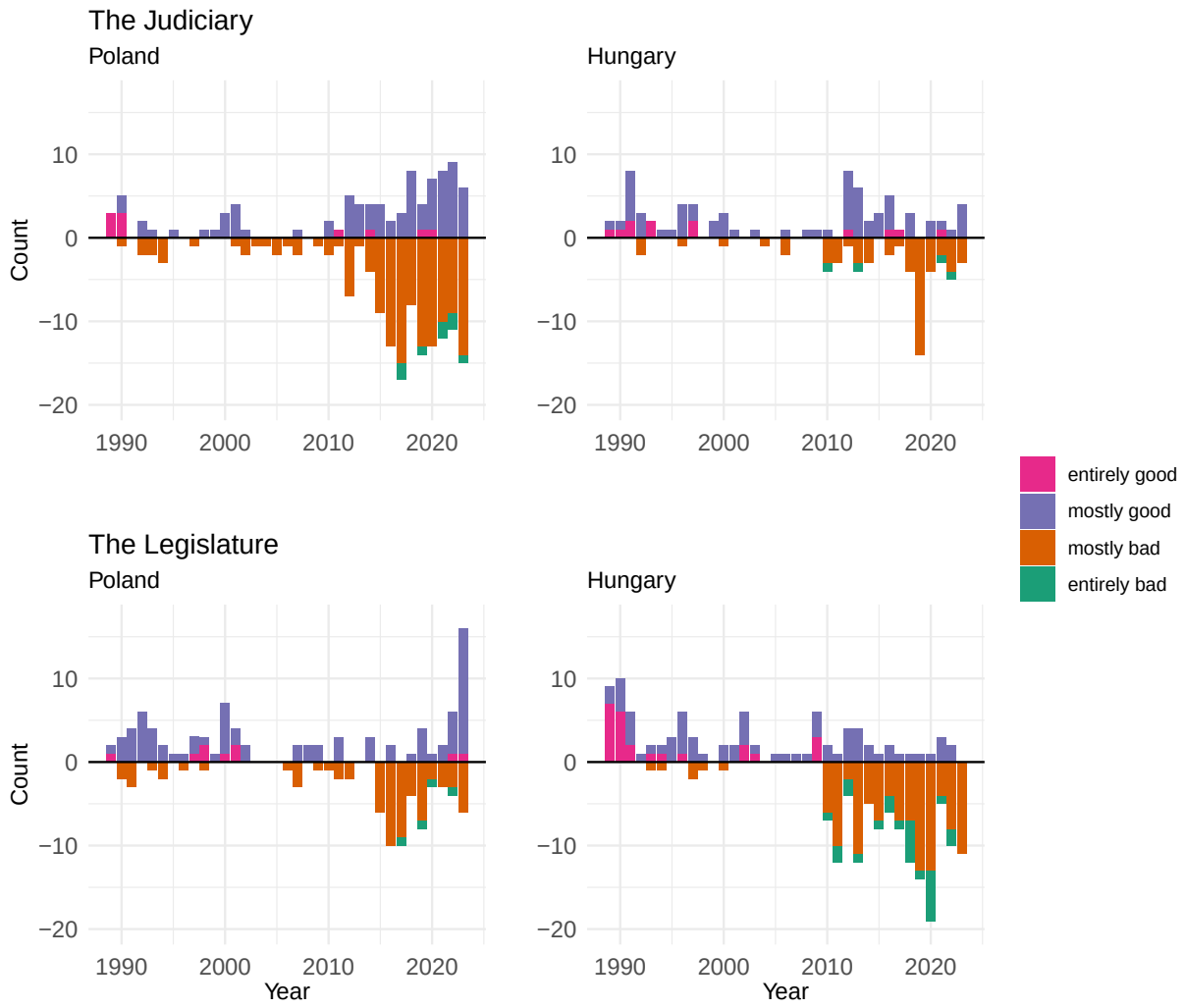


Figure 4: Positive and negative events in Poland and Hungary, isolating events related to the judiciary and the legislature.

https://doi.org/10.33774/apsa-2026-hgrxk **ORCID:** https://orcid.org/0000-0003-1955-281X Content not peer-reviewed by APSA. **License:** All Rights Reserved

The patterns that we show are not simple artifacts of how we constructed the sentiment scores. We demonstrate this by randomly sampling fifty percent of the events that we extracted and annotated for each year and allowing the democratic sentiment score associated with each event to vary by ± 1 point – effectively allowing something judged 'neutral' to be 'mostly good' or 'mostly bad', and something judged 'mostly good' to be 'entirely good' or 'neutral'. We did this 10,000 times for both countries, constructing confidence intervals around the average value for each country-year. Panel A in Figure 5 shows the intervals for Poland and Hungary, with gray boundaries denoting the maximum and minimum score for all samples. Dotted lines mark the lower bound for a 99-percent confidence interval for the first year in the sample. As the figure shows, there are six years where one might judge with confidence a sample of actions taken by the PiS as being significantly less democratic (below α=0.01) compared to the range of likely values in 1989, when both countries were in the midst of negotiating a transition from communist rule. The actions reported by Freedom House were dramatically worse in 2016 and 2017 and in 2019 and 2020 and, to a lesser extent, in 2021 and 2022. In Hungary, the sentiment attributed to illiberal activity has been significantly lower than the 1989 values since 2004. There is only one year (2018) in which activities by the PiS are significantly worse than 2014, the year before they took office; in Hungary, there are four years significantly worse than 2009, with consistently worse activity between 2018 and 2020.

https://doi.org/10.33774/apsa-2026-hgrxk **ORCID:** https://orcid.org/0000-0003-1955-281X Content not peer-reviewed by APSA. **License:** All Rights Reserved

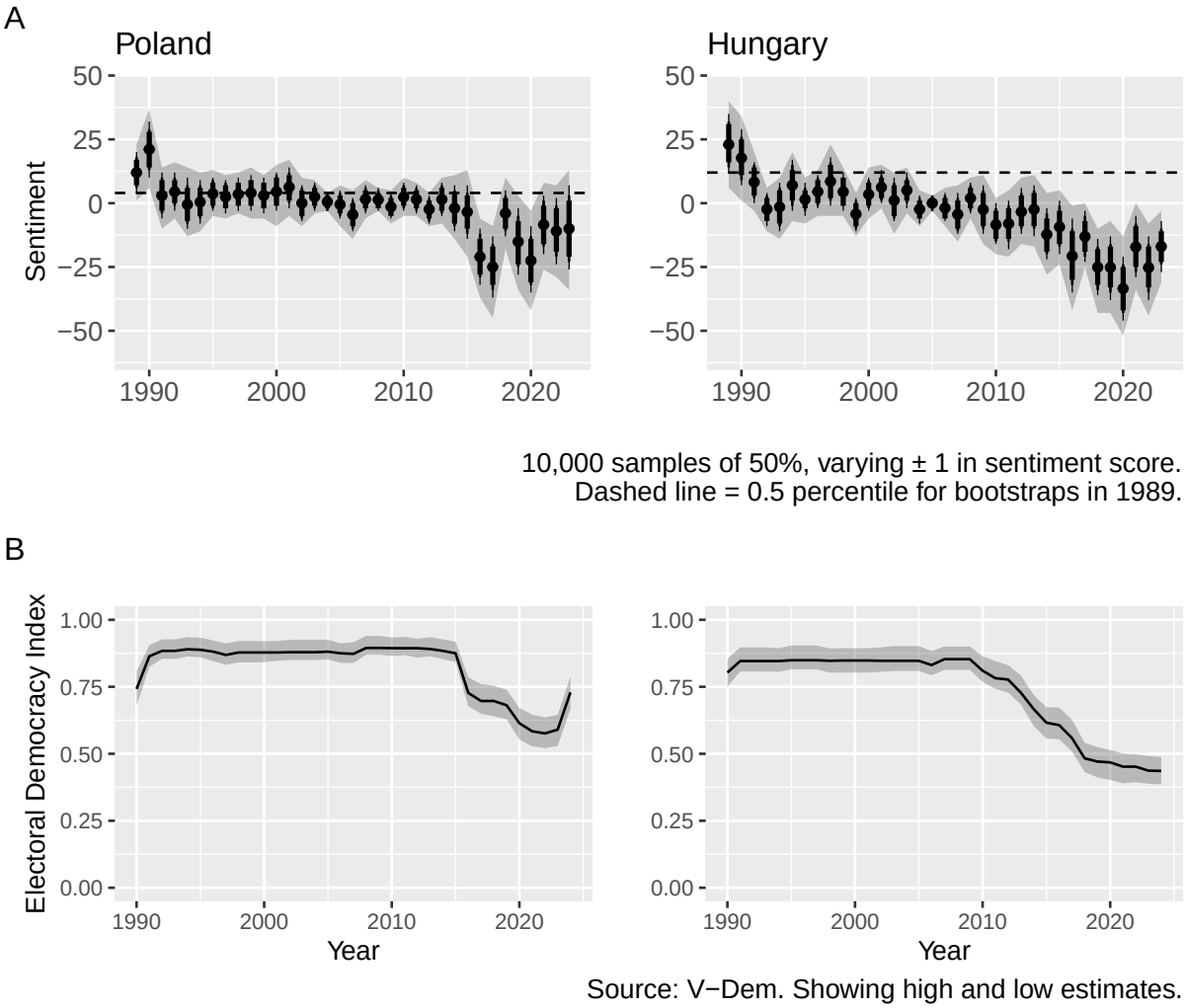


Figure 5: Bootstrapped events (panel A) and EDI (V-Dem; panel B) for Poland and Hungary.

https://doi.org/10.33774/apsa-2026-hgrxk **ORCID:** https://orcid.org/0000-0003-1955-281X Content not peer-reviewed by APSA. **License:** All Rights Reserved

Thus, while the period in which the PiS held a majority of both the Sejm and Senat showed substantial regress, it was not *consistent*, neither was it as slow and steady as the decline under Fidesz. This aligns with Bernhard (2021), who concluded that "[w]hereas it is not unreasonable to talk about a Fidesz dictatorship in Hungary, it would certainly be premature to talk in such terms about Poland" (pg. 605). It also matches roughly what might be concluded from V-Dem's Electoral Democracy Index (panel B). A key difference, however, is that these conclusions can be traced through structured evidence and reflect potential impacts of specific events – not the general state of democracy. This matters directly for concerns raised by Little and Meng (2024), who ask whether indices reflect shifting institutional realities or shifting coder perceptions. More broadly, the comparison illustrates a key methodological contribution of event-level data: the qualitative literature documents that Fidesz consolidated power through the legislature while PiS targeted the judiciary, but that distinction is obscured in summary democracy scores that aggregate across institutional themes – as well as in institution-level indices, as shown in Figure A.17. Demonstrating that our pipeline recovers it without prior selection based on what backsliding should look like helps to demonstrate that the evidentiary layer we produce has validity. The bootstrapped intervals go further, showing that these conclusions are robust to reasonable variation in how individual events are scored – something summary indices, which aggregate expert judgments rather than discrete actions, do not support.

Discussion

The core contribution of this project is in encouraging a shift to using qualitative data to create 'auditable events'. This project demonstrates that LLM-based pipelines can systematically transform qualitative narratives into structured, research-grade event data at scale – a significant methodological advance for political science. This is valuable, insofar as it helps avoid influence from historical observations and instead link conclusions to traceable actions. Democracy measurement has long struggled with compression, as composite scores flatten heterogeneous institutional changes into single numbers without preserving information about actors, targets, mechanisms, or timing. This approach provides middle ground between detailed narratives and aggregate indices. Prospective

https://doi.org/10.33774/apsa-2026-hgrxk **ORCID:** https://orcid.org/0000-0003-1955-281X Content not peer-reviewed by APSA. **License:** All Rights Reserved

events remain auditable, with users able to trace aggregate patterns back to specific actions and verbatim evidence and accept or reject them.

The strategies developed and outlined here offer several improvements that we hope subsequent efforts will build upon. The first is increasing scalability while also maintaining transparency. Using our pipeline approach, we processed 562,022 sentences across 228 countries over 35 years and extracted 198,165 discrete, structured events. Running the final version of our prompts was cost-efficient compared to manual coding at equivalent scale, and our fully documented decisions enables scrutiny of interpretive choices. The second is disaggregation – using multiple prompts with specific tasks helps to decompose compound actions into constituent parts and supports institutional-level analysis. This has important value for downstream analyses such as sequencing analysis, insofar as it can help separate out which institutions are targeted and in what order. It helps to preserve heterogeneity that indices often compress away.

The third major contribution is in its demonstration of consistency and construct validity. We show high stability across runs (87-90 percent sentence agreement, ∼90 percent event agreement for Nicaragua) that is comparable to or exceeds human-human agreement rates. Linking interpretations to verbatim quotations helps to ensure traceability to source material, and structured prompts and deterministic decoding (e.g., temperature=0) minimizes instruction-following failures. Lastly, we show that this is approach produces valid outputs. We see strong alignment with human coders on highlight extraction and a 99-percent match with DEED events that is traceable to Freedom House sources. The sentiment scores associated with events track known patterns and bear out differences in the institutional themes in which backsliding occurs that the qualitative literature depicts in detail, as we show using the examples of Poland and Hungary.

There are nevertheless important limitations to what we present. The approach entails multiple layers of interpretation and is probabilistic by design, meaning perfect replicability is impossible. Our stability checks were conducted within a single production window of approximately one week; Barrie et al. (2025) demonstrate that LM variance compounds over longer time horizons, meaning our figures establish short-run robustness reproducibility as defined by Brodeur et al. (2024) but cannot speak to stability across the delays typical of real-world replication attempts. We did not

https://doi.org/10.33774/apsa-2026-hgrxk **ORCID:** https://orcid.org/0000-0003-1955-281X Content not peer-reviewed by APSA. **License:** All Rights Reserved

version-lock the model, primarily because prompts were developed iteratively over an extended period and the same model version was not guaranteed to remain available throughout. What we document here is therefore replicable only in the narrow sense that the same script, run on the same model version, should produce very similar outputs – which does not perfectly achieve computational reproducibility but comes quite close to doing so Brodeur et al. (2024).[18] We did not compare reproducibility across alternative model families (e.g., GPT-5) or alternative propriety models such as Claude and Gemini or open-source models like LLaMA and Mistral, which represents an altogether different question. Using alternative models to assess the quality of our approach is akin to using different assistants who were not trained the same way and who have different response behaviors. This is nevertheless an important question for future research in this directly to pursue.

There are also remaining technical issues. Digitization errors persist (OCR issues, missing punctuation, spacing problems), and some events have duplicates despite extensive filtering. Edge cases in language will continue to create classification challenges and past and current actions may still be misidentified. The LLM's ability to 'peek' across batches allows context bleeding even when instructed otherwise, and word choice influences democratic sentiment assessment (value-laden language). Lastly, the pipeline does not verify whether events actually occurred, nor does it use alternative sources to cross-validate. This approach can be extended to incorporate multiple sources, thanks in part to increasing digitization of qualitative source material (Fariss et al., 2015). Next steps could also involve technical refinements such as verbalized sampling to show uncertainty in outputs (Zhang et al., 2025).

Despite the areas for improvement given by the current limitations of what we describe and present here, this paper develops some important conceptual parallels to how a similar task might be accomplished with human coders. Many limitations attributed to LLMs also characterize human coding but are less transparently acknowledged. Human inter-coder reliability can be poor and prone to misclassification, and decisions are often difficult to document and hard to reproduce (Brodeur et al., 2024; Edgell et al., 2025; Mikhaylov et al., 2012). A key difference is in producing well-documented decisions that make interpretive choices explicit and auditable, even if not perfectly

[18] Barrie et al. (2025) document how back-end changes to commercial models can silently alter behavior without announcement, rendering even version-specified pipelines non-operational over time.

https://doi.org/10.33774/apsa-2026-hgrxk **ORCID:** https://orcid.org/0000-0003-1955-281X Content not peer-reviewed by APSA. **License:** All Rights Reserved

replicable. Thus, evaluating the potential for LLMs to augment data analysis might be better served by focusing not on whether they can perfectly replicate human judgment but whether they produce defensible interpretations that are transparent, auditable, and fit for purpose. This reorientation is consistent with Bisbee and Spirling (2025), who argue that researchers using LLM-coded data should focus less on replicating results across model versions – since improvements in model performance are unlikely to dramatically alter substantive conclusions – and more on assessing the robustness of findings to potential misclassification rather than correcting outputs against a human benchmark. Our pipeline shows high robustness reproducibility and computational reproducibility for downstream users but not recreate reproducibility (Brodeur et al., 2024).

The process of developing and testing an auditable pipeline using LLMs builds on the extensive work by Baron et al. (2024) to produce event data suitable for evaluating patterns of democratic erosion. In pushing the boundary of accomplishing a goal like this at scale, we hope to see future extensions that correct and improve upon what we accomplished here. Creating a large-scale record of events traceable to rich qualitative materials is important for advancing comparative research and enabling 'creative comparisons' across dynamic political contexts. Such data would facilitate evaluation of the criteria underlying expert-derived scores, and could further serve as a systematic ledger of both democratizing and antidemocratic actions affecting different institutions – actions that may reinforce or offset one another. For example, while the executive branch was undermined by Park Geun-hye's corruption scandal and her removal from office in 2017, the Constitutional Court's upholding of her impeachment that same year demonstrated the resilience of South Korea's institutional checks and balances. Events of this kind offer multiple avenues for structuring qualitative analysis, including process tracing, case matching, and the identification of cases with similar or divergent starting points or sequences. Disaggregated records of descriptive events also support inductive comparative theory-building – particularly valuable for phenomena whose causal structure remains theoretically underspecified – and enable analytical extensions such as using Named Entity Recognition (NER) for actor-based analysis. More broadly, using LLMs to extract structured data from narratives at scale offers a promising way to better bridge quantitative data and analysis with the qualitative evidence that underlies them.

https://doi.org/10.33774/apsa-2026-hgrxk **ORCID:** https://orcid.org/0000-0003-1955-281X Content not peer-reviewed by APSA. **License:** All Rights Reserved

Conclusion

Qualitative data thus remain valuable for grounding and creating a shared understanding of what numerical trends mean (Gerring, 2012). The development of LLMs as general-purpose models, which are particularly good at processing language, can make it easier to work with qualitative data on a large scale but must be carefully considered. In this paper, we show the promise of using LLMs to develop a pipeline that is based on carefully structured prompts, deterministic decoding, verbatim evidence requirements, and decomposed tasks. Working from Freedom House country narratives spanning 228 countries and 35 years, we used a four-stage process – segmented into digitization, text extraction, event extraction, and annotation tasks – that converted 562,022 sentences into 198,165 discrete, annotated event records that are verifiable in the source text. In doing so, we provide a transparent, scalable evidentiary layer that reconnects the numerical summaries that cross-national research depends on with the discrete institutional actions that actually constitute democratic change.

We use several validation checks for a set of test cases to demonstrate that the pipeline produces outputs with strong construct validity. GPT-4o's sentence selections align with human coders at rates comparable to human-human agreement, and in some cases exceed it. The extracted events correspond with 99 percent of human-coded benchmark events traceable to the same Freedom House sources, while yielding substantially greater coverage and finer disaggregation of compound actions. Model-produced sentiment scores also track known trajectories of democratization and backsliding in ways that mirror both Freedom House ratings and V-Dem's Electoral Democracy Index, even if they are not strongly correlated. Using Poland and Hungary as examples, the output show that democratic erosion in Hungary concentrated in the legislature while Poland's backsliding focused on judicial institutions – a distinction noted by qualitative scholars that is not as easily shown in aggregate scores. Bootstrapped uncertainty intervals show that the conclusion that Hungary progressively backslid while Poland did not holds across thousands of resamples and reasonable variations in how individual events are scored. That the pipeline recovers this pattern from structured event data, rather than narrative interpretation, demonstrates that it can help uncover institutionally specific mechanisms of backsliding.

https://doi.org/10.33774/apsa-2026-hgrxk **ORCID:** https://orcid.org/0000-0003-1955-281X Content not peer-reviewed by APSA. **License:** All Rights Reserved

There are nonetheless substantial limitations to what we present, and we do not claim to have solved the measurement problem that motivates this project. We readily acknowledge that our use of GPT-4o relies on a proprietary black box that changes over time. What we document here is only 'replicable' in a narrow sense: it requires rerunning the same script on the same version. This is version control, not full replication. The design of LLMs also means that perfect replicability is nearly impossible. However, the fully documented pipeline with all prompts provided and decisions made transparent is intended to encourage the next iteration of this approach.

At the same time, scrutinizing the ability of LLMs as a measurement tool to perfectly match the subjective assessments of human coders overlooks its potential to produce data that humans *could* but are unable to produce at scale. Many of these issues also characterize human coding but are less transparently acknowledged. One way to go forward, therefore, is to focus on making interpretive choices explicit and auditable, if not perfectly reproducible. Even in the face of skepticism about which events matter and how serious are their implications for democracy, data of this type enable one to scrutinize individual events and interpretation about them and to incorporate uncertainty. If replicated, the approach that we described here could yield a different set of empirical conclusions. Though LLMs are fragile and non-exact (Barrie et al., 2025), the results that we describe suggest ways to structure them and treat their outputs as approximations of human-coded data.

To that end, this project elucidates one way forward for improving on retrieving and understanding the events that underlie expert judgments about democratic developments that numerical indices do not clearly depict. The pipeline for using LLMs should be understood as a complement to, rather than a replacement for, the qualitative and quantitative traditions it sits between. Event data like this cannot exist without qualitative work and are also not as sophisticated as quantitative measures. LLMs may never be able to automate analysis of nuance like humans can, but they can be a useful tool for creating outputs for analysis. Carefully applied, it can bridge the gap between abstract judgment and the real-world events that inform it by focusing on the sort of concrete actions that Little and Meng (2024) call for as evidence of backsliding. Where backsliding consists of 'death by a thousand cuts' (Bermeo, 2016), there may be no one action or institutional theme that one can point to. If so, understanding it requires cataloging in detail the incremental actions and events that constitute democratic decline.

https://doi.org/10.33774/apsa-2026-hgrxk **ORCID:** https://orcid.org/0000-0003-1955-281X Content not peer-reviewed by APSA. **License:** All Rights Reserved

References

Alizadeh, M., Kubli, M., Samei, Z., Dehghani, S., Zahedivafa, M., Bermeo, J. D., Korobeynikova, M., & Gilardi, F. (2024). Open-source LLMs for text annotation: A practical guide for model setting and fine-tuning. *Journal of Computational Social Science*, *8*(1), 17. https://doi.org/10.1007/s42001-024-00345-9

Bánkuti, M., Halmai, G., & Scheppele, K. L. (2012). Hungary's illiberal turn: Disabling the constitution. *Journal of Democracy*, *23*(3), 138–146. https://doi.org/10.1353/jod.2012.0054

Baron, H., Blair, R. A., Gottlieb, J., & Paler, L. (2024). An events-based approach to understanding democratic erosion. *PS: Political Science & Politics*, *57*(2), 208–215. https://doi.org/10.1017/S1049096523001026

Barrie, C., Palmer, A., & Spirling, A. (2025, November). *Replication for language models: Problems, principles, and best practice for political science* (tech. rep.). Working Paper. https://arthurspirling.org/documents/BarriePalmerSpirling_TrustMeBro.pdf

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922

Bermeo, N. (2016). On democratic backsliding. *Journal of Democracy*, *27*(1), 5–19.

Bernhard, M. (2021). Democratic backsliding in poland and hungary. *Slavic Review*, *80*(3), 585–607. https://doi.org/10.1017/slr.2021.145

Bisbee, J., Clinton, J. D., Dorff, C., Kenkel, B., & Larson, J. M. (2024). Synthetic Replacements for Human Survey Data? The Perils of Large Language Models. *Political Analysis*, *32*(4), 401–416. https://doi.org/10.1017/pan.2024.5

Bisbee, J., & Spirling, A. (2025). *What to do when humans are no longer the gold standard: Large language models, state of the art and robustness* [Working paper. First draft: June 13, 2025. This draft: July 7, 2025].

https://doi.org/10.33774/apsa-2026-hgrxk **ORCID:** https://orcid.org/0000-0003-1955-281X Content not peer-reviewed by APSA. **License:** All Rights Reserved

Bojarski, Ł. (2021). Civil society organizations for and with the courts and judges—struggle for the rule of law and judicial independence: The case of poland 1976–2020. *German Law Journal*, *22*(8), 1344–1384. https://doi.org/10.1017/glj.2021.72

Bowman, S. R. (2023, April). Eight Things to Know about Large Language Models. https://doi.org/10.48550/arXiv.2304.00612

Brodeur, A., Esterling, K., Ankel-Peters, J., Bueno, N. S., Desposato, S., Dreber, A., Genovese, F., Green, D. P., Hepplewhite, M., de la Guardia, F. H., Johannesson, M., Kotsadam, A., Miguel, E., Velez, Y. R., & Young, L. (2024). Promoting reproducibility and replicability in political science. *Research & Politics*, *11*(1). https://doi.org/10.1177/20531680241233439

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*.

Chen, J., de Jong, K., Poole, A., Burakowski, J., Elderson Nosti, E., Windt, J., & Wang, C. (2026). Uncovering political bias in large language models using parliamentary voting records [arXiv:2601.08785]. *Proceedings of the 2026 ACM XXX Conference*.

Chenoweth, E., & Lewis, O. A. (2013). Unpacking nonviolent campaigns: Introducing the NAVCO 2.0 dataset. *Journal of Peace Research*, *50*(3), 415–423.

Coppedge, M., Gerring, J., Altman, D., Bernhard, M., Fish, S., Hicken, A., Kroenig, M., Lindberg, S. I., McMann, K., Paxton, P., Semetko, H. A., Skaaning, S.-E., Staton, J., & Teorell, J. (2011). Conceptualizing and measuring democracy: A new approach. *Perspectives on Politics*, *9*(2), 247–267.

Coppedge, M., Gerring, J., Knutsen, C. H., Lindberg, S. I., Teorell, J., Altman, D., Angiolillo, F., Bernhard, M., Cornell, A., Fish, M. S., Fox, L., Gastaldi, L., Gjerløw, H., Glynn, A., Good God, A. F., Grahn, S., Hicken, A., Kinzelbach, K., Marquardt, K. L., … Ziblatt, D. (2025). V-dem codebook v15 [Version 15, March 2025]. https://www.v-dem.net/data/reference-documents/

Diamond, L. (2015). Facing up to the democratic recession. *Journal of Democracy*, *26*(1), 141–155.

Edgell, A. B., Lachapelle, J., & Maerz, S. F. (2025). Achieving transparency, traceability, and readability with human-coded data. *PS: Political Science & Politics*, *58*(2), 346–351. https://doi.org/10.1017/S1049096524000714

https://doi.org/10.33774/apsa-2026-hgrxk **ORCID:** https://orcid.org/0000-0003-1955-281X Content not peer-reviewed by APSA. **License:** All Rights Reserved

Fariss, C. J., Linder, F. J., Jones, Z. M., Crabtree, C. D., Biek, M. A., Ross, A.-S. M., Kaur, T., & Tsai, M. (2015). Human rights texts: Converting human rights primary source documents into data. *PLOS ONE*, *10*(9), 1–19. https://doi.org/10.1371/journal.pone.0138935

Gerring, J. (2012). Mere Description. *British Journal of Political Science*, *42*(4), 721–746.

Giannone, D. (2010). Political and ideological aspects in the measurement of democracy: The freedom house case. *Democratization*, *17*(1), 68–97.

Gilardi, F., Alizadeh, M., & Kubli, M. (2023). Chatgpt outperforms crowd-workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, *120*(30), e2305016120. https://doi.org/10.1073/pnas.2305016120

Grimmer, J., Roberts, M. E., & Stewart, B. M. (2022). *Text as data: A new framework for machine learning and the social sciences*. Princeton University Press.

Haggard, S., & Kaufman, R. (2021). *Backsliding: Democratic regress in the contemporary world*. Cambridge University Press.

Halterman, A., & Keith, K. A. (2025). Codebook llms: Evaluating llms as measurement tools for political science concepts. *Political Analysis*, 1–17. https://doi.org/10.1017/pan.2025.10017

Heseltine, M., & Clemm von Hohenberg, B. (2024). Large language models as a substitute for human experts in annotating political text. *Research & Politics*, *11*(1), 1–10. https://doi.org/10.1177/20531680241236239

Jenne, E. K., & Mudde, C. (2012). Hungary's illiberal turn: Can outsiders help? *Journal of Democracy*, *23*(3), 147–155. https://doi.org/10.1353/jod.2012.0057

Kalyvas, S. N., & Balcells, L. (2010). International system and technologies of rebellion: How the end of the cold war shaped internal conflict. *The American Political Science Review*, *104*(3), 415–429. Retrieved December 18, 2025, from http://www.jstor.org/stable/40863761

Karatnycky, A. (2002). The 2001 freedom house survey: What's behind the scores? *Journal of Democracy*, *13*(1), 99–110.

Knutsen, C. H., Marquardt, K. L., Seim, B., Coppedge, M., Medzihorsky, J., Edgell, A. B., Pemstein, D., Teorell, J., Gerring, J., & Lindberg, S. I. (2024). Conceptual and measurement issues in assessing democratic backsliding. *PS: Political Science & Politics*, *57*(2), 162–177.

Krekó, P., & Enyedi, Z. (2018). Explaining eastern europe: Orbán's laboratory of illiberalism. *Journal of Democracy*, *29*(3), 39–51. https://doi.org/10.1353/jod.2018.0043

https://doi.org/10.33774/apsa-2026-hgrxk **ORCID:** https://orcid.org/0000-0003-1955-281X Content not peer-reviewed by APSA. **License:** All Rights Reserved

LaFree, G., & Dugan, L. (2007). Introducing the global terrorism database. *Terrorism and Political Violence*, *19*(2), 181–204.

Le Mens, G., & Gallego, A. (2025). Positioning political texts with large language models by asking and averaging. *Political Analysis*, *33*(3), 274–282. https://doi.org/10.1017/pan.2024.29

Levitsky, S., & Ziblatt, D. (2018). *How democracies die*. Crown.

Little, A. T., & Meng, A. L. (2024). Measuring democratic backsliding. *PS: Political Science & Politics*, *57*(2), 149–161.

Liu, W., Xiang, M., & Ding, N. (2025). Active use of latent tree-structured sentence representation in humans and large language models. *Nature Human Behaviour*. https://doi.org/10.1038/s41562-025-02297-0

Mellon, J., Bailey, J., Scott, R., Breckwoldt, J., Miori, M., & Schmedeman, P. (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), 1–7. https://doi.org/10.1177/20531680241231468

Mikhaylov, S., Laver, M., & Benoit, K. R. (2012). Coder reliability and misclassification in the human coding of party manifestos. *Political Analysis*, *20*(1), 78–91. https://doi.org/10.1093/pan/mpr047

Muliavka, V., Kwiatkowska, A., & Plisiecki, H. (2025). In defence of democracy: Parliamentary opposition and democratic resilience in poland. *East European Politics*, *41*(1), 117–141. https://doi.org/10.1080/21599165.2024.2415633

Munck, G. L., & Verkuilen, J. (2002). Conceptualizing and measuring democracy: Evaluating alternative indices. *Comparative Political Studies*, *35*(1), 5–34.

O'Brochta, W. (2025). How human coding is used and described. *Political Studies Review*. https://doi.org/10.1177/14789299251395515

Potter, Y., Lai, S., Kim, J., Evans, J., & Song, D. (2024). Hidden persuaders: Llms' political leaning and their influence on voters [arXiv:2410.24190v3 [cs.CL]]. *arXiv preprint arXiv:2410.24190*.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Technical Report*. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

https://doi.org/10.33774/apsa-2026-hgrxk **ORCID:** https://orcid.org/0000-0003-1955-281X Content not peer-reviewed by APSA. **License:** All Rights Reserved

Raleigh, C., Linke, A., Hegre, H., & Karlsen, J. (2010). Introducing ACLED: An armed conflict location and event data project. *Journal of Peace Research*, *47*(5), 651–660.

Rytting, C. M., Sorensen, T., Argyle, L., Busby, E., Fulda, N., Gubler, J., & Wingate, D. (2023, June). Towards Coding Social Science Datasets with Language Models. https://doi.org/10.48550/arXiv.2306.02177

Sadurski, W. (2019). Polish constitutional tribunal under pis: From an activist court, to a paralysed tribunal, to a governmental enabler. *Hague Journal on the Rule of Law*, *11*, 63–84. https://doi.org/10.1007/s40803-018-0078-1

Törnberg, P. (2025). Large language models outperform expert coders and supervised classifiers at annotating political social media messages. *Social Science Computer Review*, *43*(6), 1181–1195. https://doi.org/10.1177/08944393241286471

Treisman, D. (2024). Psychological biases and democratic anxiety: A comment on little and meng (2023). *PS: Political Science & Politics*, *57*(2), 194–197.

Waldner, D., & Lust, E. (2018). Unwelcome change: Coming to terms with democratic backsliding. *Annual Review of Political Science*, *21*, 93–113.

Weidmann, N. B., Faulborn, M., & García, D. (2026). Large language models are democracy coders with attitudes. *PS: Political Science & Politics*, *59*(1), 17–23. https://doi.org/10.1017/S1049096525101248

White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., & Schmidt, D. C. (2023, February). A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT. https://doi.org/10.48550/arXiv.2302.11382

Woldense, J., & Kroeger, A. (2024). Elite change without regime change: Authoritarian persistence in africa and the end of the cold war. *American Political Science Review*, *118*(1), 178–194. https://doi.org/10.1017/S0003055423000151

Wu, P. Y. (2025, June). Large Language Models Can Be a Viable Substitute for Expert Political Surveys When a Shock Disrupts Traditional Measurement Approaches. https://doi.org/10.48550/arXiv.2506.06540

Zhang, J., Yu, S., Chong, D., Sicilia, A., Tomz, M. R., Manning, C. D., & Shi, W. (2025). Verbalized sampling: How to mitigate mode collapse and unlock llm diversity. *arXiv preprint arXiv:2510.01171*. https://arxiv.org/abs/2510.01171

https://doi.org/10.33774/apsa-2026-hgrxk **ORCID:** https://orcid.org/0000-0003-1955-281X Content not peer-reviewed by APSA. **License:** All Rights Reserved

Zhao, C., Tan, Z., Ma, P., Li, D., Jiang, B., Wang, Y., Yang, Y., & Liu, H. (2025, August). Is Chain-of-Thought Reasoning of LLMs a Mirage? A Data Distribution Lens. https://doi.org/10.48550/arXiv.2508.01191

https://doi.org/10.33774/apsa-2026-hgrxk **ORCID:** https://orcid.org/0000-0003-1955-281X Content not peer-reviewed by APSA. **License:** All Rights Reserved