

The 2024 US Presidential Election PoSSUM Poll

Roberto Cerina, *Institute for Logic, Language and Computation, University of Amsterdam, NL*

Raymond Duch, *Nuffield College, University of Oxford, UK*

ABSTRACT


The initial predictions presented in this article confirm that presidential candidate vote-share estimates based on AI polling are broadly exchangeable with those of other polling organizations. We present our first two biweekly vote-share estimates for the 2024 US presidential election and benchmark them against those being generated by other polling organizations. Our post-Democratic National Convention top-line estimates for Trump (47%) and Harris (46%) closely track measurements generated by other polls during the month of August. The subsequent early September (post-debate) PoSSUM vote-share estimates for Trump (47%) and Harris (48%) again closely track with other national polling being conducted in the United States. An ultimate test for the PoSSUM polling method will be the final preelection vote-share results that we publish before Election Day on November 5, 2024.


We survey citizens' voting preferences to understand or explain their voting decision but also to predict election outcomes. Because we observe election outcomes on a regular basis, we are able to monitor the trends in the performance of our modeling efforts. As Jennings and Wlezien (2018) pointed out, the overall prediction error in preelection national polls actually has declined somewhat, reflecting the increasing number of polls being produced and individuals polled. Conversely, particularly during the past decade, state-level polls and some national polling organizations have performed poorly, and the results of some presidential contests have been more difficult to predict (Clinton et al. 2021; Jackson and Lewis-Beck 2022; Kennedy et al. 2018). Maintaining a low level of prediction error in preelection polling has become increasingly challenging. This article describes how we address this challenge with a method that combines recent advances in Large Language Models (LLMs) with the proliferation of social media content. To illustrate, we estimate the

vote shares of 2024 US presidential candidates on a biweekly basis using our AI polling method: PoSSUM, a Protocol for Surveying Social Media Users with Multimodal LLMs.

Election polling has faced challenges on a number of fronts but three core elements of the polling enterprise have proved particularly challenging. Election polls are now almost entirely conducted either on the telephone or online. Response rates for traditional random digit dial polls are now much less than 10% (Keeter et al. 2017; Kennedy and Hartig 2019). Similar low response rates have been reported for recruitment into online surveys (Mercer and Lau 2023; Wu et al. 2023). Selection effects imply that these samples often are not representative of the broader population. The use of increasingly unrepresentative samples contributes to systematic bias in the predictions of public opinion polling (Kennedy et al. 2018; Sturgis et al. 2016).

The foundation of traditional polling is a survey instrument that poses questions to which interviewees respond. Critical assessments of the design of these questions, the timing of the interview, and how survey respondents answer these questions suggest that the survey and interview likely bias polling results. A second possible factor contributing to prediction performance of election polls is the sincerity of voting intentions expressed by survey respondents. For example, evidence suggests that social desirability affects survey reporting of voting intention (Claassen and Ryan 2024) and likely voting turnout.

Roberto Cerina  is assistant professor in humane and social AI at the Institute for Logic, Language and Computation at the University of Amsterdam. He can be reached at r.cerina@uva.nl.

Corresponding author: Raymond Duch  is director of the Centre for Experimental Social Sciences at Nuffield College at the University of Oxford. He can be reached at raymond.duch@nuffield.ox.ac.uk.

© The Author(s), 2024. Published by Cambridge University Press on behalf of American Political Science Association. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (<http://creativecommons.org/licenses/by/4.0>), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

A third critical and increasingly challenging element of the polling exercise is weighting of the sampled respondents (Gelman 2007; Rothschild et al. 2018). Most important, nonresponse is not random, which has undermined efforts to weight survey data. This has affected the accuracy of election surveys (Clinton et al. 2021; Kennedy et al. 2018) as well as surveys conducted in other areas (Bradley et al. 2021). As a result, scholars give increasing attention to the correlation between whether and how people respond to surveys and how this correlation interacts with population size (Bailey 2023, 2024).

This article introduces an alternative AI-driven approach to polling that significantly reduces the estimation biases associated with these three features of traditional polling. Our biweekly PoSSUM estimation of the 2024 US presidential vote share provides an opportunity to test this claim. The article first describes how AI polling is likely to reshape the future of election polling. The second section describes the methodology. Results of our first two biweekly estimates of 2024 presidential vote share, benchmarked against other polling organizations, are presented in the third section. The discussion concludes in the fourth section.

THE AI FUTURE OF POLLING?

In the not-to-distant future, the entire polling enterprise will be redefined by the value added that LLMs can bring to the design, implementation, and analysis of surveys. Our PoSSUM poll of the 2024 US presidential election illustrates one direction that this AI election polling can take. Our proposed AI polling method leverages the proliferation of social media content and recent developments in LLMs while retaining the core features of a classic public opinion poll.

Population

The “target” population of interest is likely voters in the 2024 US presidential election. Our data collection is guided by a stratification frame that represents the population of the United States. We populate the relevant cells of this stratification frame with population figures from the American Community Survey (US Census Bureau 2021). The vote probabilities in these cells are estimated using multilevel regression with post-stratification (MRP) along with the results from our AI survey—an estimation strategy that Cerina and Duch (2023) and others (Lauderdale et al. 2020) have championed as a method for improving the precision of vote-share estimates.

Sampling

The classic data-collection strategy for election polling is a version of a random probability sample from the population of individuals who are eligible to vote in the US presidential election. As previously mentioned, these samples are increasingly unrepresentative and problematic. In many cases, the sample is not from the US population *per se* but rather a segment of the population. This is the case, for example, with online surveys that sample individuals who have Internet access or who have been recruited into a sample pool.

All of these methods have in common the fact that the individuals in their sample respond to interviews either in person, on the telephone, or online. Our AI polling does not require our sample of people to respond to questions. The LLMs will collect digital traces from members of the population of interest. These

digital traces will come from diverse subscribers but they hardly represent the complete population. This sampling requires that social media platforms provide sufficient information to allow the LLM to match the account holder to a cell in our stratification frame. There also must be a sufficient regular volume of political content to allow the LLM to infer an opinion or preference—in our case, likely vote choice. The LLM will parse out the digital traces that are informative. The goal is to construct a representative sample of the population of interest. Few social media platforms meet these criteria; X (formerly Twitter), with all its imperfections, does satisfy these conditions and is the basis for our online social media panel. Pfeffer et al. (2023) provide an informative overview of the X “population”: their complete 24-hour “audit” of tweets generated 375 million tweets sent by 40,199,195 accounts. During this 24-hour period, the United States accounted for 20%, or about 70 million tweets, generated by 8 million accounts. The authors’ analysis of hashtags suggests that approximately 5% had a political theme—ignoring Iranian protest hashtags that accounted for 15% at the time. For our 2024 presidential vote-share estimates, we sample from these US X accounts. Previous efforts to use X for election forecasting have failed in part because of how the X samples are constructed and subsequently deployed in forecast modeling (Huberty 2015). We address these limitations by adopting an innovative approach to sampling social media that harnesses the power of recent advances in LLMs along with MRP statistical modeling.

The AI polling method we propose can accommodate and should include diverse social media platforms such as Facebook, Instagram, and TikTok. Each of these platforms caters to distinct demographic profiles, and tapping into this diversity would reduce bias in our digital sampling frame. Progress in incorporating this diversity into our digital sample is hindered by access restrictions to the Application Programming Interfaces (APIs) of these social media platforms.

Interview

Public opinion surveys consist of a questionnaire with closed and open-ended questions that are administered by an interviewer either in person or on the telephone; alternatively, they are administered online. As discussed previously, the “interview” must be constructed and administered, and it is the source of significant measurement error (Krosnick, Presser, and Art-Sociology Building 2009). This is problematic because the accuracy of election polling is reliant on interviewees expressing sincere preferences and opinions. We avoid this particular source of measurement error with our method because LLMs do not ask questions. They unobtrusively observe digital conversations and infer preferences and opinions from the conversations—they are, for example, instructed to infer vote choice from the digital traces that they “digest.”

Although AI polling is unlikely to experience these conventional sources of measurement error, other types of measurement may be prevalent. Of particular concern for our method, from a measurement perspective, is whether (1) individuals are misrepresenting their sincere political preferences; and (2) this misrepresentation goes undetected by the LLM. For example, social pressures might lead some individuals to express “conforming” opinions within their social media networks. Our ongoing research will explore the extent to which this is the case. Although

there clearly is a hesitancy for individuals to express their political preferences on social media, our intuition is that misrepresentation of preferences is probably relatively rare (McClain 2019).

Uncertainty

A broader challenge that encompasses measurement error is to associate a measure of uncertainty with the estimates generated by AI polling. We propose a number of strategies in this regard. First, the LLM associates a speculation score with the profile estimate it generates (e.g., the profile's gender and likely vote).

Weighting

Of course, our method makes no claim to be a random probability sample. Our point of departure is quota sampling. The LLMs are instructed to identify sufficient digital information for each cell of a stratification frame. The occurrences of the cells in the population effectively “weight” the digital opinions that we collect. We recognize the limitations—we are not observing the counterfactual identical individuals with each of our sociopolitical stratification frame profiles who are not X users. These “counterfactual” individuals may not be “missing at random,” thereby introducing bias into our estimates of vote share (Bailey 2023, 2024).

THE METHOD

As with conventional polling, our data collection focuses on sampling and conducting interviews (Cerina and Duch 2024). Our approach is tailored to the X API, which uses the digital trace of X users as the mold for LLM generation. However, this general approach can be extended to any social media that allows querying of a user panel via user- and content-level queries. PoSSUM is composed of two principal LLM routines that create the digital panel and then conduct the digital interview.

Gathering a Digital Panel

To create a digital panel of X users, we rely on the tweets and search API endpoint. Users who have participated in conversations related

to the query during the past seven days (as per the limits of X's Basic API tier) are gathered to build the digital subject pool. Listing 1 is an example query for the X API. This type of query is likely to yield users who explicitly express opinions about candidates—and therefore will yield highly informative digital traces—that the LLM can annotate with confidence. However, selection effects loom large with this type of query—that is, the type of user who frequently comments on politics on X is likely to be different from one who does not, *ceteris paribus*. To account for this selection, we complement this political query with a set of queries based on currently trending topics (see <https://trends24.in/united-states>). Trending topics may be related to politics—for example, during party conventions and televised debates—although they are more likely to be associated with events such as sports, concerts, marketing campaigns, famous people, and otherwise *viral* online content. Users engaging with this set of queries are far more likely to be *normies*, who pay relatively little attention to politics and therefore can balance the high-attention selection associated with the query in listing 1. Figure 1 is an illustration of the trending topics associated with users in our digital panel.

The digital panel then is further filtered, according to a number of sequential exclusion criteria. This is done for two reasons: (1) it contributes to data quality by ensuring that the digital traces belong to real existing users within the population of interest; and (2) it improves the efficiency of the sampling by identifying hard-to-find users who are more “valuable” for the pool. We exclude from the sample users whose self-reported location information is missing and those for whom we already have gathered a digital trace within the last τ days (i.e., to avoid overreliance on frequently active users). Users who do not represent a real offline person—including accounts for organizations, services, and bots—are discarded. Users who reside outside of the United States are discarded. We again rely on the LLM's judgment, using the profile as a whole to make a determination when the self-reported location is not exhaustive or otherwise uncertain. Given the user's characteristics, we then match the user to a cell in the population,

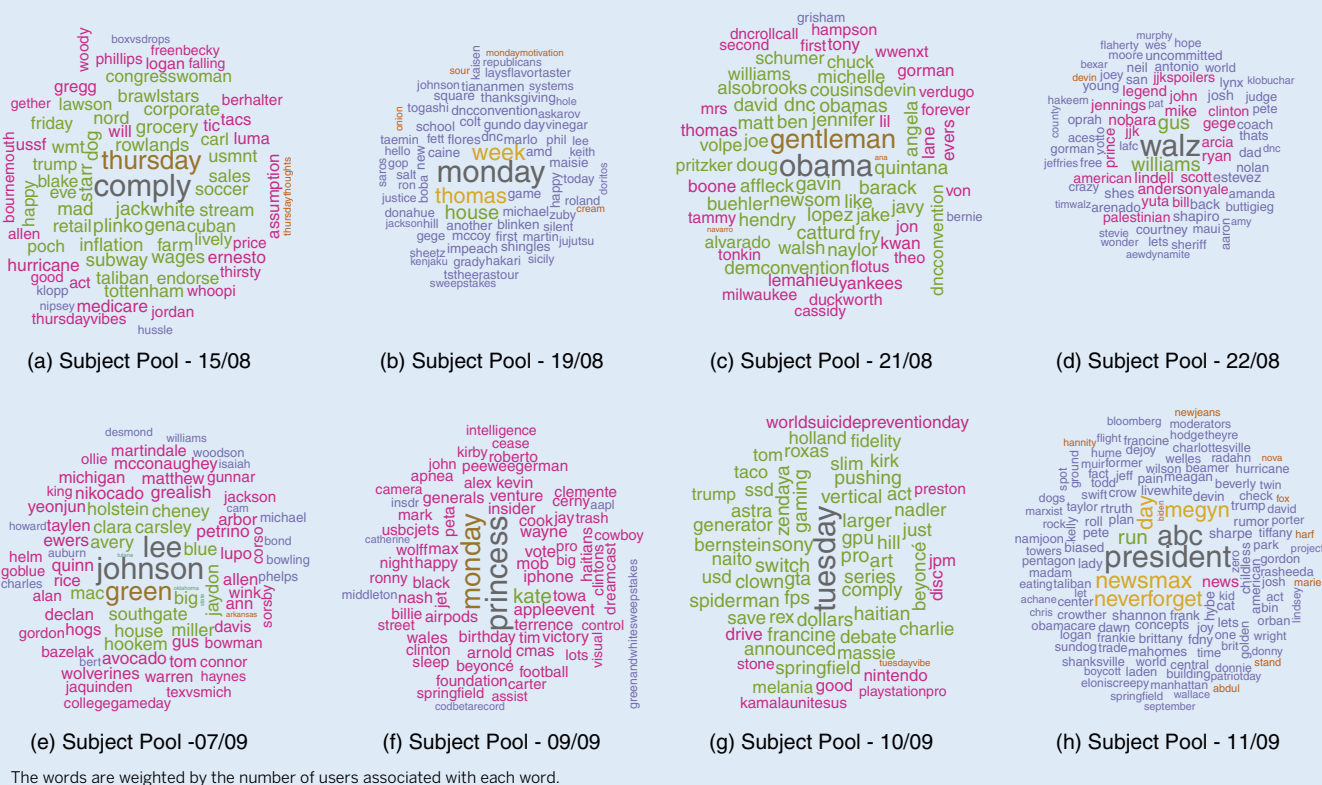
Listing 1

Search Terms for Tweets Related to Candidates Involved in the US 2024 Presidential Election

```
1 query <-
2 " (
3   Kamala OR VP OR KamalaHarris OR           # Democratic candidate terms
4   MAGA OR Trump ORrealDonaldTrump OR         # Republican candidate terms
5   Robert Kennedy OR RFK OR RobertKennedyJr OR RFKJr
6   OR KennedyShanahan24 OR Kennedy24 OR       # RFK terms
7   Cornel West OR Dr. West OR CornelWest OR   # Cornel West terms
8   Jill Stein OR DrJillStein OR               # Green candidate terms
9   ChaseForLiberty                             # Libertarian candidate terms
10  )"
11 -from:VP -from:KamalaHarris                  # Don't sample candidate profiles
12 -from:realDonaldTrump
13 -from:RobertKennedyJr
14 -from:CornelWest
15 -from:DrJillStein
16 -from:ChaseForLiberty
17 -is:retweet"
```

Figure 1

Word-Cloud Presenting Words from the “Trending” Queries to the X API, for PoSSUM Polls Fielded Between August 2015 and August 2023 and Between July 2009 and December 2009



according to a stratification frame (table 1 presents an example). If the user belongs to a cell for which a given representation quota has been filled, the user is discarded.

Digital Interview

Users who survive the inclusion criteria comprise our final survey sample. Using the users/id/tweets endpoint of the X API, we

collect the most recent m tweets for each user. We append these tweets to the profile information and pass this augmented mold to the LLM to generate plausible survey responses for a given user. m is a hyper-parameter to be tuned depending on the provenance of the subject pool. Users captured among those discussing trending topics are unlikely to frequently generate text associated with political preferences and, as such, a larger record of their digital

Table 1

Example Implementation of a Stratification Frame with Quota Counter

Cell	Sex	Age	Household Income	Race/Ethnicity	Vote 2020	Quota	Counter
1	Male	65 or Older	Up to 25k	Black	D	2	0
2	Female	25 to 34	Between 25k and 50k	White	D	3	3
3	Male	35 to 44	Between 75k and 100k	Hispanic	D	2	2
4	Female	45 to 54	Between 75k and 100k	White	D	6	6
5	Female	35 to 44	Between 25k and 50k	Black	D	1	1
...
430	Female	25 to 34	Between 25k and 50k	Asian	Stayed Home	1	0
431	Female	65 or Older	Between 50k and 75k	Hispanic	Stayed Home	1	0
432	Female	18 to 24	More Than 100k	Asian	Stayed Home	1	0
433	Male	18 to 24	Between 50k and 75k	Native	Stayed Home	1	0
434	Female	55 to 64	Between 50k and 75k	Asian	Stayed Home	1	0
435	Male	18 to 24	Between 50k and 75k	Asian	Stayed Home	1	0

Notes: This table shows a stratification frame for a target sample size $\Omega^*=1,500$. The snapshot was taken with 647 respondents still to be collected.

behavior is necessary to reasonably inform the LLM's judgment. The opposite is true for users sampled via explicitly political queries, leading to the following heuristic: $mtrending = \lambda \times mpolitics$, $\forall \lambda > 1$.

Listing 2 presents an extract from the feature-extraction prompt. A *features-object* (see listing 3) is appended to this prompt. The *features-object* is given a standard structure: it is composed of a set of elements—each element contains a *title*, which describes a survey question; a set of *categories*, which represent the potential responses; and each category is identified by a unique *symbol*.

The feature-extraction operation considers all features simultaneously and prompts the LLM to produce a joint set of imputed features for a given user. We find that for most tasks, simultaneous feature extraction is preferable to a set of independent prompts, one for each attribute of interest. Separating prompts is an intuitively attractive choice due to its preservation of full independence among extracted features. However, this is extremely inefficient in terms of tokens, given that each prompt must redescribe the background, the mold, and the operations of interest. Prompting the LLM to extract all features simultaneously, by including the full list of desired features in a single prompt, is generally a productive approach.

An important caveat specific to this type of joint extraction pertains to the order in which features are presented in the prompt. The auto-regressive nature of LLMs (LeCun 2023) implies that when multiple answers are presented in response to a given feature-extraction prompt, previous answers will affect the next-token probabilities downstream. To minimize the overall effects of auto-regression on the generated survey-object, we can randomize the order of all features in the feature-extraction prompt so that order effects on the overall sample cancel out with a sufficient number of observations. The auto-regressive nature of the LLM is another reason that we prompt an explanation *before* a given choice is made, as opposed to after—we want to avoid *post hoc* justification of the choice and instead induce the LLM to select a choice that follows from a given line of reasoning.

We innovate LLM feature extraction by prompting a *speculation score*. A classic critique of silicon samples (i.e., synthetic survey responses) is that the data-generating process of the LLM ultimately is unknown. More crucially for PoSSUM, it is uncomfortable to not know how much of the LLM's "own" knowledge—which it has acquired during its training phase—is responsible for a given estimate and how much is simply evident in the X profile and tweets.

To address this concern, we provide the LLM with instructions to generate a speculation score $S \in [0, 100]$ associated with each imputed characteristic. The wording of the prompt makes explicit that speculation refers to the amount of information in the observable data (e.g., the text of the tweets or the pixels of the profile image), which is directly useful to the imputation task and distinguishes this from other types of knowledge that the LLM might leverage. The score has a categorical interpretation that identifies "highly speculative" imputations at $S > 80$.

Model-Based Weighting

As suggested previously, some quotas will be difficult to fill given the highly unrepresentative sampling medium (i.e., the X platform). The weighting method of choice is MRP (Gelman and Little 1997; Lauderdale et al. 2020; Park, Gelman, and Bafumi 2004). We

consider this the obvious weighting choice given the sampling method: the explicit knowledge of unfilled quotas prompts a treatment of these cells as having missing dependent variables. We then can use a hierarchical model, under the ignorability assumption (Van Buuren 2018), to estimate the dependent values for the incomplete cells and stratify these estimates to obtain national- and state-level estimates. This also allows a comprehensive treatment of uncertainty at the cell level, which is liable to provide more realistic intervals on the poll's national vote-share estimates than traditional adjustments.

The target stratification frame, which is derived from the 2021 American Community Survey (US Census Bureau 2021), is extended according to the MRP procedure (Leemann and Wasserfallen 2017) to extend the stratification frame and to include the joint distribution of 2020 Vote Choice as derived from the 2022 Cooperative Election Study (Schaffner, Ansolabehere, and Shih 2023) (see table 1).

The hierarchical model used to generate estimates of the dependent variable of interest imposes structure (Gao et al. 2021) to smooth the learned effects of a model trained on AI-generated data in a sensible way. LLMs can leverage stereotypes in making their imputations (Choenni, Shutova, and van Rooij 2021), which can translate to exaggerated relationships between covariates and dependent variables. Adding structured smoothing to the model allows us to correct for this phenomenon to some degree. We regress the dependent variable—which is assigned a categorical likelihood with SoftMax link—onto sex, age, ethnicity, household income, and 2020 vote. Sex and ethnicity effects are estimated as random effects; state¹ effects are assigned an Intrinsic Conditional Auto-Regressive (ICAR) prior (Besag, York, and Mollié 1991; Donegan 2022; Morris 2018); and date, income, and age effects are given random-walk priors. Separate area-level predictors are created for each dependent variable of interest. The covariates and parameters used in the model for 2024 vote choice are presented in table 2.

We have described the three broad features of our AI polling method: recruitment, sampling, and measurement. They correspond to similar core elements that define telephone and online polling methods. To put the elements of our AI method in context, figure 2 compares our AI approach to these three core activities with those undertaken for telephone and online polling.

RESULTS

During the course of the 2024 US presidential election campaign, we are publishing biweekly vote-share estimates for the candidates. These include the national vote-share estimates for the presidential candidates as well as the vote-share breakouts at the state level, along with vote-share tables for our key sociodemographic profiles. Our national-level vote-share estimates from our August 15–23 and September 7–12, 2024, AI polls are presented in table 3. For our first August wave of the PoSSUM, we estimated that Harris had a national vote share of 46.4% compared to 47.2% for Trump. In the second wave, Harris scored 47.6% and Trump registered 46.8% (i.e., Harris is estimated to have 50.4% of the two-party vote). Table 4 breaks out these estimates by gender. As most election polling has been suggesting, Harris has a significant lead over Trump with women and Trump leads Harris among men. As indicated in table 5, race and ethnic differences between Harris and Trump supporters match those of other polling organizations:

Listing 2

Standardized Feature-Extraction Operation

```
1 I will show you a number of categories to which this user may belong to.
2 The categories are preceded by a title (e.g. "AGE:" or "SEX:" etc.) and a symbol (e.g. "A1",
   "A2" or "E1" etc.).
3 Please select, for each title, the most likely category to which this user belongs to.
4
5 In your answer present, for each title, the selected symbol.
6 Write out in full the category associated with the selected symbol.
7 The chosen symbol / category must be the most likely to accurately represent this user.
8 You must only select one symbol / category per title.
9 A title, symbol and category cannot appear more than once in your answer.
10
11 For each selected symbol / category, please note the level of Speculation involved in this
    selection.
12 Present the Speculation level for each selection on a scale from 0 (not speculative at all,
    every single element of the user data was useful in the selection) to 100 (fully
    speculative, there is no information related to this title in the user data).
13 Speculation levels should be a direct measure of the amount of useful information available
    in the user data.
14 Speculation levels pertain only to the information available in the user data -- namely the
    username, name, description, location, profile picture and tweets from this user -- and
    should not be affected by additional information available to you from any other source.
15 To ensure consistency, use the following guidelines to determine speculation levels:
16
17 0-20 (Low speculation): The user data provides clear and direct information relevant to the
    title. (e.g., explicit mention in the profile or tweets)
18 21-40 (Moderate-low speculation): The user data provides indirect but strong indicators
    relevant to the title. (e.g., context from multiple sources within the profile or tweets
    )
19 41-60 (Moderate speculation): The user data provides some hints or partial information
    relevant to the title. (e.g., inferred from user interests or indirect references)
20 61-80 (Moderate-high speculation): The user data provides limited and weak indicators
    relevant to the title. (e.g., very subtle hints or minimal context)
21 81-100 (High speculation): The user data provides no or almost no information relevant to
    the title. (e.g., assumptions based on very general information)
22
23 For each selected category, please explain at length what features of the data contributed
    to your choice and your speculation level.
24
25 Preserve a strictly structured answer to ease parsing of the text.
26 Format your output as follows (this is just an example, I do not care about this specific
    title or symbol / category):
27
28 **title: AGE**
29 **explanation: ...**
30 **symbol: A1)**
31 **category: 18-25**
32 **speculation: 90**
33
34 YOU MUST GIVE AN ANSWER FOR EVERY TITLE !
35
36 Below is the list of categories to which this user may belong to:
37
38 ---
```

The text is followed by a list of features to be extracted, such as those in Listing 3.

Listing 3
Example of a “Dependent Features” Object

```
1 dep.features <- c(  
2   'CURRENT VOTING PREFERENCES - VOTE CHOICE IN THE 2024 PRESIDENTIAL ELECTION IF THE  
3     ELECTION WERE HELD ON THE DATE OF THEIR MOST RECENT TWEET :  
4   V1) would not vote in the 2024 elections for President  
5   V2) would vote for Donald Trump, the Republican Party candidate  
6   V3) would vote for Kamala Harris, the Democratic Party candidate  
7   V4) would vote for Robert F. Kennedy Jr., who is not affiliated with any major political  
8     party  
9   V5) would vote for Jill Stein, the Green Party candidate  
10  V6) would vote for Chase Oliver, the Libertarian Party candidate  
11  V7) would vote for Dr. Cornel West, who is not affiliated with any political party  
12 )
```

Table 2
Model Predictors and Parameters for the 2024 Vote-Choice Model

Predictor	Level	Description	Index	Domain	Parameter	Prior Correlation Structure
1	Global	/	/	/	α_j	iid
/	State	State ID	l	{1, ..., 54}	λ_{sj}	Spatial (BYM2)
/	Poll	Poll ID	t	{1, ..., T}	$\eta^A P_{tj}$	Random-Walk
/	Individual	Age ID	a	{1, ..., 6}	$\eta^A A_{aj}$	Random-Walk
/		Income ID	h	{1, ..., 5}	$\eta^A H_{hj}$	Random-Walk
/		Sex ID	g	{1, 2}	$\gamma^A G_{gj}$	Unstructured + Shared Variance
/		Race ID	r	{1, ..., 6}	$\gamma^A R_{rj}$	Unstructured + Shared Variance
/		Vote20 ID	v	{1, ..., 5}	$\gamma^A V_{vj}$	Unstructured + Shared Variance
z_1	State	2020 R Share	/	\mathbb{R}	$\beta_{1j}=R$	iid
z_2		On Ballot: RFK Jr.	/	/	$\beta_{1j}=K$	
z_3		On Ballot: Jill Stein	/	/	$\beta_{1j}=G$	
z_4		2020 G Share	/	/	$\beta_{2j}=G$	
z_5		On Ballot: Chase Oliver	/	/	$\beta_{1j}=L$	
z_6		2020 L Share	/	/	$\beta_{2j}=L$	
z_7		On Ballot: Cornel West	/	/	$\beta_{1j}=W$	
z_8		2020 “Stay Home” Share	/	/	$\beta_{1j}=\text{stay_home}$	

Notes: “iid” refers to fully independent parameters or “fixed” effects (see Gelman et al. 2013). “Unstructured + shared variance” priors refer to classic random-intercepts. Random-walk and spatial correlation structures are explained in detail in the article text.

Trump has a lead over Harris with whites; Harris has a Black and Hispanic lead over Trump, and this appears to be growing. The PoSSUM national presidential vote-share estimates, along with demographic breakouts, align with similar estimates by the leading US polling organizations.

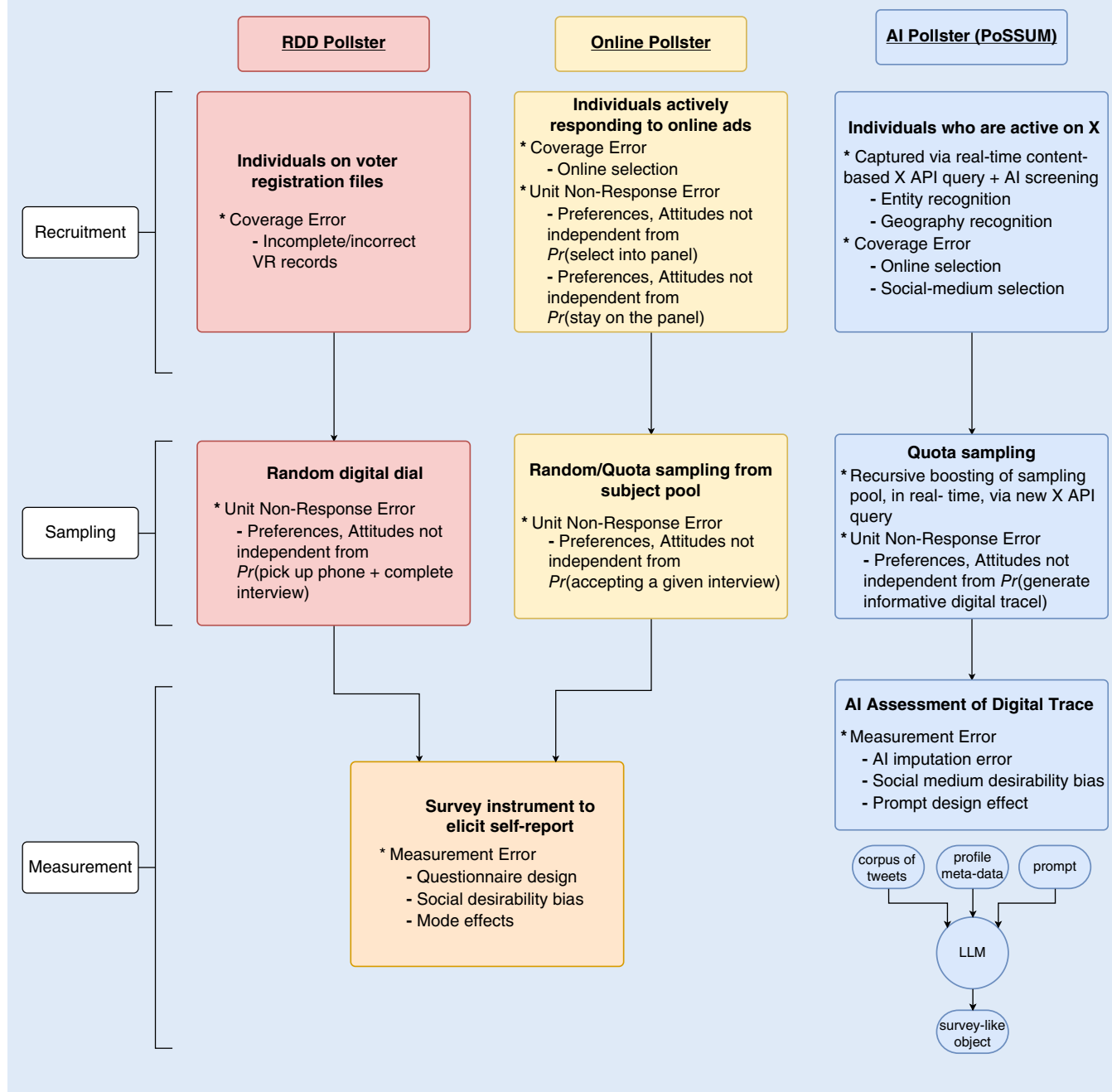
To benchmark our estimates against those of other major US presidential polls, we analyze the vote-share cross-tabulations produced by these polling organizations. This allows us to benchmark our estimates on a biweekly basis. The results for our first two polls are presented in figure 3. Each polling estimate includes a 95% confidence interval. Note that the line in each figure is the overall average for the vote-share estimates of all of the polling organizations. In the case of the Trump vote share, our PoSSUM MRP estimate is slightly higher than this

average in the August poll and almost identical to this average in the September poll. Our vote-share estimate for Harris is lower than most other measurements in both the August and September polls.²

As described previously, the PoSSUM 2024 presidential election study constructs a national sample of the US voting population. It is feasible, however, by using our MRP modeling strategy to generate state-level estimates of candidate vote share. Given that the sampling strategy was not designed to generate representative samples of individual state voting populations, we expect state-level vote-share estimates to be imprecise. Nevertheless, the state-level breakouts provide an additional indication of the robustness of our AI polling method. Figure 4 presents state-level vote-share differences for the two Republican and Democratic

Figure 2

Election Polling: Random Digit Dial, Online, and AI Polling



candidates (i.e., Republican vote share minus Democratic vote share). Posterior distributions are shown for states where polls have been fielded in a comparable period and are published on the *FiveThirtyEight* state-level polling database. There are some states in which the estimates are implausible—Maine, in particular, although its estimates are based on a total of four users across both samples and, as such, should be discounted. We aim to aggregate samples from our biweekly polls, accounting for temporal dynamics in the MRP, to improve state-level coverage. For the important swing states—with the possible exception of Wisconsin—the results track those of other major polling organizations. The dotted

vertical line in the state figures represent these simple polling averages for the state. If we consider Arizona, for example, the polling organization average difference between Republicans and Democrats essentially is zero. We are estimating a 2.2% lead for the Republicans and a probability of a Republican win of 0.80. Although the AI sampling strategy was not designed for estimating vote share at the state level, our state breakouts are generally reasonable, providing further evidence of the robustness of the AI polling method. Our Electoral College vote-share estimates based on the state forecasts from the September poll are 301 for Trump versus 237 for Harris.

Table 3

PoSSUM Poll Estimates of National Presidential Candidates' Vote Share

Pop.	Vote2024	08/15 to 08/23	09/07 to 09/12
LV	Harris (D)	46.4 (44.2, 48.3)	47.6 (45.4, 50)
LV	Trump (R)	47.2 (45.1, 49.3)	46.8 (44.4, 49.6)
LV	RFK Jr. (Ind)	3.7 (2.4, 5.3)	3.0 (1.7, 4.8)
LV	Stein (G)	1.1 (0.4, 2.5)	0.4 (0.1, 1.0)
LV	West (Ind)	0.2 (0.0, 0.7)	0.8 (0.2, 2.1)
LV	Oliver (L)	1.0 (0.5, 2.0)	0.9 (0.4, 1.7)
A	Abstention	30.0 (27.6, 32.2)	24.6 (21.4, 27.6)
A	Turnout	70.0 (67.8, 72.4)	75.4 (72.4, 78.6)

Table 4

PoSSUM Poll Estimates of 2024 Presidential Vote Choice by Sex

Pop.	Vote2024	08/15 to 08/23	09/07 to 09/12
Female			
LV	Harris (D)	51.3 (48.4, 53.7)	52.1 (49.2, 55.1)
LV	Trump (R)	43.4 (40.6, 45.9)	43.1 (40.3, 46.4)
LV	RFK Jr. (Ind)	3.3 (1.9, 5.1)	2.4 (1.0, 4.6)
LV	Stein (G)	1.1 (0.4, 3.0)	0.5 (0.1, 1.6)
LV	West (Ind)	0.1 (0.0, 0.6)	0.9 (0.2, 2.3)
LV	Oliver (L)	0.5 (0.0, 1.6)	0.4 (0.0, 1.2)
A	Abstention	27.3 (24.1, 30.5)	22.1 (17.8, 25.9)
A	Turnout	72.7 (69.5, 75.9)	77.9 (74.1, 82.2)
Male			
LV	Harris (D)	41.0 (38.4, 43.1)	42.6 (40.0, 45.3)
LV	Trump (R)	51.6 (49.0, 54.3)	51.1 (48.1, 54.3)
LV	RFK Jr. (Ind)	4.3 (2.6, 6.3)	3.5 (2.0, 5.7)
LV	Stein (G)	1.0 (0.3, 2.5)	0.2 (0.0, 0.8)
LV	West (Ind)	0.2 (0.0, 0.9)	0.7 (0.2, 2.0)
LV	Oliver (L)	1.5 (0.7, 3.0)	1.3 (0.6, 2.7)
A	Abstention	32.8 (30.1, 35.4)	27.4 (24.0, 30.2)
A	Turnout	67.2 (64.6, 69.9)	72.6 (69.8, 76.0)

CONCLUSION

The PoSSUM 2024 US presidential election vote project explores the feasibility of replacing conventional election polling estimates with an AI survey application. Our goal is to provide the only detailed and open-sourced AI polling estimates of the 2024 US presidential election candidate vote shares. On a biweekly basis during the US presidential campaign, we publish our vote-share estimates at the national and state levels. Additionally, we harmonize estimates being generated by other polling organizations and benchmark them against our detailed estimates.

This article identifies a number of the most serious challenges currently facing election polling. We make the case that LLMs combined with rapidly growing social media content are the solution to the serious challenges facing conventional

Table 5

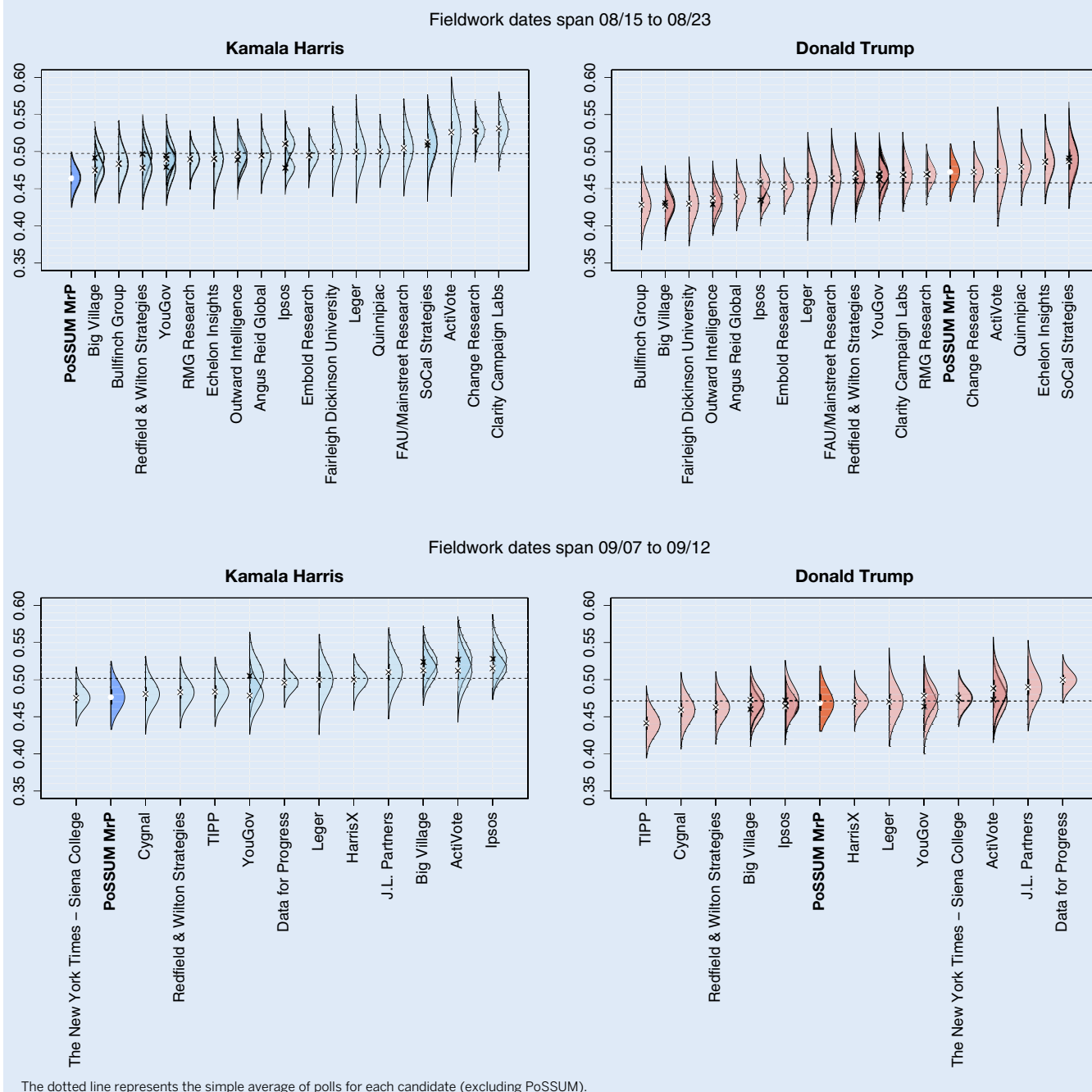
PoSSUM Poll Estimates of 2024 Presidential Vote Choice by Race/Ethnicity

Pop.	Vote2024	08/15 to 08/23	09/07 to 09/12
White			
LV	Harris (D)	40.5 (38.4, 42.4)	41.1 (38.9, 43.5)
LV	Trump (R)	53.2 (50.9, 55.4)	54.2 (51.7, 57.1)
LV	RFK Jr. (Ind)	4.2 (2.6, 6.0)	2.5 (1.3, 4.3)
LV	Stein (G)	0.7 (0.3, 1.9)	0.2 (0.1, 0.8)
LV	West (Ind)	0.1 (0.0, 0.6)	0.8 (0.2, 1.9)
LV	Oliver (L)	0.9 (0.4, 1.9)	0.7 (0.3, 1.5)
A	Abstention	28.0 (25.5, 30.3)	22.6 (19.4, 25.7)
A	Turnout	72.0 (69.7, 74.5)	77.4 (74.3, 80.6)
Black			
LV	Harris (D)	78.1 (72.0, 83.4)	80.0 (73.9, 85.0)
LV	Trump (R)	16.7 (11.6, 21.7)	11.6 (6.6, 17.2)
LV	RFK Jr. (Ind)	1.2 (0.1, 4.0)	4.2 (1.8, 8.4)
LV	Stein (G)	1.5 (0.3, 4.8)	0.6 (0.1, 2.2)
LV	West (Ind)	0.5 (0.1, 2.0)	1.5 (0.4, 4.4)
LV	Oliver (L)	1.0 (0.2, 2.7)	1.0 (0.2, 3.2)
A	Abstention	37.7 (33.2, 42.1)	31.0 (24.0, 37.0)
A	Turnout	62.3 (57.9, 66.8)	69.0 (63.0, 76.0)
Hispanic			
LV	Harris (D)	59.2 (52.7, 64.5)	61.0 (53.5, 67.1)
LV	Trump (R)	35.4 (30.2, 41.3)	33.9 (27.6, 42.0)
LV	RFK Jr. (Ind)	1.7 (0.2, 5.5)	2.7 (0.5, 5.7)
LV	Stein (G)	1.4 (0.2, 5.2)	0.4 (0.0, 2.2)
LV	West (Ind)	0.2 (0.0, 0.7)	0.5 (0.1, 1.6)
LV	Oliver (L)	1.0 (0.2, 3.4)	0.9 (0.2, 2.4)
A	Abstention	38.0 (32.3, 43.1)	32.5 (24.9, 39.1)
A	Turnout	62.0 (56.9, 67.7)	67.5 (60.9, 75.1)
Asian			
LV	Harris (D)	61.9 (49.4, 68.9)	67.4 (59.4, 75.3)
LV	Trump (R)	30.8 (24.8, 41.5)	24.6 (14.3, 33.5)
LV	RFK Jr. (Ind)	1.8 (0.2, 6.0)	4.6 (0.9, 11.7)
LV	Stein (G)	2.5 (0.5, 13.6)	0.4 (0.1, 2.4)
LV	West (Ind)	0.1 (0.0, 0.6)	0.6 (0.1, 1.9)
LV	Oliver (L)	0.8 (0.1, 2.6)	1.2 (0.3, 3.9)
A	Abstention	25.7 (16.9, 32.8)	23.0 (13.6, 30.3)
A	Turnout	74.3 (67.2, 83.1)	77.0 (69.7, 86.4)

polling today. Increasingly unrepresentative samples are a serious challenge for election polling. We address this challenge with a sampling method that leverages voluminous social media content with the rapidly increasing capabilities of LLMs. Of growing concern for election polling is the declining quality of the data generated from a conventional survey interview with humans. There are no humans interviewed in our AI polls. LLMs unobtrusively observe, collect, and analyze human opinions that are expressed by human subjects in social media conversations. Conventional election predictions require a strategy for weighting the data that are generated from increasingly unrepresentative samples. Weighting is accomplished

Figure 3

Benchmarking PoSSUM 2024 US Presidential Vote-Share Estimates with Major Polling Houses



transparently by our PoSSUM method because vote probabilities are estimated using MRP with a stratification frame that guides the LLM in creating our digital sample.

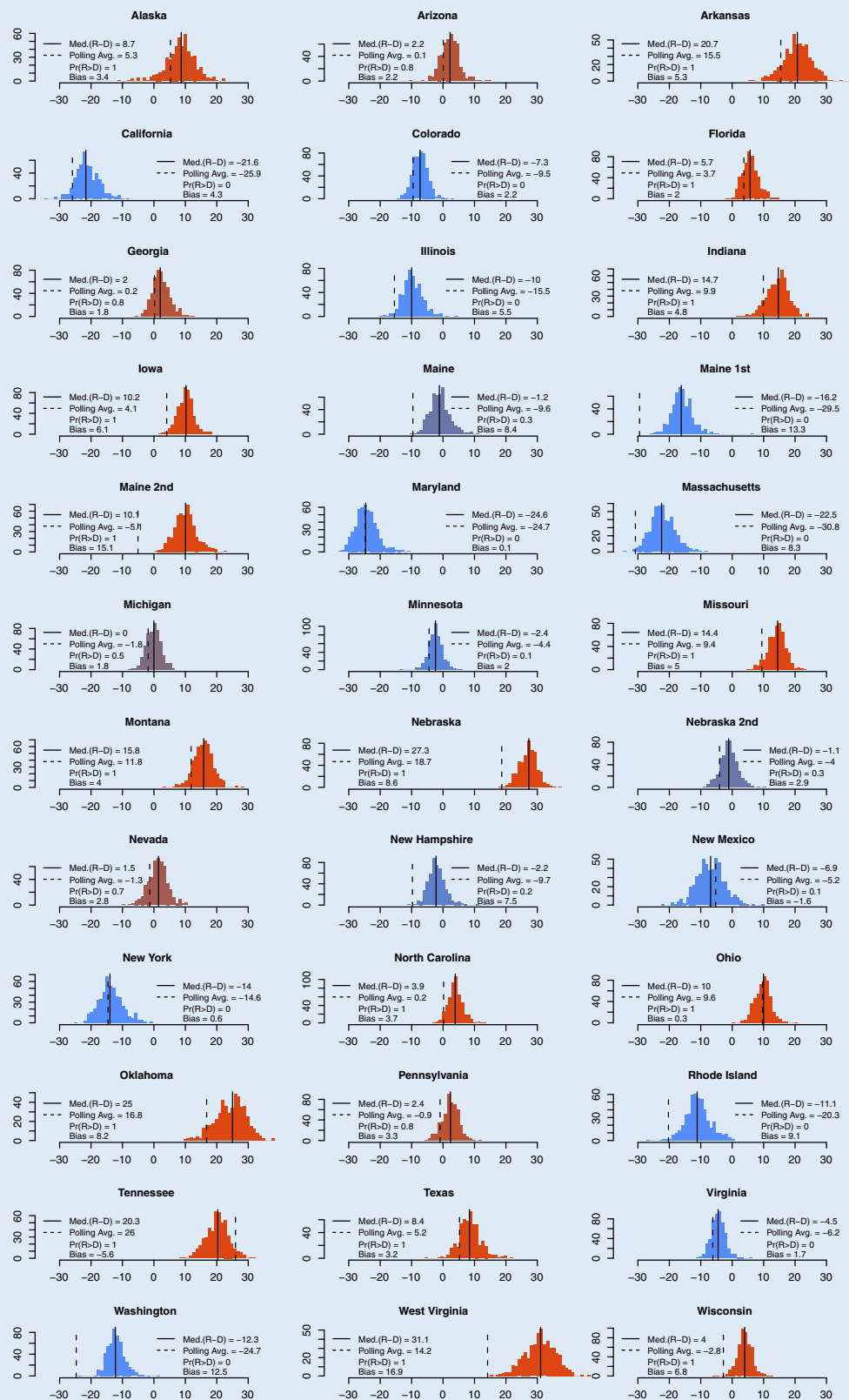
Our initial predictions confirm that presidential candidate vote-share estimates based on AI polling are broadly exchangeable with those of other polling organizations. We present our first two biweekly vote-share estimates for the 2024 US presidential election and benchmark them against those being generated by other polling organizations. Our post-Democratic National Convention presidential vote-share estimates for

Trump (47.2%) and Harris (46.4%) closely track results generated by other polls during the month of August. The subsequent early September (post-debate) PoSSUM vote-share estimates for Trump (46.8%) and Harris (47.6%) again closely track with other national polling being conducted in the United States. An ultimate test for the PoSSUM polling method will be the final preelection vote-share results that we publish before Election Day on November 5, 2024.

LLMs will have an increasingly important role in how we conduct preelection polling. The methods we describe in this

Figure 4

Benchmarking PoSSUM 2024 US Presidential Vote-Share Estimates State Breakouts



The dotted line represents the simple polling average for that state. The x-axis presents the Republican lead in the district. States are ordered alphabetically.

article and the open-sourced code being made available to readers are an important foundation for facilitating the integration of AI into our election polling strategies.

ACKNOWLEDGMENTS

We are grateful for the generous research funding support provided by the University of Oxford “Talking to Machines Project,” the Swiss National Science Foundation (Grant No. 100018M-21519), and Nuffield College Oxford.

DATA AVAILABILITY STATEMENT

The editors have granted an exception to the data policy for this manuscript. In this case, replication code and data are available to reproduce its figures and tables, but there are substantively small differences between the replication and the printed results. This exception was granted because the authors affirmed that these differences are attributable to randomness in the sampling procedure that generates draws from Bayesian posterior distributions and do not change the conclusions of the manuscript.

CONFLICTS OF INTEREST

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

NOTES

1. Because we have an interest in being able to estimate the number of electoral votes won by each candidate, we treat the congressional districts of Nebraska and Maine as separate states.
2. Note: Estimates from the first August poll were reweighted to account for the latest ballot-access information as of September 16, 2024.

REFERENCES

- Bailey, Michael A. 2023. “A New Paradigm for Polling.” *Harvard Data Science Review* 5 (3). <https://doi.org/10.1162/99608f92.9898eede>.
- Bailey, Michael A. 2024. *Polling at a Crossroads: Rethinking Modern Survey Research Methodological Tools in the Social Sciences*. Cambridge, UK: Cambridge University Press.
- Besag, Julian, Jeremy York, and Annie Mollié. 1991. “Bayesian Image Restoration, with Two Applications in Spatial Statistics.” *Annals of the Institute of Statistical Mathematics* 43 (1): 1–20.
- Bradley, Valerie C., Shiro Kuriwaki, Michael Isakov, Dino Sejdinovic, Xiao-Li Meng, and Seth Flaxman. 2021. “Unrepresentative Big Surveys Significantly Overestimated US Vaccine Uptake.” *Nature* 600 (7890): 695–700.
- Cerina, Roberto, and Raymond Duch. 2023. “Artificially Intelligent Opinion Polling.” <https://doi.org/10.48550/arXiv.2309.06029>.
- Cerina, Roberto, and Raymond Duch. 2024. “Replication Data for ‘The 2024 US Presidential Election PoSSUM Poll.’” *PS: Political Science & Politics* Harvard Dataverse. <https://doi.org/10.7910/DVN/NMCTXV>.
- Choenni, Rochelle, Ekaterina Shutova, and Robert van Rooij. 2021. “Stepmothers Are Mean and Academics Are Pretentious: What Do Pretrained Language Models Learn About You?” *arXiv Preprint arXiv:2109.10052*.
- Claassen, Ryan L., and John Barry Ryan. 2024. “Biased Polls: Investigating the Pressures Survey Respondents Feel.” *Acta Politica*. <https://doi.org/10.1057/s41269-024-00356-4>.
- Clinton, Joshua, Jon Cohen, John Lapinski, and Mark Trussler. 2021. “Partisan Pandemic: How Partisanship and Public Health Concerns Affect Individuals’ Social Mobility During COVID-19.” *Science Advances* 7 (2): eabd7204.
- Donegan, Connor. 2022. “Flexible Functions for ICAR, BYM, and BYM2 Models in Stan.” *GitHub*. <https://github.com/ConnorDonegan/Stan-ICAR>.
- Gao, Yuxiang, Lauren Kennedy, Daniel Simpson, and Andrew Gelman. 2021. “Improving Multilevel Regression and Post-Stratification with Structured Priors.” *Bayesian Analysis* 16 (3): 719.
- Gelman, Andrew. 2007. “Struggles with Survey Weighting and Regression Modeling.” *Statistical Science* 22 (2): 153–64.
- Gelman, Andrew, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. 2013. *Bayesian Data Analysis*. Boca Raton, FL: Chapman & Hall/CRC Press.
- Gelman, Andrew, and Thomas C. Little. 1997. “Post-Stratification into Many Categories Using Hierarchical Logistic Regression.” *Statistics Canada* 1997002. www150.statcan.gc.ca/n1/en/catalogue/12-001-X19970023616.
- Huberty, Mark. 2015. “Can We Vote with Our Tweet? On the Perennial Difficulty of Election Forecasting with Social Media.” *International Journal of Forecasting* 31 (3): 992–1007.
- Jackson, Natalie, and Michael Lewis-Beck. 2022. *Forecasting the Party Vote in the 2020 Election*. Lanham, MD: Rowman & Littlefield.
- Jennings, Will, and Christopher Wlezien. 2018. “Election Polling Errors Across Time and Space.” *Nature Human Behaviour* 2 (4): 276–83.
- Keeter, Scott, Nick Hatley, Courtney Kennedy, and Arnold Lau. 2017. “What Low Response Rates Mean for Telephone Surveys.” Washington, DC: Pew Research Center. www.pewresearch.org/methods/2017/05/15/what-low-response-rates-mean-for-telephone-surveys.
- Kennedy, Courtney, Mark Blumenthal, Scott Clement, Joshua D. Clinton, Claire Durand, Charles Franklin, et al. 2018. “An Evaluation of the 2016 Election Polls in the United States.” *Public Opinion Quarterly* 82 (1): 1–33.
- Kennedy, Courtney, and Hannah Hartig. 2019. “Response Rates in Telephone Surveys Have Resumed Their Decline.” Washington, DC: Pew Research Center. <https://coillink.org/20.500.12592/w9j4zj>.
- Krosnick, Jon, Stanley Presser, and Art-Sociology Building. 2009. “Question and Questionnaire Design.” In *Handbook of Survey Research*, ed. James D. Wright and Peter V. Marsden, 3. San Diego, CA: Elsevier.
- Lauderdale, Benjamin E., Delia Bailey, Jack Blumenau, and Douglas Rivers. 2020. “Model-Based Preelection Polling for National and Subnational Outcomes in the US and UK.” *International Journal of Forecasting* 36 (2): 399–413.
- LeCun, Yann. 2023. “Do Large Language Models Need Sensory Grounding for Meaning and Understanding?” In *Workshop on Philosophy of Deep Learning, NYU Center for Mind, Brain, and Consciousness and the Columbia Center for Science and Society*. New York University: Courant Institute & Center for Data Science.
- Leemann, Lucas, and Fabio Wasserfallen. 2017. “Extending the Use and Prediction Precision of Subnational Public Opinion Estimation.” *American Journal of Political Science* 61 (4): 1003–22.
- McClain, Colleen. 2019. “70% of U.S. Social Media Users Never or Rarely Post or Share About Political, Social Issues.” Washington, DC: Pew Research Center.
- Mercer, Andrew, and Arnold Lau. 2023. “Comparing Two Types of Online Survey Samples: Opt-In Samples Are About Half as Accurate as Probability-Based Panels.” Washington, DC: Pew Research Center.
- Morris, Mitzi. 2018. “Spatial Models in Stan: Intrinsic Auto-Regressive Models for Areal Data.” *GitHub Repository*. <https://mc-stan.org/users/documentation/case-studies/icar-stan>.
- Park, David K., Andrew Gelman, and Joseph Bafumi. 2004. “Bayesian Multilevel Estimation with Post-Stratification: State-Level Estimates from National Polls.” *Political Analysis* 12 (4): 375–85.
- Pfeffer, Juergen, Daniel Matter, Kokil Jaidka, Onur Varol, Afra Mashhadi, Jana Lasser, et al. 2023. “Just Another Day on Twitter: A Complete 24 Hours of Twitter Data.” *Research Gate*. DOI:10.48550/arXiv.2301.11429.
- Rothschild, David, Sharad Goel, Houshmand Shirani-Mehr, and Andrew Gelman. 2018. “Disentangling Bias and Variance in Election Polls.” *Journal of the American Statistical Association* 113 (522): 607–14.
- Schaffner, Brian, Stephen Ansolabehere, and Marissa Shih. 2023. “Cooperative Election Study Common Content, 2022.” Harvard Dataverse. <https://doi.org/10.7910/DVN/PR4L8P>.
- Sturgis, Patrick, Nick Baker, Mario Callegaro, Stephen Fisher, Jane Green, Will Jennings, et al. 2016. “Report of the Inquiry into the 2015 British General Election Opinion Polls.” London: Market Research Society and British Polling Council.
- US Census Bureau. 2021. “American Community Survey 5-Year Estimates.” Washington, DC: US Census Bureau.
- Van Buuren, Stef. 2018. *Flexible Imputation of Missing Data*. Boca Raton, FL: CRC Press.
- Wu, Patrick Y., Jonathan Nagler, Joshua A. Tucker, and Solomon Messing. 2023. “Large Language Models Can Be Used to Estimate the Latent Positions of Politicians.” New York University: Center for Social Media and Politics.