COMPUTATIONAL SOCIAL SCIENCE

The Media Bias Detector: A framework for annotating and analyzing the news

Samar Haider[1]†*, Amir Tohidi[1]†, Jenny S. Wang[2]†, Timothy Dörr[3], David M. Rothschild[4], Chris Callison-Burch[1], Duncan J. Watts[1,3,5]

News organizations introduce bias into their coverage via the choices they make about which topics to cover (or ignore) and how to frame the issues they do decide to cover. Here, we introduce the Media Bias Detector, a scalable computational framework that integrates large language models (LLMs) with near-real-time news scraping to extract structured annotations, including political lean, tone, topics, article type, and major events, across hundreds of articles per day. We quantify these dimensions of coverage at the sentence level, the article level, and the publisher level, expanding the ways in which researchers can analyze selection and framing bias in the modern news landscape. We also release an interactive web platform for convenient exploration of these data and an accompanying dataset covering more than 140,000 articles published in 2024 by 10 prominent publishers. Last, we present some results derived from this dataset that illustrate how the MBD can uncover correlates of bias in news coverage.

INTRODUCTION

The importance of the media in disseminating information and shaping public discourse is undeniable. As the primary vehicle for delivering news and framing societal debate, the media influences what people think about and how they think about it (*1*, *2*). Even in an era in which U.S. adults often get their news from social media and mobile apps, media organizations remain important. News websites or apps are preferred by 23% of adults versus 18% who favor social media (*3*), and most links that are shared on social media tend to originate from mainstream media sources (*4*). Given its reach and influence, the reliability of information produced by news organizations continues to be an important concern. In contrast to social media, mainstream media organizations typically adhere to professional journalistic standards of fact-checking and accuracy, and hence tend not to be the primary sources of what is often referred to as "misinformation" or "fake news." Nonetheless, the information they communicate to the public can still be biased in at least two ways that have long been studied by communication and media scholars (*5*–*7*).

First, "selection bias" holds that media organizations actively choose which topics, events, issues, and perspectives to cover as well as how prominently or extensively to cover them (*8*). For example, although Hillary Clinton's use of a private email server while serving as U.S. Secretary of State was arguably a matter of public interest that ought to have been covered by the media in the lead-up to the 2016 presidential election, the appearance of 10 front-page articles in The New York Times in the 8 days leading up to the election represented an editorial choice to focus on this one issue over all others (*9*). More broadly, the attention that media organizations pay to topics such as crime or health risks has been shown to be unrelated to, or even inversely related to, the corresponding rates expressed by objective administrative data (*10*–*12*), suggesting again that the extent of coverage of these issues represents a choice by news organizations themselves rather than a reflection of external reality. In both cases, the information contained in the stories that are published might be entirely factually accurate and may even be presented in a neutral, objective manner. Nonetheless, as theories of "agenda setting" have long held, selection bias can still distort public opinion by driving attention to some issues at the cost of others that might be equally or more important (*1*, *13*). Moreover, because the counterfactual issues are, by definition, not being paid attention to, the distorting influence of selection bias can be hard to detect.

Second, "framing bias" contends that, conditional on having selected a topic to cover, journalists and editors can exploit additional degrees of freedom in how they present the constituent facts and issues, again without necessarily saying anything that would fail a fact check (*14*). Framing bias can involve choices about the inclusion or omission of specific details, perspectives or narratives, the tone or language used, the context provided, and the inclusion or omission of background information (*2*, *11*, *15*). Such decisions can lead a reader to reach a desired conclusion that, while not necessarily false, is only one of potentially many equally true, and potentially qualitatively different, conclusions. For example, news organizations are often accused of covering elections through the frame of sporting contests (also known as "horse-race coverage"), which emphasizes the popularity of candidates, rather than through the frame of policy differences and other issues that are likely to affect voters post-election (*9*, *16*). As with selection bias, framing bias can be subtle and hard to detect from the available data. For example, "false equivalence" or "bothsidesism" is a type of framing bias in which manifestly unequal views are presented in a balanced and evenhanded manner (*17*, *18*). In such cases, the bias can be difficult to identify because it is actively masquerading as neutrality.

As the media landscape continues to expand and simultaneously fragment, these potential sources of media bias and their attendant consequences for public understanding of controversial issues make the development of scalable tools to measure bias more critical than ever. However, the methods that have traditionally been used to study bias were developed when news cycles were slower and media consumption was more centralized, allowing researchers to manually analyze coverage patterns across a small set of publications. For

[1]Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA, USA. [2]Harvard Business School, Boston, MA, USA. [3]Annenberg School for Communication, University of Pennsylvania, Philadelphia, PA, USA. [4]Microsoft Research, New York, NY, USA. [5]Department of Operations, Information, and Decisions, University of Pennsylvania, Philadelphia, PA, USA.
*Corresponding author. Email: samarh@engineering.upenn.edu
†These authors contributed equally to this work.

example, traditional content analysis relies on manual annotation (*19*–*21*), which has difficulty scaling even to thousands of articles. Meanwhile, computational techniques such as keyword frequency analysis and topic modeling (*22*–*24*) scale much better, but struggle to capture the contextual nuances necessary for identifying complex forms of bias (*25*). Last, commercial bias-rating tools such as AllSides, Ad Fontes, and NewsGuard measure bias by assigning static, publisher-level ratings along a one-dimensional left-right spectrum (*26*–*28*). Although useful, these classifications necessarily overlook within-publisher variation across different topics or events or how these patterns evolve over time, all of which are questions of potential interest to researchers.

Recent advancements in large language models (LLMs) present new opportunities to move beyond these limitations (*29*, *30*). Unlike traditional natural language processing methods, LLMs can capture nuanced semantic relationships within text (*31*, *32*). By leveraging vast training corpora that encompass a wide variety of published content, these models can identify subtle variations in tone, emphasis, and narrative structure that shape media bias (*33*). Although the resulting ratings remain imperfect and require careful human validation, especially as news cycles shift (*34*, *35*), carefully administered LLMs can now generate human-quality ratings for even very complex and subtle judgments (*36*, *37*). Equally important, although the cost of using commercial LLMs still presents some scaling limitations, it is feasible to rate hundreds to thousands of articles per day, allowing for near-real-time, article-level classification across multiple publishers, and we expect costs to continue to fall.

Here, we introduce the Media Bias Detector, a scalable, near-real-time system that measures elements of framing and selection bias at multiple levels, from individual articles to publisher-wide trends. The Media Bias Detector continuously collects clean and comprehensive data from a set of major news publishers across the political spectrum. We also describe a unique dataset derived from the Media Bias Detector and illustrate how these data can be used to analyze media coverage during a critical election year. Here, we focus on data collected during 2024 for the 10 publishers included in the initial launch of the Media Bias Detector; however, we note that our set of publishers has grown (currently to 21) as has the time period of our data, and we anticipate both will continue to grow over time. Last, we make the Media Bias Detector methodology and data publicly available to support academic research on media bias, and also provide an interactive online tool (https://mediabiasdetector.seas.upenn.edu) for users to explore and analyze coverage patterns dynamically (*38*). Figure 1 illustrates the core design principles of our framework and shows screenshots of our dashboard.

At the core of our methodology is a multistage pipeline that systematically tracks and analyzes news articles. Although television and cable news also play substantial roles in shaping public discourse, we currently focus on text-based digital news where the unit of analysis is each individual news article. We capture homepage snapshots from major publishers at regular intervals, recording article content, as well as metadata on homepage placement and prominence to measure selection bias. Using LLMs, we classify articles by topic, political lean, tone, and news type. In addition, we break our analysis down to the sentence level, capturing nuanced aspects of framing bias such as sentence type and tone. We also perform event clustering to group articles together based on their content on a daily basis and track which events received more coverage than others, and by which publishers.

While existing media bias ratings organizations judge media outlets as a whole, placing them on a discrete left-right spectrum based on the entirety of their reporting as well as any political alignment their pundits or owners may have expressed, our approach evaluates each article independently and without knowledge of the publisher or their history. To analyze articles, we use state-of-the-art LLMs such as GPT-4o to extract a large set of simple labels. By focusing on labels that are tractable for current models (such as the topic or tone of a news article) and by validating them with human raters (see below), we ensure a high degree of confidence in the outputs of our framework and the resulting dataset. In addition to extracting numerical labels, we also elicit natural language analyses for each article from the model to summarize its content and aid human oversight. Last, we allow users to freely combine our data in various ways so they can use it to answer a variety of research questions about the media. This flexibility is facilitated by the variety of labels we extract for each news article and the number of news articles we have collected in our dataset.

To ensure accuracy and transparency, we incorporate systematic human oversight at multiple stages. Trained annotators regularly validate a random sample of the model's outputs every day, correcting biases and ensuring consistency in the LLM's classification. This human-in-the-loop approach balances scalability with quality control, addressing key limitations of both fully manual and fully automated methods. We also conducted an in-depth data validation exercise to benchmark the effectiveness of LLMs at complex tasks such as identifying political lean and tone from news articles. Furthermore, our methodology is transparent and publicly documented, with detailed information about our LLM prompting strategies and validation procedures available to researchers who use our system.

We argue that the Media Bias Detector enables new avenues of research into media bias that were previously difficult or impossible to study beyond the scale of a handful of articles. As we describe in Results, our data reveal interesting patterns in contemporary news coverage and offer relatively large-scale, quantitative evidence to earlier qualitative claims of bias. In addition, the scale and granularity of our data enable deeper investigation of these research questions, revealing nuanced patterns and generating insights that extend beyond prior research findings. For example, in coverage of political topics we find a systematic tendency for media outlets to focus more on criticizing their ideological opposition than supporting aligned positions, extending previous research suggesting that partisan media effects are primarily driven by negative coverage of opposition candidates (*39*). In another example, we find that headlines often present issues in a manner that leans more Republican than the associated article text, complementing recent research showing that biased headlines can substantially affect readers' ability to make inferences and recall facts from news stories (*40*). In a third example, our data replicate and extend prior claims of negativity bias in the media, which can in turn shape how the public perceives the reality of the world.

The remainder of the paper is structured as follows: First, we provide a detailed explanation of our system for real-time analysis of media bias, which combines the processing power of LLMs with systematic human oversight. In this section, we also present evidence validating our methodology through extensive testing, demonstrating high accuracy in topic classification, event clustering, and labeling article lean and tone using LLMs. Second, we use our curated dataset to explore a range of questions related to media practices and biases, showcasing the potential use cases of our dataset for

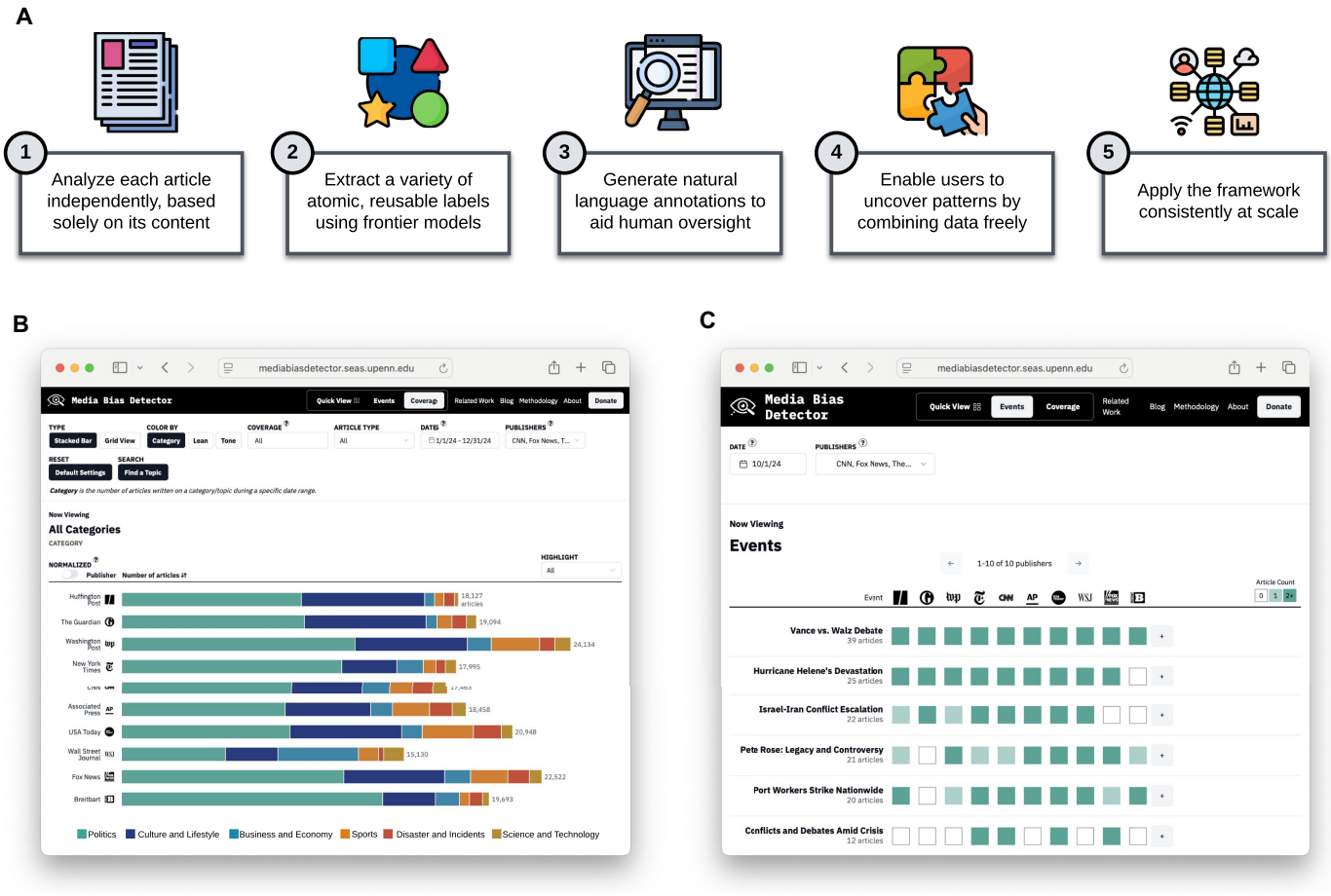

Fig. 1. The Media Bias Detector design philosophy. (**A**) The core guiding principles behind the design of our framework. We analyze each article individually to allow direct, data-driven comparisons between publishers regarding their selection and framing of news topics. By extracting simple labels using state-of-the-art LLMs, we aim to ensure a high level of reliability in our dataset. As we have a variety of data points for each article, our framework allows users to freely combine them to gain new insights into the media. (**B**) The coverage page gives users a long-term view of what topics the media chooses to cover. Users can set filters to narrow the time frame and click through on each news topic to compare the distribution of subtopics within it. (**C**) The event page, on the other hand, offers a more immediate view of the daily news cycle and shows the top news events of the past day along with the amount of attention given to them by each publisher.

gaining deeper insights into media. Last, we discuss the practical implications of our work and outline how other researchers can build on our transparent methodology to advance the study of media bias.

MATERIALS AND METHODS

System architecture

As already noted, our multistage data pipeline is built upon LLMs, which we use for both their capacity to produce deep contextual representations of text documents and their instruction-tuned ability to generate labels without requiring extensive training data.

Prior work has found that LLMs perform well in generative tasks such as summarization (*41*, *42*), as well as discriminative tasks such as sentiment analysis and topic classification (*43*, *44*), where these tasks can even be performed in a zero-shot setting (*45*). These findings suggest that LLMs can be well-suited for complex content analysis tasks like media bias detection, where identifying patterns in context, tone, and subtle language cues is essential. Critically, LLMs also scale effectively to large volumes of content that would be unmanageable with manual annotation alone (*37*). As a result, LLMs can competently and efficiently extract detailed information such as sentence-level labels and classifications by topic, subtopic, article type, tone, and political lean at the scale of many publishers in near-real-time (*46*).

The Media Bias Detector uses LLMs in two ways: First, for labeling news articles and sentences using zero-shot prompting; and second, for generating contextual embeddings to cluster articles about major events and highlight important facts within them. Figure 2 provides a high-level overview of our pipeline, which comprises three primary components: data collection, data labeling, and data validation. In the following sections, we discuss each of these in detail.

Data collection

Our dashboard currently covers 21 major news publishers chosen to capture a mix of audience size (see Fig. 3), agenda-setting influence, and ideological diversity: Associated Press, Breitbart News, CNN, Fox News, HuffPost, The Guardian, The New York Times, The Wall Street Journal, The Washington Post, USA Today, Los Angeles Times, Chicago Tribune, Star Tribune, BBC, Financial Times, Reuters, Bloomberg, CNBC, New York Post, Newsmax, and The Daily Wire.

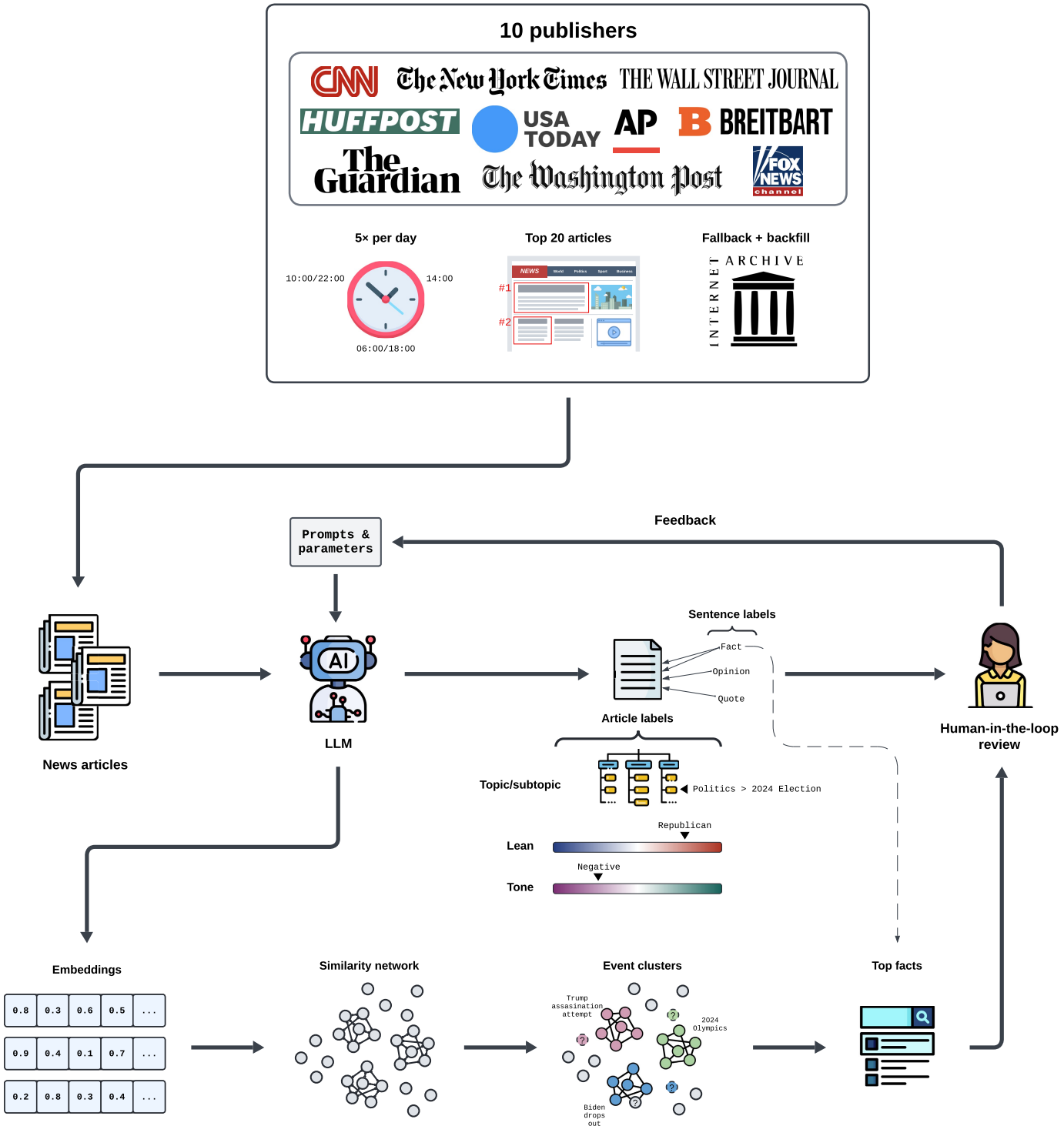

Fig. 2. The Media Bias Detector data pipeline. For each of the 10 publishers, we take a snapshot of their homepage five times per day and scrape the top-ranking news articles. We then use LLMs to extract multiple levels of structured labels at the article level (topic, subtopic, lean, tone, and type) and the sentence level (type, tone, and focus). In addition, we obtain vector embeddings for all articles to cluster them based on the events they focus on. We then cluster all sentences from articles about a particular event to highlight the top facts surrounding it. Our regular human-in-the-loop process adds oversight to this framework and ensures that the generated labels are accurate.

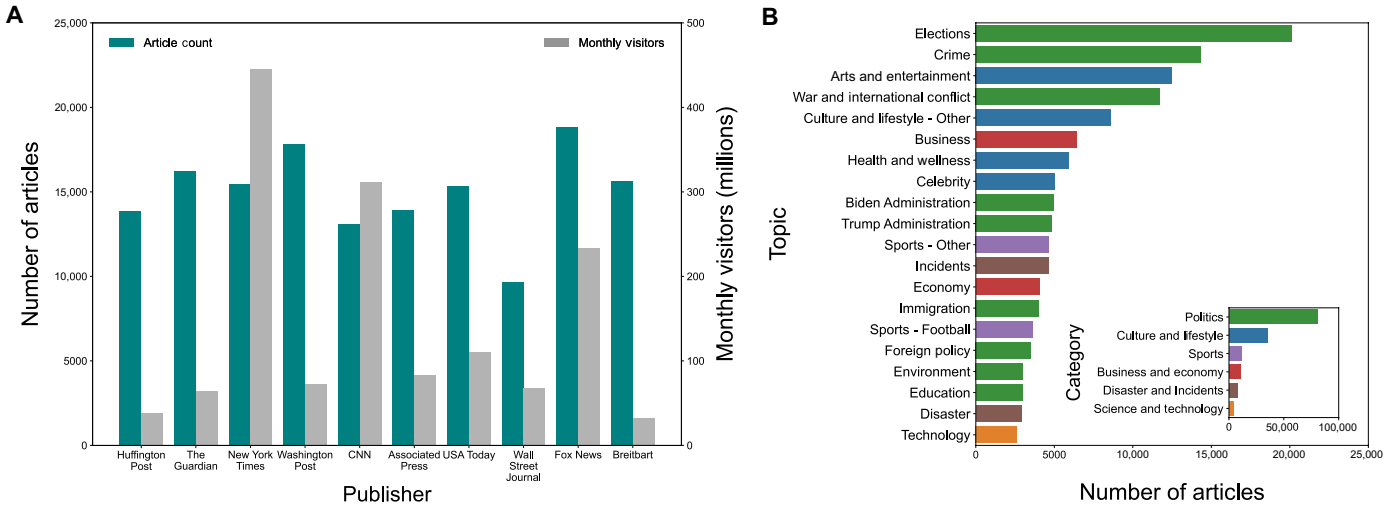

Fig. 3. Overall statistics of our dataset. (**A**) Total number of unique articles in our dataset from each publisher in 2024, compared to their monthly traffic for May 2025 (*47*). (**B**) The most popular news topics across this dataset, colored by category (shown in the inset). As expected in an election year, most of the news was dominated by politics, which comprises more than half of all articles in our dataset.

Of these, the first 10 were included in the initial launch in June 2024 and the remaining 11 were added in May 2025. As indicated earlier, we expect the collection of publishers will continue to increase over time, including other publication modes such as television, podcasts, and social media.

The justification for starting with a relatively small number of prominent publishers and growing the collection over time is twofold. First, our objective of characterizing bias requires us to capture all relevant text while minimizing extraneous content such as ad copy or boilerplate text that could generate spurious results. This goal is complicated by the highly variable formatting of news websites, particularly dynamic mixed-mode features such as "live" articles that increasingly serve as top stories on several publishers' websites. Focusing on just a few publishers allows us to customize our approach to individual websites and hence obtain a level of precision that would be infeasible at the scale of hundreds or thousands of publishers. Second, media consumption, especially by elites, is highly skewed such that the handful of prominent publishers that we do cover arguably play a larger agenda-setting role than even thousands of publishers with tiny audiences.

Every day, we take a snapshot of each publisher's homepage five times, at 6:00 a.m., 10:00 a.m., 2:00 p.m., 6:00 p.m., and 10:00 p.m. Eastern Time (ET). Then, we use headline placement on these homepage snapshots to identify the top 30 most prominent articles displayed to readers. We determine this position on the page as a combination of distance from the top, font size of the title text, and the presence and size of images accompanying the article. Currently, we disregard content that is not a news article, such as videos, podcasts, and photo galleries. We then scrape the text, title, and date of publication for every article in these top 30 (although we collect the data for the top 30 articles, for cost purposes the labeled data that we release on the dashboard and review here are currently restricted to the top 20 articles). We then preprocess the text to remove advertisements, boilerplate content, and superfluous text unrelated to the article. In addition, we deduplicate articles based on normalized URLs to account for superficial variations that may cause the same article to appear under technically distinct URLs, such as trailing slashes, query parameters, fragments, hostnames, and schemes. Articles whose normalized URLs are identical are treated as duplicates, and only one copy is retained in our dataset.

Our data collection started on 1 January 2024, for an initial 10 publishers (Associated Press, Breitbart News, CNN, Fox News, HuffPost, The Guardian, The New York Times, The Wall Street Journal, The Washington Post, and USA Today) and an additional 11 publishers (Los Angeles Times, Chicago Tribune, Star Tribune, BBC, Financial Times, Reuters, Bloomberg, CNBC, New York Post, Newsmax, and The Daily Wire) were added in May 2025. In planned future work, we will backdate our coverage of the later additions to January 2024 by using screenshots drawn from the Internet Archive (https://archive.org) to replace the "live" screenshots in the first step of the data acquisition process described above (we will also use this procedure to backdate all 21 publishers before January 2024). We note, however, that the illustrative results presented here are based only on the initial 10 publishers for the period 1 January to 31 December 2024, during which time we collected over 140,000 unique news articles. Figure 3A shows the total number of articles per publisher, along with their most recent monthly traffic statistics (*47*). Because all publications are being scraped at the same times, for the same number of articles, the differences in the number of unique articles reflect the speed with which they rotate their content.

Data labeling

Measuring media bias is an inherently challenging task even for humans, in part due to the lack of a universally accepted set of standard metrics (*48*). Unlike outright falsehoods, which can be fact-checked against verifiable sources (*49*–*51*), bias is inherently subjective and difficult to quantify (*52*). Readers often perceive articles as "biased" on the grounds that they disagree with them, but in doing so, they are (usually implicitly) treating their own subjective opinions as the "truth" (*53*). In many cases of interest, such as contentious social and political issues on which subjective views vary widely, there is no objective or even intersubjective ground truth for measuring bias in the news. As noted earlier, for example, even ostensibly neutral articles can be perceived as biased on the grounds that they ought not

to have been neutral. Unfortunately, however, in such cases the appropriate lack of neutrality is highly subjective. A second challenge is that the distinction between selection and framing bias is often blurry in practice. Consider, for example, a journalist who, intentionally or unintentionally, writes a negative story about the state of the economy by emphasizing facts about high inflation while omitting facts about high employment and wage increases. From one perspective, this example looks like framing bias: Having selected the topic of the economy, the journalist is framing the story around inflation rather than employment. From another perspective, however, it resembles selection bias, just operating at the level of facts rather than topics.

In response to these challenges, our approach is guided by two general principles. First, although our goal is to quantify bias, defined as systematic patterns of selection and framing effects that are observed across collections of articles or topics, we note that measuring bias directly requires some "unbiased" reference point to which these patterns can be compared. Given that such a reference point is generally unavailable, we instead focus on properties of an individual article's text, such as whether its tone is positive or negative or whether it aligns with one political party's position or another, that are easier to evaluate directly. Although, as we explain in more detail below, even these properties can be subjective, our view is that two readers with different opinions about a topic may agree on the tone and lean of an article even as they disagree on what an unbiased article would have said, and hence on the article's bias. By comparing patterns in these observable metrics between publishers at one period in time, or by tracking shifts in these patterns within a single publisher over time, we can detect and surface the signatures of potential bias; however, we refrain from assigning explicit "bias scores" to individual stories or publishers.

Second, we also avoid making a categorical distinction between selection and framing bias, instead labeling our data at both the sentence and article level and retaining the flexibility to aggregate our metrics at a variety of levels such as events, topics, publishers, and even groups of publishers. In this way, we can capture the effects of selection and framing without having to settle on a single operationalization of these constructs ex-ante. Although our approach is in some respects less satisfying than definitive-seeming bias scores and cannot detect certain instances of bias (e.g., false equivalency at the article level), we believe that it is more robust and scalable than methods that assume a notion of objective truth. Moreover, in instances where objective data are available (e.g., published economic indicators), our data can potentially be combined with such data to yield more definitive claims of bias for those specific cases. Figure 4 shows the various data points we extract from each news article, a sample of which can be found in text S1.

Topic labeling

Each day, we use the LLM to classify the content of every news story at three levels: category (politics, culture and lifestyle, business and economy, disaster and incidents, sports, and science and technology); topic (e.g., within politics: crime, elections, etc.); and subtopic (within crime: violent crime, political corruption, etc.). Given the large number of articles per day and the many possible ways to classify their content, we created a top-down category-topic-subtopic hierarchy via an iterative human-in-the-loop process.

First, a selection of coauthors hand-coded the topic and subtopic of all of the front-page articles in The New York Times and The Washington Post from 1 September 2022, to the date of the midterm elections on 8 November 2022 (*11*). Second, we asked the LLM to freely generate these labels for 1 month of news. Then, with a mix of humans and LLMs, we checked which topics and subtopics clearly fell under our initial set from the 2022 study, and what new topics and subtopics needed to be generated. Third, we followed an iterative process of asking the LLM to pick its labels from this list or otherwise suggest others if the list did not work.

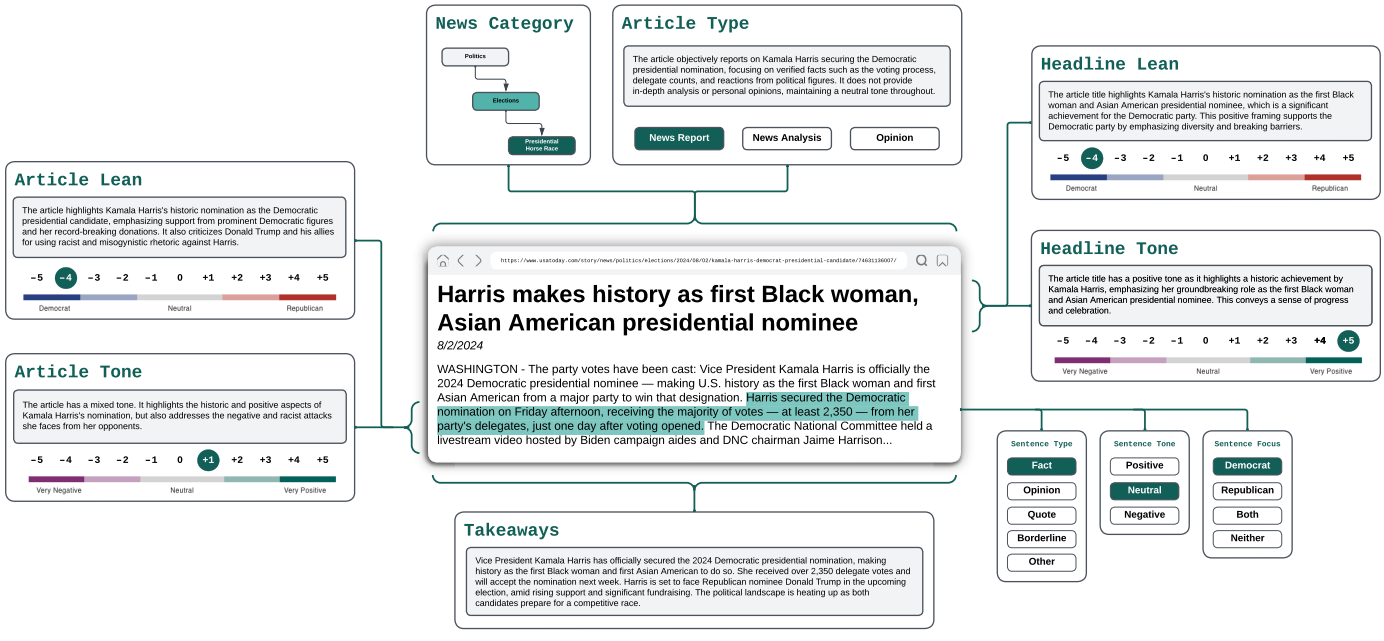

Fig. 4. Data points extracted from each article. All labels and qualitative descriptions are obtained using LLMs, and can be aggregated across articles to investigate a broad range of research questions about media coverage.

After several iterations, we landed on an initial list of topics and subtopics. Currently, the list contains 33 topics and 162 subtopics across the six categories, where we add new topics and subtopics when human editors see evidence of articles not cleanly falling within our current list. Each article is assigned to exactly one topic and subtopic, and each predefined topic is mapped to one of these six categories, defaulting to politics when an article may fall in several categories. Clearly, many articles could plausibly be classified into more than one topic, and many other topics and subtopics could have been included beyond the ones we chose. However, our preference for single assignment to a relatively short list of topics has the dual benefit of improving classification accuracy and simplifying aggregation. Figure S1 shows the full list of topics and subtopics developed in this process, and figs. S2 and S3 show the topic and subtopic labeling prompts that ask the LLM to choose from within this hierarchy, respectively. The most popular news topics in 2024 by number of articles are shown in Fig. 3B. As expected, most of the coverage was dominated by politics, in particular the U.S. Presidential Election.

Last, for each article we also generate a short summary of its main takeaways using GPT-4o-mini (see fig. S4 for the prompt). This summary distills the article into a few sentences that cover the key points a reader is likely to remember, and can be used as a compressed representation of the article content for downstream analysis. These summaries also facilitate manual exploration of the dataset by allowing users to quickly skim the takeaways instead of having to read the full article to understand its gist.

Article lean, tone, and type

Aside from topic categorization, the three article-level properties that we measure are political lean, tone, and type. Political lean, which we also refer to as lean, is defined as the extent to which the article either explicitly or implicitly aligns with the viewpoints, policies, or concerns of Republicans versus Democrats. Lean can manifest itself as overt partisan bias, such as when an article directly criticizes a political party or its representatives, but can also be more subtle, such as when certain topics or perspectives are inherently more aligned with one party's preferences or policies. For example, GPT-4o tends to code articles supporting social justice movements as Democratic-leaning, even when the article neither endorses nor critiques either party, on the grounds that such issues are more explicitly associated with Democrats than Republicans.

Lean, in other words, is itself a complex and multidimensional concept that is also somewhat subjective. To improve consistency and transparency, we first ask the LLM to generate a natural language analysis of the article's political lean and then provide an appropriate numerical score on an 11-point Likert scale in the range of [−5, +5] (see fig. S5 for the prompt). Following the chain-of-thought prompting approach (*54*), this encourages the model to articulate an analysis of the article before scoring, which both improves the consistency of its outputs and also provides an interpretable summary of the article's lean. Moreover, as we will describe below (see "Article lean, tone, and type" under "Data Validation"), human raters agree substantially with GPT-4o's lean and tone ratings.

We caution that we cannot rule out that both GPT-4o and human ratings are themselves biased in the sense that a right-leaning rater might code lean systematically differently from a left-leaning rater. Nonetheless, we believe that lean ratings are more reliable than asking directly about bias in the sense that it is easier (both for LLMs and humans) to determine that a particular article is pro-Republican or pro-Democratic than what its "true" orientation ought to be, and hence how it differs from the truth.

In addition to lean, we also measure tone, defined as the extent to which the article overall has a positive versus negative tone. As with lean, the tone of an article, or of a body of news coverage, also represents an editorial choice that can shape readers' opinions. For example, talking about the economy in a pessimistic tone can be interpreted as a publisher taking a negative stance with respect to the incumbent administration, and by extension, the governing political party. Although tone tends to be simpler and more easily interpretable than lean, it too is open to a certain amount of subjectivity. As with lean, therefore, we also use chain-of-thought reasoning, asking the LLM both for a human-interpretable analysis of the article and a numerical rating in the range [−5, +5] (see fig. S6 for the prompt). Figure 4 shows an example of the output from both prompts illustrating the surprising sophistication and depth of the explanations. In the "Article lean, tone, and type" section, we provide results on the validation of these labels. We also tested for robustness of these ratings by rerunning the same prompt five times on a set of 50 articles spanning diverse topics and the full range of lean and tone labels. The average Pearson's r was 0.99 for lean and 0.97 for tone (fig. S16). Of the 50 articles, the lean and tone ratings were identical across all runs for 35 and 37 articles, respectively. In cases with any variation, the confidence intervals were almost always within 1 scale point of the mean.

We also repeat the above labeling process separately for the articles' headlines to study how they may deviate from the actual article text in terms of lean and tone, which is itself a form of media bias. We use the same wording for both prompts with the addition of the article's topic to give the model more context to interpret the headline. The prompts for headline lean and tone labeling can be found in figs. S7 and S8, respectively.

Last, we also label the article's type as a news report, news analysis, or an opinion piece. We consider an article to be a news report if it provides fact-based, objective reporting on an event. Meanwhile, a news analysis article is a more in-depth examination accompanied by context. An opinion piece is more subjective than either of these, and can reflect personal views or beliefs of the writer. Classifying the type of an article is important for studying media bias because different types carry different expectations of objectivity. We generally expect news reports to be impartial, whereas opinion pieces are understood to represent a particular viewpoint (although we note that readers are not necessarily aware of this distinction). These labels can thus be used in conjunction with lean and tone to help understand where in a publisher's reporting bias shows up most strongly. As with lean and tone, we ask the LLM to provide a natural language analysis in addition to the categorical label (see fig. S9 for the prompt). Since the boundaries between these categories can often overlap—for example, a news analysis may blend factual reporting with interpretive framing in ways that bring it close to an opinion piece—the LLM's analysis serves as a useful check on borderline cases.

Sentence type, tone, and focus

In addition to article-level labels, we also label every sentence in an article with three features: type, tone, and focus. These sentence-level annotations complement the article-level labels by enabling more fine-grained analysis of how bias manifests within the structure of individual news stories.

The type of a sentence can be one of "fact," "opinion," "borderline," "quote," or "other." We do not ask the model to evaluate the truth or falsity of any claims, and hence by fact, we mean only that it has the form of a factual claim or statement, not that it is a true fact (*55*). Opinion refers to sentences that express a viewpoint or belief.

Borderline captures sentences that blend factual reporting with interpretive or evaluative language. Quote identifies direct quotations, which are handled separately in our quote extraction pipeline ("Quote extraction" section). Last, other is a residual category for sentences that do not neatly fit into any of the above types.

The tone of a sentence is defined in a similar way to tone at the article level but is assigned only one of three values: "positive," "negative," or "neutral." We use this coarser rating scale to reflect the more limited information available at the sentence level. However, even these coarse labels can be useful when aggregated, for example, to study the proportion of positive or negative sentences in coverage of a particular topic or individual.

Last, focus identifies whether the sentence refers to a particular political party, with possible values of "Democrat," "Republican," "neither," or "both." Focus differs from lean in that it only asks what is being referred to, not whether the reference is "pro" or "con," and hence is simpler to answer reliably (we do not code individual sentences for lean, as these ratings do not appear to be reliable).

Although these labels are extracted at the sentence level, their coding still requires the surrounding context of the article to make accurate judgments. Therefore, we pass the entire article to the model in the form of an enumerated list of sentences, which we obtain by using the Stanza toolkit (*56*) for segmentation. The full sentence labeling prompt can be found in fig. S10.

Quote extraction

Quotations serve as critical elements of news articles, providing firsthand perspectives to readers while simultaneously enabling journalists to frame their stories by choosing whom to quote and how. This allocation of voice, determining whose perspectives are amplified through direct quotation, represents a potentially important yet understudied dimension of media influence. A publisher that consistently quotes individuals associated with one political party over another, for example, implicitly frames its coverage of an issue through their lens.

To study this dimension of bias, we extract structured information for all quotes in each news article using GPT-4o. We focus on direct quotations because they form a clearly defined and reliably identifiable element of news articles. Furthermore, we perform this extraction as a dedicated task independently of the sentence labeling described above because quotes can span multiple sentences and can be accompanied by important contextual information that may be found in a separate sentence or in a different part of the article altogether. In particular, we extract the name of the person being quoted as well as their occupation and affiliation if that information is available anywhere in the article (see fig. S11 for the prompt). We also ask the LLM to classify the person's domain (e.g., Politics, Sports, etc.) and the capacity in which they are being quoted (e.g., a spokesperson, an expert, etc.) based on the context of the article. Together, these fields allow us to analyze whose perspectives are amplified by the media.

Event clustering

Comparing how different publishers cover the same event is fundamental to studying selection and framing bias, but doing so requires us to first identify which articles are about the same event. While the news categorization described earlier allows comparisons over long-running topics, event-level analysis enables a more granular analysis of coverage in a shifting news cycle. To this end, we develop a clustering framework that groups articles into precise, interpretable event clusters every day.

The clustering pipeline, illustrated at the bottom of Fig. 2, operates in three stages Fig. 2 In the first stage, we use OpenAI's text-embedding-3-large model to obtain vector embeddings for all news articles published on a given day. We then compute cosine similarity between all pairs of articles and represent them as nodes in a network where the edge weight between any two articles is equal to their similarity. We drop edges below an empirically determined threshold of 0.85, chosen through manual inspection of cluster quality during development, and keep only connected components containing at least three articles across all publishers. These connected components form cohesive article clusters around major news events. Importantly, this approach does not constrain the number or size of clusters: On days where the news cycle is dominated by one major story, we might observe a skewed distribution with one large cluster, while on other days with dispersed coverage, a larger number of smaller clusters may emerge.

In the second stage, we refine these initial clusters with LLM-based verification of their quality. First, we use GPT-4o-mini to assign a thematic title to each cluster based on the headlines of the articles in it (see fig. S12 for the prompt). This step yields a list of the day's major event themes. We then perform two complementary refinement steps. To improve recall, we give GPT-4o the headlines of all articles from that day that were not assigned to any cluster and ask it to assign the appropriate ones to each of the event themes (see fig. S13 for the prompt). To improve precision, we give the model the list of headlines within each event cluster and ask it to flag any that are not related to the theme (see fig. S14 for the prompt). These steps ensure that no related articles are incorrectly left out of a cluster, and no unrelated articles are incorrectly included within a cluster. The cluster assignments after this refinement step yield our final event themes and their constituent articles for that day.

In the third stage, we extract the key facts associated with each event to study which facts certain publishers choose to include (or omit) and how they frame them. We begin by retaining only those sentences within an article that were labeled as facts in the sentence labeling pipeline ("Sentence type, tone, and focus" section). Then, we apply the same embedding-based clustering process described above to all factual statements across all articles in that event, this time at the sentence level using the smaller text-embedding-3-small model and a lower cosine similarity threshold of 0.8. This yields clusters of highly similar statements coming from different publishers, where larger clusters represent the most widely repeated and crucial pieces of information surrounding the event. We then summarize each fact cluster, which contains variations of the same core information, into a single, representative, synthetic statement using GPT-4o-mini (see fig. S15 for the prompt). These summary facts capture the most widely reported and important pieces of information about a particular event, and can be used to study bias in fact selection between publishers at the event level. Figure S26 shows a sample of the output from our event clustering pipeline.

Data validation

To evaluate the quality of labels generated by our system, we built a custom data annotation platform and conducted separate data labeling exercises for each facet of our data: article lean/tone/type, article topic/subtopic, sentence type/tone/focus, and event clusters. Each of the four data dimensions was independently labeled by three annotators. We used Pearson's r to compute correlation for the numerical article lean and tone ratings and Cohen's κ to compute agreement for the remaining categorical labels. We represent the interannotator agreement as the mean of all pairwise agreement scores between the three annotators, and model-annotator agreement as the mean of all

pairwise agreement scores between the model and each of the three annotators. To evaluate whether the level of model-human agreement is consistent with human-human agreement, we perform a bootstrap test with 10,000 resamples to show that their confidence intervals overlap and that the agreement between the model and humans is statistically indistinguishable from that between the human annotators themselves. The following sections describe the results for each dimension of the data separately, with full results presented in Fig. 5.

Article topic and subtopic

We sampled 512 news articles with equal representation from all topics and subtopics where possible and asked three annotators to independently assign appropriate labels to them at both levels. The full set of 512 articles was used for the topic-level analysis. For the subtopic-level analysis, however, we restricted the sample to 243 articles where all annotators and the model had selected the same topic, and compared the subtopic labels within that subset. We computed pairwise Cohen's κ and found strong interannotator agreement at both the topic [mean = 0.668, 95% confidence interval (CI; 0.634 to 0.702)] and the subtopic level [mean = 0.734, 95% CI (0.691 to 0.778)]. Comparing these labels with the model's outputs, we find that it achieves comparable κ at both the topic [mean = 0.649, 95% CI (0.614 to 0.683)] and subtopic level [mean = 0.751, 95% CI (0.708 to 0.793)]. These results show that the model closely tracks human-level performance as the model-annotator agreement was statistically indistinguishable from interannotator agreement for both topics and subtopics.

In most cases, we found that misclassifications were not outright wrong, but that the article existed at the intersection of multiple topics and could reasonably have been classified as either. Figures S22 and S23 show the confusion matrices for topic- and subtopic-level classification, respectively. For example, an article on the involvement of the U.S. in a foreign conflict might be classified as about "foreign policy" but also as about "war and international conflict." We continue to evolve our news topic and subtopic hierarchy to accommodate shifts in the news cycle as well as to update it retrospectively and make it more robust to future changes.

Article lean, tone, and type

To evaluate our lean, tone, and type labels for news articles, we divided our 11-point Likert scale for lean and tone into five buckets each ([−5, −4], [−3, −2], [−1, 0, +1], [+2, +3], and [+4, +5]) and sampled two articles from each of the 25 possible lean-tone bucket combinations while ensuring a balanced distribution over the underlying 11-point values, yielding a total of 50 articles. We recruited three PhD students in communication who had a deep understanding of the media and the current U.S. political landscape to read each of these articles and code their lean, tone, and type. The full interface for these tasks can be seen in figs. S18 and S19, and the codebook provided to the annotators can be found in text S2.

To evaluate interannotator agreement for the numerical labels, we computed pairwise Pearson's *r* and found high correlations between the annotators for both lean [mean = 0.756, 95% CI (0.602 to 0.875)] and tone [mean = 0.772, 95% CI (0.667 to 0.858)]. The model's average correlations with the human annotators closely tracked the interannotator values for both lean [mean = 0.800, 95% CI (0.682 to 0.890)] and tone [mean = 0.786, 95% CI (0.692 to 0.862)]. For news type, the pairwise interannotator κ indicated moderate agreement [mean = 0.518, 95% CI (0.368 to 0.665)], reflecting the inherently subjective nature of the task, while the model's labels achieved comparable agreement as well [mean = 0.577, 95% CI (0.435 to 0.711)].

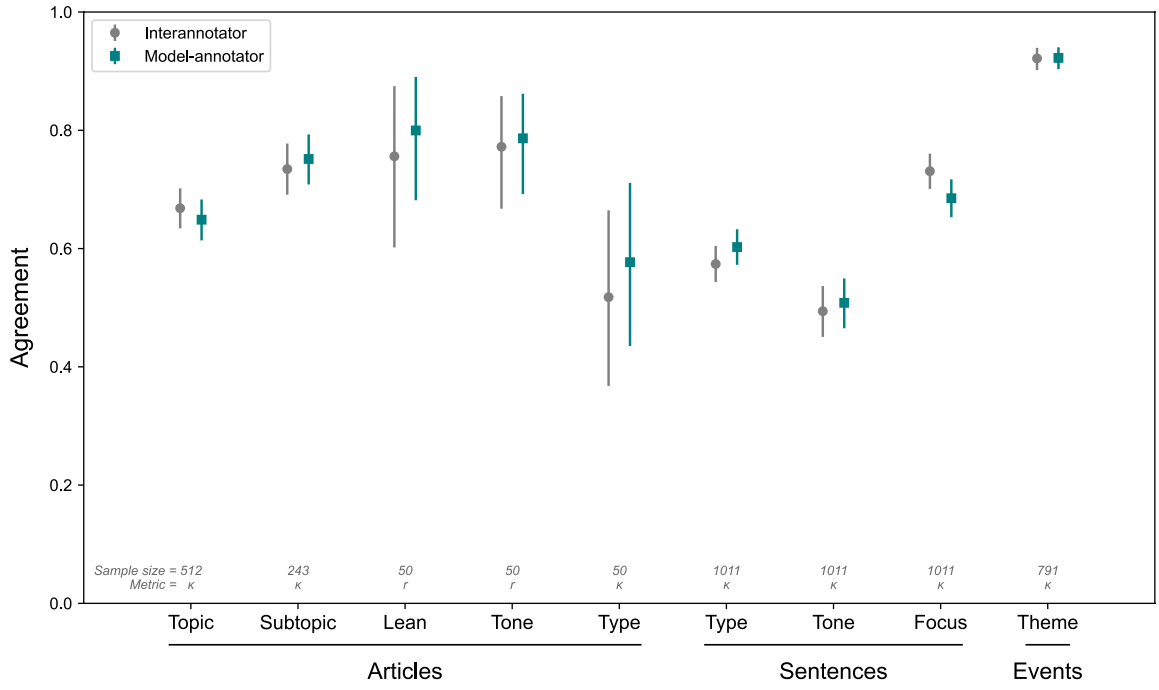

Fig. 5. Validation results across tasks. Average pairwise interannotator and model-annotator agreement across article, sentence, and event-level validation tasks. The strong correspondence between human and model agreement, along with overlapping confidence intervals, indicates that the model's labels are highly consistent with human labels.

In all three cases, the model's agreement with human labels was indistinguishable from the interannotator agreement. Figure S24 shows the scatter plots for lean and tone ratings and the confusion matrix for article type labels.

Sentence type, tone, and focus

To validate sentence labels, we used a subset of 25 articles from the article lean and tone annotation task and asked a separate set of annotators to label the type, tone, and focus of every sentence. These 25 articles contained a total of 1011 sentences, for an average of 40.44 sentences per article. The full task interface is shown in fig. S21 and the codebook provided to the annotators can be found in text S3.

Most sentences in the news articles were neutral facts that did not mention any particular political party, although this fraction varied for different article types and topics. The interannotator κ was moderate for sentence type [mean = 0.574, 95% CI (0.543 to 0.604)] and tone [mean = 0.494, 95% CI (0.451 to 0.537)] and high for focus [mean = 0.731, 95% CI (0.701 to 0.761)]. This pattern likely reflects that focus, which involves identifying concrete references to political actors or parties, is a more unambiguous feature than type and tone, which require more subjective judgments that depend on contextual interpretation. The model's agreement with the annotators showed a similar trend, with moderate values for type [mean = 0.603, 95% CI (0.572 to 0.633)] and tone [mean = 0.508, 95% CI (0.465 to 0.549)] and high κ for focus [mean = 0.685, 95% CI (0.653 to 0.717)]. Figure S25 shows the confusion matrices for all three sentence labels.

Event clusters

To evaluate the accuracy of our event clustering pipeline, we asked three annotators to read all headlines from a given day and assign them to the most relevant event theme out of the set generated by our system. Figure S20 shows the full interface for this task. We picked 1 October 2024, which was the date of the U.S. Vice-Presidential Debate and represented a mix of major and minor events being covered in the news. There were a total of 791 articles published on that day, with 361 of those distributed across 22 major events, ranging from new escalations in the Israel-Iran conflict (75 articles) and the VP debate (65 articles) to Julian Assange's plea and a Verizon outage (three articles each). The full list of events can be found in table S1.

The average interannotator agreement as measured by Cohen's κ was very strong [mean = 0.922, 95% CI (0.902 to 0.939)], and the model-annotator agreement was equally high [mean = 0.922, 95% CI (0.903 to 0.940)], where both were statistically indistinguishable from each other. In addition, we used precision and recall to evaluate the quality of the clusters, where higher precision indicates that the clusters contain fewer unrelated articles, and higher recall indicates that the clusters do not miss any relevant articles. Because our framework uses a high cosine similarity threshold, which ensures that only highly similar articles are clustered together, we achieve a precision of 0.941. In addition, the automated prompt-based cluster refinement essentially mimics a human review phase to filter out unrelated articles and achieves a high recall of 0.918. This results in an overall F1 score of 0.919, showing the effectiveness of our system for generating news clusters that are highly cohesive yet also have comprehensive coverage.

RESULTS

In this section, we apply our curated dataset from 2024 to a handful of questions about media practices and biases. Because of length constraints, we do not attempt to answer these questions comprehensively, but rather use these exercises to illustrate the richness of the dataset and demonstrate its potential to serve as a resource for future research. For example, our dataset allows researchers to answer many more focused questions by examining different subsets, enabling targeted investigations into specific aspects of media coverage, topical biases, or temporal trends.

The imbalance between policy and election horse race coverage

News organizations operate under constraints of limited time and space, meaning they need to make careful choices about which stories to cover. This gatekeeping power has been extensively studied in the social sciences, particularly under the framework of agenda setting, which posits that while news media may not directly dictate what people think, they substantially influence what people think about (*1*). In making editorial decisions, news organizations assess the newsworthiness of events based on various factors such as prominence, timeliness, and proximity (*57*). This assessment is particularly important given that news organizations are often private entities driven by profit or political motives and aim to conform to their audience's prior beliefs (*58*). Furthermore, these editorial choices have a major impact on readers, whose understanding of the world around them is shaped by the news they consume. Ideally, news organizations should help voters make informed choices by informing them about policy discussions. However, existing literature has highlighted that media outlets tend to focus predominantly on the "horse race" elements of elections, emphasizing candidates' polling numbers, strategic positioning, and projected wins, which can overshadow crucial policy issues (*59*, *60*).

Our data offer a more comprehensive quantitative analysis of the coverage of horse race dynamics compared with other critical policy issues in the lead-up to the 2024 U.S. presidential election than has previously been possible. Specifically, using our topic and subtopic classification system, we carried out a detailed comparison of the amount of coverage dedicated to the presidential horse race versus policy discussions. Figure 6 illustrates this disparity, showing that horse race coverage of the election overwhelmingly dominated all news outlets, even receiving more attention than all policy issues combined. Beyond this dominant topic, Fig. 6 also reveals systematic differences in other topics across publishers (*61*, *62*). For example, outlets like Breitbart and, to a lesser extent, Fox News allocated substantially more coverage to Immigration Policy and Reform in comparison to left-leaning media such as CNN or The Guardian.

Political lean varies across news topics

As discussed earlier, news producers can also influence public perception through their framing of issues (*7*), defined broadly as choices in language and context as well as the inclusion or omission of certain details and perspectives (*2*, *58*). Here, we study lean, the extent to which a news article supports Democratic or Republican viewpoints and policies,. Current methods typically reduce news publishers to a single value of metric by assigning them a political lean rating that ranges from "left" to "right" (*26*, *27*, *63*). While these static classifications are useful, they often fail to capture the nuances of bias within a single publisher, across various topics, or even from one article to another. Understanding how partisan alignment shifts depending on the subject matter can substantially enhance our comprehension of media dynamics and discourse.

As described above (see the "Article lean, tone and type" section), the Media Bias Detector provides lean labels at the article level,

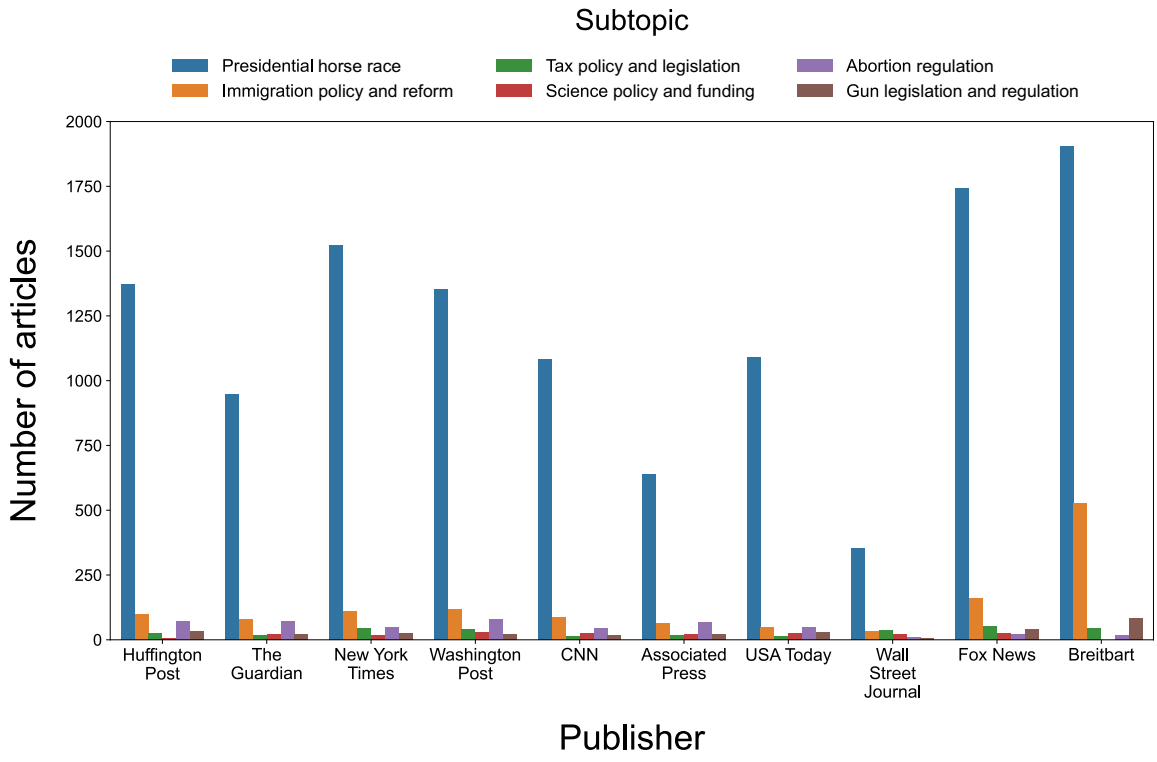

Fig. 6. The imbalance between policy and election horse-race coverage. The media primarily covered the election as a horse race instead of focusing on more substantive policy discussions to inform their readers of each candidate's and their party's agenda. Across all publishers in 2024, more than four times as many articles were focused on the horse race as on all policy topics combined, with a notable share of policy coverage driven by Breitbart's emphasis on immigration.

allowing for a more nuanced analysis of media outlets' political lean. Figure 7 presents the average political lean across publishers and topics, with negative values indicating a Democratic lean and positive values indicating a Republican lean. Although the overall pattern aligns with commonly held perceptions of these publishers' political lean, Fig. 7 reveals variations in lean not only across different outlets but also within publishers on different topics.

For example, although it is generally true that an outlet's stance on one issue tends to correlate with its stance on others, the extent of this alignment can vary considerably depending on the topic. For instance, The New York Times exhibits a consistently left-leaning position on topics such as the environment, whereas it maintains a relatively neutral stance on technology. In contrast, Fox News presents a more right-leaning stance on both topics. Figure 7 also shows that The Wall Street Journal appears to be the most neutral outlet within this set of publishers, maintaining a relatively balanced stance on most topics. Although The Wall Street Journal is often perceived as having a conservative lean, its apparent neutrality here likely stems from its relatively heavy focus on business and the economy, which, along with sports, travel, technology, and health and wellness, tend to be less partisan than political topics. We also reemphasize that neutrality does not imply a lack of bias. Whereas the former can be evaluated directly from the text of an article, the latter requires knowledge of some (typically unobservable) ground truth, which could potentially favor one "side" over the other. In such a case, a neutral stance would in fact reflect bias, as is often asserted in accusations of "false balance" or "bothsidesism." We also emphasize that even claims about neutrality are inherently relative and can only be made in the context of other publishers or, alternatively, the same publisher over time. The characterization of The Wall Street Journal as "the most neutral" publisher is therefore specific to this particular sample of publishers and likely will not generalize to other samples.

With these caveats in mind, we observe that the overall ordering of publishers based on our framework nonetheless aligns closely with past work that has focused on assigning ideological slant to news publishers based on their consumption patterns or crowd-sourced labels, as shown in fig. S27. Here, we compare the average political lean score for each publisher across all articles with estimates from (*64*) and (*19*). Our estimates align closely with both studies, showing strong correlations of $R^2 = 0.946$ with (*64*) and $R^2 = 0.976$ with (*19*), demonstrating the consistency of our approach with established findings. We also show that this lean is not static but fluctuates over time (fig. S28), highlighting changes in news outlets' reporting in response to major events.

Most news is negative, especially politics

As discussed earlier, the tone in which stories are presented can be another indicator of framing bias. Prior work has found that tone plays a critical role in agenda-setting (*65*) and influences the salience of negative and positive elements within the public debate (*66*). In politics, for example, voters have been found to be influenced by the tone of news coverage surrounding an election candidate (*67*), while in economic reporting, news sentiment predicts macroeconomic variables like consumption and output (*68*).

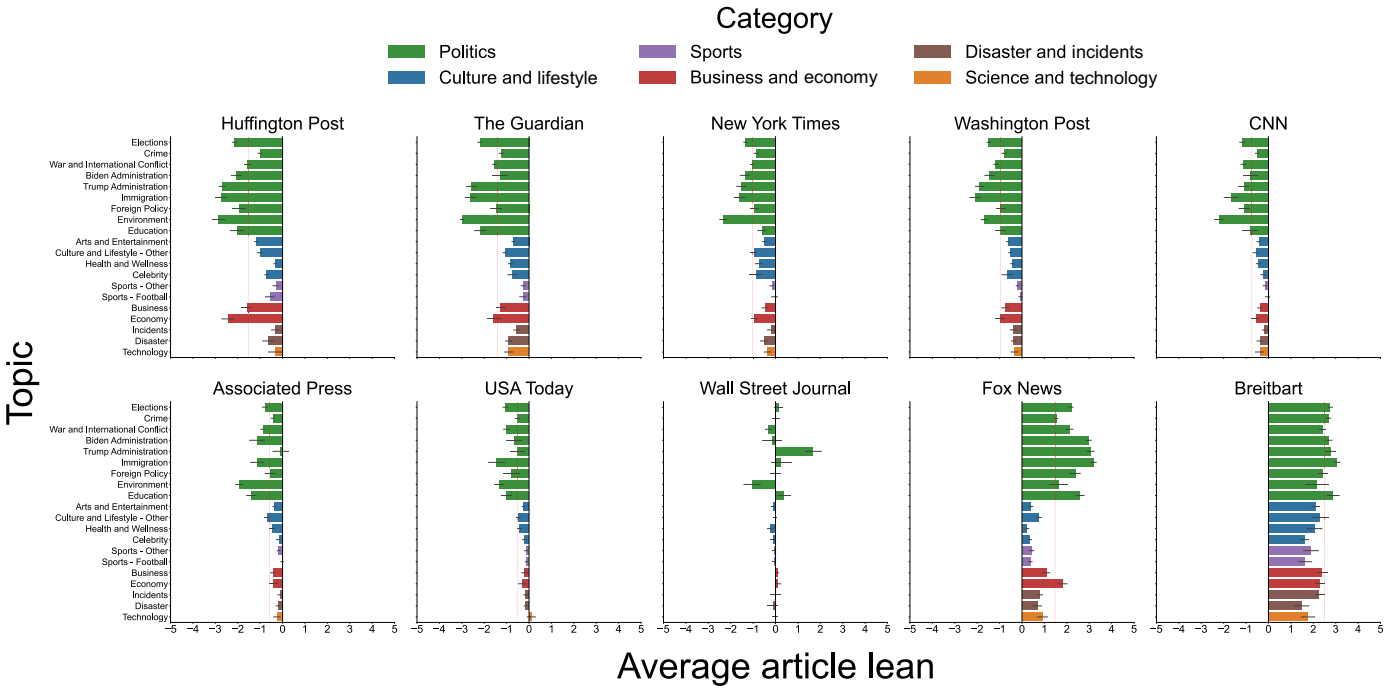

Fig. 7. Political lean varies across news topics. This figure illustrates the average (indicated by the vertical red line) political lean of selected news outlets, categorized by topic. Negative values indicate a Democratic lean, whereas positive values suggest a Republican lean. The red line shows the average lean for a publisher across all topics; the variation highlights the nuanced and topic-dependent nature of media bias. Variations in the degree of political alignment provide insights into each outlet's editorial stance and the complexity of bias representation in media coverage.

Here, we study negativity bias, an important case of framing bias in the media, which has been extensively documented in both the production and consumption of news (*69*, *70*). This focus on the negative could arise from both the selection and framing of news. On the selection side, it often involves prioritizing stories inherently perceived as negative, such as those involving conflict, crime, and disasters (*71*). On the framing side, any given story can be presented in a more or less negative light through the deliberate use of specific words and phrases (*72*). The tone classifications in our dataset capture both of these aspects, providing a comprehensive view of how negativity is embedded in news content across a wider range of topics and publishers than has previously been possible.

Figure 8 depicts the average tone (left *y* axis) of articles across topics as well as the number of articles on that topic (right *y* axis). We find that the majority of the news content analyzed contained negative elements, aligning with trends that are well documented in the aforementioned literature. However, we also extend this finding by showing how tone varies within political news, which is consistently negative yet also receives the highest volume of coverage. This combination of high negativity and high visibility in political news has important implications for democratic discourse, as it may contribute to political cynicism and disengagement (*72*). Moreover, our cross-topic analysis demonstrates how this negativity bias extends beyond politics (e.g., religion and economy), suggesting a systemic feature of contemporary news production that affects how the public perceives various aspects of social life.

Headlines are not always faithful to their articles

A headline is arguably the most important part of any news story and sets the stage for the rest of the article. For many casual readers, the headline might be their only takeaway about an event, an effect that is exacerbated by the news feeds common on social media platforms that focus almost exclusively on headlines (*73*). Consequently, any misalignment in lean or tone between news headlines and the accompanying articles (*74*) raises concerns for at least two reasons. First, many individuals decide to share information based solely on the headline, without reading the full news article. A recent study on social media sharing demonstrated that more than 75% of the news articles shared on Facebook were not clicked on before being forwarded (*75*). Second, even when people do read the entire article, the headline continues to exert a substantial influence on opinion formation. If the information presented in the headline is incongruent with the article's content, individuals often struggle to adjust their initial impressions (*76*) and their ability to recall facts presented in the article is negatively affected (*40*).

The separate labeling of headlines and article texts with identical prompts in our framework enables us to explore this phenomenon across both dimensions of tone and political lean. Figure 9 presents the results broken down by news category. In Fig. 9A, the average headline lean is represented on the *y* axis, and the average article lean on the *x* axis, with positive values indicating an increasingly pro-Republican lean (red) and negative values indicating an increasingly pro-Democratic lean (blue). The dark blue marker shows the mean across all articles, and the size of each scatter point reflects the number of articles within that category.

Our results reveal that news headlines systematically lean more pro-Republican than the articles themselves, as evidenced by the position of the dark blue marker, the overall average, which is above the diagonal in Fig. 9A. This pattern is most pronounced in the "business and economy" category, where articles are fairly neutral overall

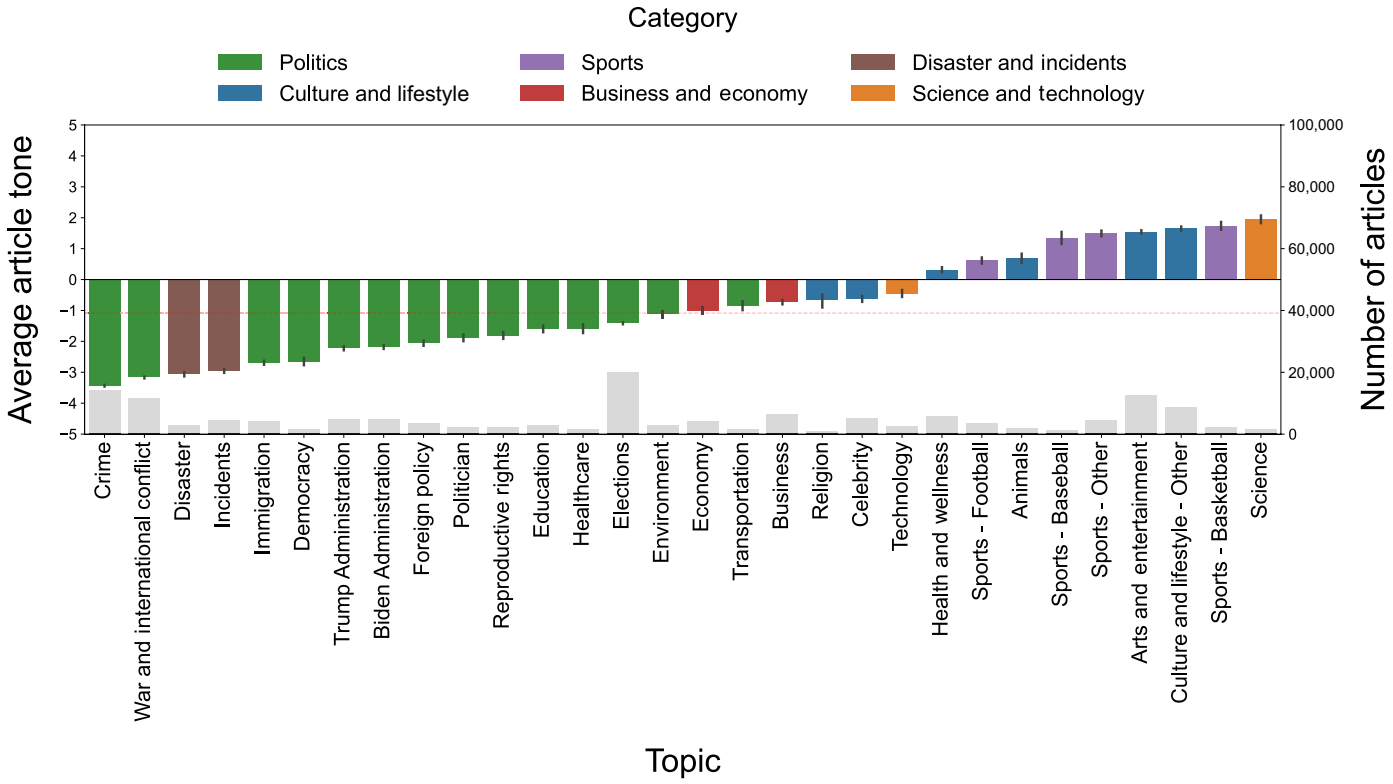

Fig. 8. Most news is negative, especially politics. The colored bars represent the average tone of articles that fall in the top 20 topics by number of articles (shown in gray), sorted from most negative to most positive. Negative tone is prevalent, both by number of topics and by the total volume of articles. Except for news about weather and natural disasters, all of the remaining topics that are more negative than the average (red line) are related to politics, highlighting the pervasive nature of negative elements in media coverage.

but are often accompanied by headlines that lean more pro-Republican. Similarly, for the "politics" category, the articles lean slightly pro-Democratic but the headlines are effectively neutral, a trend that can be even more pronounced for specific events (fig. S29). Figure 9B presents the same comparison for tone, with average headline tone depicted on the *y* axis and average article tone on the *x* axis. Although the overall mean lies close to the diagonal, suggesting a close correspondence of tone between headlines and articles on average, this aggregate measure masks substantial heterogeneity across topics. That is, topics with inherently positive content (sports, culture and lifestyle, and health) tend to have headlines that are less positive than their corresponding articles. Conversely, for typically negative topics like politics and disasters, the headline tone either aligns with or is slightly more positive than the article text. Comparing headline and article labels on both of these dimensions also shows that the observed pro-Republican lean is not due to differences in length or style of headlines compared to the full articles (since the same shift is not seen in tone), but may instead reflect editorial choices about how to frame the news.

Media outlets' partisan focus is inversely correlated with their lean

Partisan news outlets use various tactics to advance their ideological agendas, including their choices about which political actors to cover. One intriguing example that has been studied previously is the tendency of partisan media to focus more on critical coverage of opposing political groups than on positive coverage of their ideological allies (*39*). Our data can address this question by applying a combination of focus and lean measures at a larger scale than has previously been possible with human-annotated data. Figure 10 compares the political lean of articles, which measures the extent to which each article supports the viewpoints and policies of one party over the other, with the focus of the corresponding article for the 10 publishers in our dataset. As noted earlier, focus is measured on a sentence-by-sentence basis, indicating whether each sentence references the Democratic Party, the Republican Party, both, or neither. From these labels, we compute an article-level focus score by averaging the individual sentence scores, with Republicans coded as +1, Democrats as −1, and neither or both as 0. Figure 10 reveals a negative correlation between focus and lean, indicating that more left-leaning publishers tend to talk more about Republicans than right-leaning ones, and that the average focus across all outlets is skewed toward Republicans. We note that our focus measure conflates positive and negative mentions; thus, although our findings are directionally consistent with the hypothesis that partisan outlets focus more on opposing parties, they do not rule out the possibility that outlets also engage in "cheerleading" for their own side.

DISCUSSION

In this paper, we have introduced the Media Bias Detector, which leverages LLMs to analyze news media at the scale of hundreds of

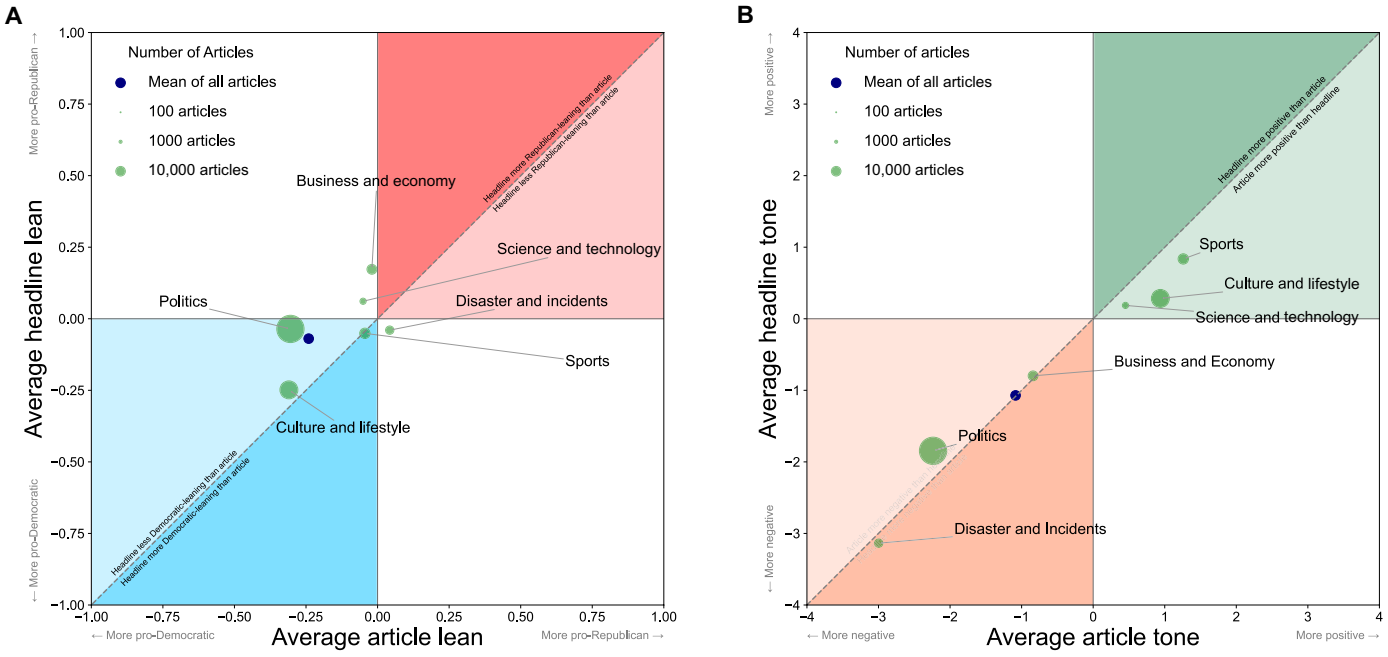

Fig. 9. Headlines are not always faithful to their articles. Each point represents average headline (*y* axis) versus article (*x* axis) lean or tone for a topic category. Positive lean values indicate pro-Republican stances; negative values indicate pro-Democratic stances. All content is labeled on a [−5, +5] scale (axes adjusted separately for clarity). (**A**) Headlines lean slightly more pro-Republican than their articles on average (blue point above diagonal), most notably for business and economy topics. (**B**) Headlines are more faithful to article tone, with the average (blue point) falling on the diagonal.

articles per day and in near-real-time to support and accelerate research on media bias. Our work builds on broader efforts to quantify the impact of news on democracy (*77*), in particular by constructing robust, scalable data infrastructure for studying how news moves from production to consumption and absorption. In addition, the Media Bias Detector exemplifies the related goal of building web-based interactive visualizations to communicate insights to nonacademic stakeholders (*38*). By making our data and dashboards widely accessible, we hope to enable data-driven transparency in journalism, allowing stakeholders, including policymakers, educators, and media professionals, to engage with the findings in meaningful ways.

The Media Bias Detector complements existing media datasets such as PeakMetrics, GDELT, and Media Cloud that already provide researchers with high-volume collections of online news articles, enabling large-scale studies of media coverage (*78*); indeed, much larger than we have presented here. Although these datasets collectively constitute a valuable resource, like all such efforts they suffer from limitations. As noted earlier, for example, scraping thousands or tens of thousands of websites daily, or even more frequently, over long periods inevitably introduces numerous problems with data missingness and quality. Articles that are easier to extract, such as those from publishers without paywalls, tend to be overrepresented, while those from publishers with sophisticated paywalls are more sparse. Furthermore, these very large-scale scraping efforts generally treat content across the entire website equally, regardless of whether it is featured prominently on the homepage or buried deep within the category pages. In contrast, the Media Bias Detector focuses on effectively capturing the top stories from each publisher's homepage at regular intervals with high reliability. As such, our approach prioritizes capturing stories that receive substantial exposure, creating opportunities to study how publishers shape public focus through their editorial choices.

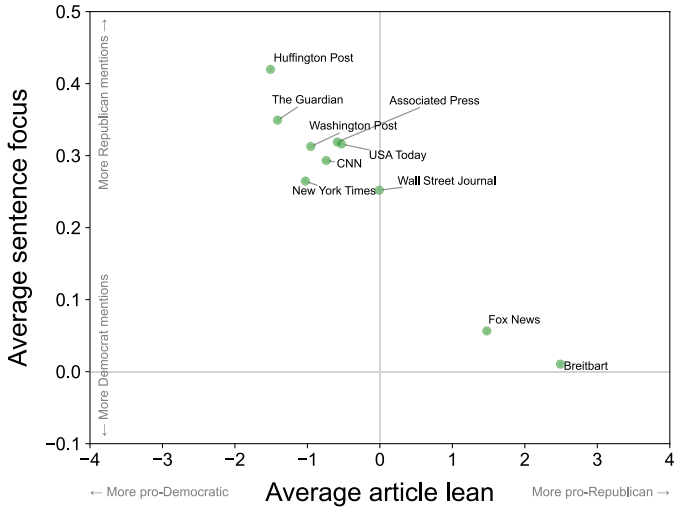

Fig. 10. Media outlets focus more on the opposing party than on their own side. Association between publishers' political lean and coverage of each political party. The *x* axis shows the average political lean of all articles by a publisher (positive values represent more pro-Republican lean and negative values represent more pro-Democratic lean) and the *y* axis shows the ratio of sentences that mention Republicans (positive) versus Democrats (negative). The more a publisher talks about a particular party, the more they lean toward the opposition.

Symmetrically, our approach also reveals which topics are absent from the headlines, allowing researchers to study patterns of omission, cases where major political actions, policy decisions, or social issues receive substantial coverage from some outlets but are largely overlooked by others. By leveraging LLMs and measuring multiple quantities at the article and sentence level, our approach generates

higher-quality and more granular labels than have been possible until very recently. Last, the continuous human-in-the-loop monitoring ensures that our system is able to adapt to the news cycle and avoid concept drift. In this way, we ensure that our dataset is always up to date with the events and topics being covered in the media and classifies them accurately. Although our current focus is on text-based news from mainstream media, our framework could also be adapted for social media by updating our prompts to more effectively label short-form text and by incorporating interaction data and network features such as likes and shares to measure the popularity and influence of various pieces of content.

Naturally, our approach also has important limitations. First, our data collection process does not provide exhaustive coverage of all online news. We limited ourselves to 21 major publishers (10 in the first stage of the project), but there are many more local and online news sources whose coverage may be of interest depending on the research question being studied. Further, by limiting ourselves to the top-ranking news articles from these publishers and only collecting them at regular intervals throughout the day, we may miss articles that either do not show up in the top 20 or appear only briefly and disappear before we are able to scrape them. The time frame of our collection is also limited to January 2024 onward, which limits the depth of temporal analysis that can be done with these data (although efforts are currently underway to backfill before 2024). In some cases, we also face issues with paywalls or broken links, which lead to gaps in our coverage.

Our data labeling and validation pipelines also have limitations. First, we use proprietary models that may become unavailable in the future, requiring us to switch to a different model, thus affecting the consistency of our data. Although we used state-of-the-art models at the time of writing, future models may be more capable on these tasks and may require relabeling past data. Beyond GPT-4o and GPT-4o-mini, we did not compare other models such as Claude, Gemini, or Grok, an exercise that could be helpful in understanding how models from different families and of different sizes analyze news articles' political lean and tone. Our approach also used zero-shot prompts throughout the pipeline and did not attempt to fine-tune any models. We also note that the level of agreement varies considerably between labeling tasks. As shown in Fig. 5, some tasks, such as article type, sentence tone, and sentence focus, show systematically less agreement than others. Fortunately, the model-annotator and interannotator agreement are statistically indistinguishable in all cases, suggesting that these differences are inherent to the tasks themselves and not a deficiency of the model. Moreover, when aggregated across hundreds or thousands of articles, these random misclassifications are less likely to have an impact and hence the overall patterns remain reliable even for "difficult" tasks.

A final limitation is one that we have already mentioned: that in the absence of clear, objective ground truth, bias is inherently difficult to measure. Some experts have concluded that the difficulty is so great as to render the detection of content that is misleading on account of bias rather than verifiable falsehoods unsuitable for scientific inquiry (*79*). Although we are sympathetic to this critique, we remain optimistic that something useful can be accomplished despite these challenges. Specifically, our approach of measuring "elemental" quantities such as topic, tone, lean, and focus is explicitly designed to avoid making judgments about the "true" or desired state of the world, and hence to result in more robust labels. For example, it is possible for raters with different biases to agree that an article is about the economy, has a positive tone, talks more about Democrats than Republicans, and hence is overall favorable to Democrats while maintaining very different views about its accuracy, informativeness, and even whether it should have been published at all.

As previously noted, a downside of this design principle is that we are unable to make direct assessments of article or even publisher bias, a limitation that media critics and political partisans alike may find frustrating. We maintain, however, that any attempt to impose a single notion of truth or objectivity against which bias can be measured will itself inevitably be biased, undermining the integrity of the project. Moreover, by measuring many of these elemental quantities both at the sentence and article level, and by allowing analysts the flexibility to aggregate them to yet higher levels (e.g., publisher, groups of publishers, the entire collection, etc.), we believe our approach does generate valuable insights about media bias, albeit indirectly. For example, our demonstration that news publishers devote markedly more coverage to horse race stories than to policy issues does not settle the question of how much coverage ought to have been devoted to each, but it does call into question public declarations by news organizations that their goal is to inform citizens about the most important issues in an upcoming election (*9*). Likewise, our demonstration that the political lean of headlines is systematically different from the text of the corresponding stories does not settle which one, if either, is correct, but the difference does show that writers of articles and headlines alike exploit their flexibility to cast the same story in different lights. Last, our demonstration that article lean is negatively correlated with partisan focus does not determine which of our publishers is more or less biased than the others, but it does identify a possible mechanism (attacking the opposition rather than supporting one's own side) by which all publishers express their bias.

In summary, by integrating systematic data collection, LLM-driven analysis, and human-in-the-loop validation, we have laid a scalable foundation for advancing research on media bias. Although no single study can fully encapsulate the complexities of media influence, and no single dataset can comprehensively define media bias, we believe this work represents a substantial step forward. By continually refining our methodology and expanding our dataset to encompass past and future time periods, we aim to enhance both academic understanding and public awareness of media bias dynamics through an empirical, data-driven approach.

Supplementary Materials

This PDF file includes:

Figs. S1 to S29
Supplementary Text S1 to S3
Table S1

REFERENCES

1. M. E. McCombs, D. L. Shaw, The agenda-setting function of mass media. *Public Opin. Q.* **36**, 176–187 (1972).
2. R. M. Entman, Framing: Toward clarification of a fractured paradigm. *J. Commun.* **43**, 51–58 (1993).
3. C. St. Aubin, J. Liedke, News platform fact sheet. *Pew Research Center* (2024), https://web.archive.org/web/20241231045522/https://www.pewresearch.org/journalism/fact-sheet/news-platform-fact-sheet/.
4. S. Wu, J. M. Hofman, W. A. Mason, D. J. Watts, Who says what to whom on twitter, in *Proceedings of the 20th international conference on World wide web* (2011), pp. 705–714.
5. N. J. Stroud, Media use and political predispositions: Revisiting the concept of selective exposure. *Polit. Behav.* **30**, 341–366 (2008).

6. M. A. Baum, T. Groeling, New media and the polarization of American political discourse. *Polit. Commun.* **25**, 345–365 (2008).
7. D. Chong, J. N. Druckman, Framing theory. *Annu. Rev. Polit. Sci.* **10**, 103–126 (2007).
8. D. Bourgeois, J. Rappaz, K. Aberer, Selection bias in news coverage: Learning it, fighting it. *Companion Proc. Web Conf.* **2018**, 535–543 (2018).
9. D. J. Watts, D. M. Rothschild, Don't blame the election on fake news. Blame it on the media. *Columbia Journalism Review* (2017); https://www.cjr.org/analysis/fake-news-media-election-trump.php.
10. J. F. Sheley, C. D. Ashkins, Crime, crime news, and crime views. *Public Opin. Q.* **45**, 492–506 (1981).
11. D. M. Rothschild, E. Pickens, G. Heltzer, J. Wang, D. J. Watts, Warped Front Pages. *Columbia Journalism Review* (2023); https://www.cjr.org/analysis/election-politics-front-pages.php.
12. C. Isch, Media bias in portrayals of mortality risks: Comparison of newspaper coverage to death rates. *Soc. Sci. Med.* **364**, 117542 (2025).
13. J. W. Dearing, E. Rogers, *Agenda-setting* (Sage publications, 1996).
14. F. Morstatter, L. Wu, U. Yavanoglu, S. R. Corman, H. Liu, Identifying framing bias in online news. *ACM Trans. Soc. Comput.* **1**, 1–18 (2018).
15. T. Groseclose, J. Milyo, A measure of media bias. *Q. J. Econ.* **120**, 1191–1237 (2005).
16. S. J. Farnsworth, B. Andrew, S. Soroka, A. Maioni, The media: All horse race, all the time. *Policy Options* (2007); https://irpp.org/wp-content/uploads/assets/po/realignment-in-quebec/farnsworth.pdf.
17. E. Haas, False equivalency: Think tank references on education in the news media. *Peabody J. Educ.* **82**, 63–102 (2007).
18. J. M. Shapiro, Special interests and the media: Theory and an application to climate change. *J. Public Econ.* **144**, 91–108 (2016).
19. C. Budak, S. Goel, J. M. Rao, Fair and balanced? Quantifying media bias through crowdsourced content analysis. *Public Opin. Q.* **80**, 250–271 (2016).
20. S. Lim, A. Jatowt, M. Färber, M. Yoshikawa, Annotating and analyzing biased sentences in news articles using crowdsourcing, in *Proceedings of the Twelfth Language Resources and Evaluation Conference* (2020), pp. 1478–1484.
21. A. Mitchell, J. Gottfried, G. Stocking, K. E. Matsa, E. Grieco, Covering President Trump in a polarized media environment. *Pew Research Center* (2017); https://www.pewresearch.org/journalism/2017/10/02/covering-president-trump-in-a-polarized-media-environment/.
22. M. Gentzkow, J. M. Shapiro, What drives media slant? Evidence from U.S. Daily newspapers. *Econometrica* **78**, 35–71 (2010).
23. J. Grimmer, B. M. Stewart, Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Polit. Anal.* **21**, 267–297 (2013).
24. P. DiMaggio, M. Nag, D. Blei, Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of U.S. government arts funding. *Poetics* **41**, 570–606 (2013).
25. M. Iyyer, P. Enns, J. Boyd-Graber, P. Resnik, Political ideology detection using recursive neural networks, in *Proceedings of the 52nd annual meeting of the Association for Computational Linguistics (volume 1: long papers)* (2014), pp. 1113–1122.
26. AllSides, https://www.allsides.com/media-bias/media-bias-rating-methods (2025).
27. NewsGuard, https://www.newsguardtech.com/ (2025).
28. Ad Fontes Media, https://adfontesmedia.com/interactive-media-bias-chart/.
29. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need. *Adv. Neural Inf. Proces. Syst.* **30**, 59995999–6009 (2017).
30. A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, Language models are unsupervised multitask learners. *OpenAI blog* (2019); https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.
31. T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners. *Adv. Neural Inf. Proces. Syst.* **33**, 1877–1901 (2020).
32. OpenAI, GPT-4 technical report. arXiv:2303.08774 [cs.CL] (2023).
33. C. W. Kuo, K. Chu, N. AlDahoul, H. Ibrahim, T. Rahwan, Y. Zaki, Neutralizing the narrative: AI-powered debiasing of online news articles. arXiv:2504.03520 [cs.CL] (2025).
34. S. Feng, C. Y. Park, Y. Liu, Y. Tsvetkov, From Pretraining Data to Language Models to Downstream Tasks: Tracking the Trails of Political Biases Leading to Unfair NLP Models, in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (2023), pp. 11737–11762.
35. T. Horych, C. Mandl, T. Ruas, A. Greiner-Petter, B. Gipp, A. Aizawa, T. Spinde, The promises and pitfalls of LLM annotations in dataset labeling: A case study on media bias detection. In *Findings of the Association for Computational Linguistics: NAACL 2025* (Association for Computational Linguistics, 2025); pp. 1370–1386.
36. F. Gilardi, M. Alizadeh, M. Kubli, ChatGPT outperforms crowd workers for text-annotation tasks. *Proc. Natl. Acad. Sci. U.S.A.* **120**, e2305016120 (2023).
37. C. Ziems, W. Held, O. Shaikh, J. Chen, Z. Zhang, D. Yang, Can large language models transform computational social science? *Comput. Linguist.* **50**, 237–291 (2024).
38. J. S. Wang, S. Haider, A. Tohidi, A. Gupta, Y. Zhang, C. Callison-Burch, D. Rothschild, D. J. Watts, Media Bias Detector: Designing and Implementing a Tool for Real-Time Selection and Framing Bias Analysis in News Coverage, in *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (2025), pp. 1–27.
39. G. Smith, K. Searles, Who let the (attack) dogs out? New evidence for partisan media effects. *Public Opin. Q.* **78**, 71–99 (2014).
40. U. K. H. Ecker, S. Lewandowsky, E. P. Chang, R. Pillai, The effects of subtle misinformation in news headlines. *J. Exp. Psychol.-Appl.* **20**, 323–335 (2014).
41. Y. Liu, K. Shi, K. He, L. Ye, A. Fabbri, P. Liu, D. Radev, A. Cohan, On Learning to Summarize with Large Language Models as References, in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)* (2024), pp. 8647–8664.
42. T. Zhang, F. Ladhak, E. Durmus, P. Liang, K. McKeown, T. B. Hashimoto, Benchmarking large language models for news summarization. *Trans. Assoc Comput. Linguist.* **12**, 39–57 (2024).
43. S. Rathje, D.-M. Mirea, I. Sucholutsky, R. Marjieh, C. E. Robertson, J. J. Van Bavel, GPT is an effective tool for multilingual psychological text analysis. *Proc. Natl. Acad. Sci. U.S.A.* **121**, e2308950121 (2024).
44. C. M. Pham, A. Hoyle, S. Sun, P. Resnik, M. Iyyer, Topicgpt: A prompt-based topic modeling framework, in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)* (2024), pp. 2956–2984.
45. Y. Chae, T. Davidson, Large language models for text classification: From zero-shot learning to instruction-tuning. *Sociol. Methods Res.* **55**, 501–567 (2023).
46. O. Kiss, M. Mobius, T. Rosenblat, Assembling News Like Legos (2023); https://www.markusmobius.org/publications/assembling-news-legos.
47. Press Gazette, Top 50 news websites in the US: Top sites see steep traffic falls in May. *Press Gazette* (2025), https://web.archive.org/web/20250701020023/https://pressgazette.co.uk/media-audience-and-business-data/media_metrics/most-popular-websites-news-us-monthly-3/.
48. T. Spinde, S. Hinterreiter, F. Haak, T. Ruas, H. Giese, N. Meuschke, B. Gipp, The media bias taxonomy: A systematic literature review on the forms and automated detection of media bias. arXiv:2312.16148 [cs.CL] (2023).
49. D. M. J. Lazer, M. A. Baum, Y. Benkler, A. J. Berinsky, K. M. Greenhill, F. Menczer, M. J. Metzger, B. Nyhan, G. Pennycook, D. Rothschild, M. Schudson, S. A. Sloman, C. R. Sunstein, E. A. Thorson, D. J. Watts, J. L. Zittrain, The science of fake news. *Science* **359**, 1094–1096 (2018).
50. S. Vosoughi, D. Roy, S. Aral, The spread of true and false news online. *Science* **359**, 1146–1151 (2018).
51. G. Pennycook, D. G. Rand, The psychology of fake news. *Trends Cogn. Sci.* **25**, 388–402 (2021).
52. S. Park, J. Y. Park, J.-h. Kang, M. Cha, The presence of unexpected biases in online fact-checking. *The Harvard Kennedy School Misinformation Review* (2021); https://misinforeview.hks.harvard.edu/article/the-presence-of-unexpected-biases-in-online-fact-checking/.
53. E. Pronin, T. Gilovich, L. Ross, Objectivity in the eye of the beholder: Divergent perceptions of bias in self versus others. *Psychol. Rev.* **111**, 781–799 (2004).
54. J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. V. Le, D. Zhou, Chain-of-thought prompting elicits reasoning in large language models. *Adv. Neural Inf. Proces. Syst.* **35**, 24824–24837 (2022).
55. A. Mitchell, J. Gottfried, M. Barthel, N. Sumida, Distinguishing between factual and opinion statements in the news. Pew Research Center's Journalism Project, https://www.pewresearch.org/journalism/2018/06/18/distinguishing-between-factual-and-opinion-statements-in-the-news/ (2018).
56. P. Qi, Y. Zhang, Y. Zhang, J. Bolton, C. D. Manning, Stanza: A Python Natural Language Processing Toolkit for Many Human Languages, in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (2020), pp. 101–108.
57. P. J. Shoemaker, S. D. Reese. *Mediating the message: Theories of influences on mass media content* (Longman, 1996).
58. M. Gentzkow, J. M. Shapiro, Media bias and reputation. *J. Polit. Econ.* **114**, 280–316 (2006).
59. S. Iyengar, H. Norpoth, K. S. Hahn, Consumer demand for election news: The horserace sells. *J. Polit.* **66**, 157–175 (2004).
60. R. Vliegenthart, H. G. Boomgaarden, P. Van Aelst, C. H. De Vreese, Covering the US presidential election in Western Europe: A cross-national comparison. *Acta Politica* **45**, 444–467 (2010).
61. S. Mullainathan, A. Shleifer, The market for news. *Am. Econ. Rev.* **95**, 1031–1053 (2005).
62. K. Welbers, W. Van Atteveldt, J. Kleinnijenhuis, N. Ruigrok, J. Schaper, News selection criteria in the digital age: Professional norms versus online audience metrics. *Journalism* **17**, 1037–1053 (2016).
63. Ground News, https://ground.news/rating-system (2025).
64. E. Bakshy, S. Messing, L. A. Adamic, Exposure to ideologically diverse news and opinion on Facebook. *Science* **348**, 1130–1132 (2015).

65. T. Sheafer, How to evaluate it: The role of story-evaluative tone in agenda setting and priming. *J. Commun.* **57**, 21–39 (2007).
66. C. De Vreese, H. Boomgaarden, Valenced news frames and public support for the EU. *Communications* **28**, 361–381 (2003).
67. J. N. Druckman, M. Parkin, The impact of media bias: How editorial slant affects voters. *J. Polit.* **67**, 1030–1049 (2005).
68. A. H. Shapiro, M. Sudhof, D. J. Wilson, Measuring news sentiment. *J. Econ.* **228**, 221–243 (2022).
69. G. Lengauer, F. Esser, R. Berganza, Negativity in political news: A review of concepts, operationalizations and key findings. *Journalism* **13**, 179–202 (2012).
70. C. E. Robertson, N. Pröllochs, K. Schwarzenegger, P. Pärnamets, J. J. Van Bavel, S. Feuerriegel, Negativity drives online news consumption. *Nat. Hum. Behav.* **7**, 812–822 (2023).
71. T. G. L. A. Van Der Meer, M. Hameleers, I Knew it, the world is falling apart! Combatting a confirmatory negativity bias in audiences' news selection through news media literacy interventions. *Digit. Journal.* **10**, 473–492 (2022).
72. M. Trussler, S. Soroka, Consumer demand for cynical and negative news frames. *Int. J. Press Polit.* **19**, 360–379 (2014).
73. K. Searles, J. T. Feezell, Scrollability: A new digital news affordance. *Polit. Commun.* **40**, 670–675 (2023).
74. F. Marquez, How accurate are the headlines? *J. Commun.* **30**, 30–36 (1980).
75. S. S. Sundar, E. C. Snyder, M. Liao, J. Yin, J. Wang, G. Chi, Sharing without clicking on news in social media. *Nat. Hum. Behav.* **9**, 156–168 (2025).
76. N. Carcioppolo, D. Lun, S. J. McFarlane, Exaggerated and questioning clickbait headlines and their influence on media learning. *J. Media Psychol* **34**, 30–41 (2022).
77. D. J. Watts, D. M. Rothschild, M. Mobius, Measuring the news and its impact on democracy. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e1912443118 (2021).
78. H. Roberts, R. Bhargava, L. Valiukas, D. Jen, M. M. Malik, C. Bishop, E. Ndulue, A. Dave, J. Clark, B. Etling, R. Faris, A. Shah, J. Rubinovitz, A. Hope, C. D'Ignazio, F. Bermejo, Y. Benkler, E. Zuckerman, Media Cloud: Massive Open Source Collection of Global News on the Open Web, in *Proceedings of the International AAAI Conference on Web and Social Media* (2021), vol. 15, pp. 1034–1045.
79. D. Williams, Misinformation researchers are wrong: There can't be a science of misleading content. *Conspicuous Cognition* (2024), https://www.conspicuouscognition.com/p/misinformation-researchers-are-wrong.

Acknowledgments: We thank Y. Zhang, G. Jennings, E. Pickens, and Y. Duan for developing the data collection and annotation infrastructure, and A. Patel and B. Li for helpful discussions on the methodology. We also thank A. Gupta for helping organize and coordinate the project, as well as C. Isch, B. Howland, N. Fasching, G. Jeon, B. Pierce, U. Dutta, and J. Nguyen for assistance with data annotation. D.J.W. and D.M.R. acknowledge the long collaboration with M. Mobius, during which they worked on pre-LLM methods to cluster news articles and display them visually on a web-based dashboard as part of what was called "Project Ratio," initially at Harmony Labs and subsequently at Microsoft Research (https://microsoft.com/en-us/research/project/project-ratio/). As described in detail in Materials and Methods, we used OpenAI's GPT-4o and GPT-4o-mini models to generate structured annotations of news articles at both the article and sentence level. At the article level, we classified topic, subtopic, political lean, tone, and type, while at the sentence level, we classified type, tone, and focus. In addition, we used these models to generate summaries of news articles and event clusters. Samples of these outputs were systematically validated by human annotators to ensure accuracy.
Funding: This research was developed with funding from Richard Jay Mack, the Knight Foundation (Grant ID G-2023-67859), and the Defense Advanced Research Projects Agency's (DARPA) SciFy program (agreement no. HR00112520300). The views expressed are those of the authors and do not reflect the official policy or position of the Department of Defense or the U.S. Government. **Author contributions:** Conceptualization: D.J.W., C.C.-B., D.M.R., and S.H. Methodology: D.J.W., C.C.-B., D.M.R., S.H., A.T., J.S.W., and T.D. Software: S.H. and J.S.W. Validation: S.H., A.T., J.S.W., and T.D. Formal analysis: S.H. and A.T. Investigation: S.H., A.T., J.S.W., and T.D. Resources: D.J.W. and C.C.-B. Data curation: S.H., A.T., and J.S.W. Writing–original draft: D.J.W., S.H., A.T., J.S.W., and T.D. Writing–review and editing: D.J.W., D.M.R., S.H., and A.T. Visualization: S.H. Supervision: D.J.W., C.C.-B., and D.M.R. Project administration: D.J.W., C.C.-B., and D.M.R. Funding acquisition: D.J.W. and C.C.-B. **Competing interests:** The authors declare they have no competing interests. **Data, code, and materials availability:** All data and code needed to evaluate and reproduce the results in the paper are present in the paper and/or the Supplementary Materials, and are available at the following links: Zenodo: https://doi.org/10.5281/zenodo.21028090. GitHub: https://github.com/Watts-Lab/media-bias-detector.

Submitted 28 July 2025
Accepted 15 June 2026
Published 9 September 2026
10.1126/sciadv.aea7456

ScienceAdvances

The Media Bias Detector: A framework for annotating and analyzing the news

Samar Haider, Amir Tohidi, Jenny S. Wang, Timothy Dörr, David M. Rothschild, Chris Callison-Burch, and Duncan J. Watts

Sci. Adv. **12** (37), eaea7456. DOI: 10.1126/sciadv.aea7456