POLITICS

Special Issue on Forecasting the 2024 US Elections

The 2024 U.S. Presidential Election PoSSUM Poll

Roberto Cerina
Institute for Logic, Language and Computation
University of Amsterdam
r.cerina@uva.nl

Raymond Duch
Nuffield College
University of Oxford
raymond.duch@nuffield.ox.ac.uk

September 30, 2024

Abstract

The initial predictions presented in this essay 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 bi-weekly vote share estimates for the 2024 U.S. presidential election, and benchmark against those being generated by other polling organizations. Our post-Democratic convention national 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 other national polling being conducted in the U.S. An ultimate test for the PoSSUM polling method will be the final pre-election vote share results that we publish prior to election day November 5, 2024.

This is a "preproof" accepted article for PS: Political Science & Politics. This version may be subject to change during the production process.

DOI: 10.1017/S1049096524000982

Introduction

We survey citizens' voting preferences to understand, or explain, their voting decision but also to predict election outcomes. Since 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) point out, the overall prediction error in pre-election national polls has actually declined somewhat reflecting the rising number of polls being produced and individuals polled. On the other hand, particularly over 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 pre-election polling has become increasingly challenging. This essay 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. As an illustration we estimate the vote shares of 2024 U.S. presidential candidates on a bi-weekly basis using our artificially intelligent polling method – PoSSUM, a **Pro**tocol for **S**urveying **S**ocial-media **U**sers with **M**ultimodal 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 over the telephone or online. Response rates for traditional random digit dial (rdd) polls are now well below 10% (Keeter et al., 2017; Kennedy and Hartig, 2019). Similarly 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 are often 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 the how survey respondents answer these questions suggest that the survey/interview likely biases polling results. A 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 reported voting intention (Claassen and Ryan, 2024) and likely voting turnout.

A third critical, and increasingly challenging, element of the polling exercise is weighting of the sampled respondents (Gelman, 2007; Houshmand Shirani-Mehr and Gelman, 2018). Most importantly non-response 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) but also surveys conducted in other areas (Bradley et al., 2021). As a result scholars pay 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 essay introduces an alternative AI-driven approach to polling that significantly reduces the estimation biases associated with these three features of traditional polling. Our bi-weekly PoSSUM estimation of the 2024 U.S. presidential vote share provides an opportunity to test this claim. This essay proceeds by first describing how AI polling is likely to reshape the future of election polling. A section describing the methodology then follows. We then present the results of our first two bi-weekly estimates of 2024 presidential vote share, benched against other polling organizations. We then conclude the discussion.

The AI Future of Polling?

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

Population The "target" population of interest is likely voters in the 2024 U.S. presidential elections. Our data collection is guided by a stratification frame that represents the population of the U.S. We populate the relevant cells of this stratification frame with population figures from the American Community Survey. 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 we (Cerina and Duch, 2023) and others (Lauderdale et al., 2020a) 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 U.S. election. As we pointed out, these samples are increasingly unrepresentative and problematic. In many cases, the sample is not from the U.S. 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 an 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 phone or over the internet. 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 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 needs to 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 will be 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 U.S. accounted for 20% or about 70 million tweets generated by 8 million accounts. The authors' analysis of hashtags suggests that about 5% had a political theme (ignoring Iranian protest hashtags that account for 15% of hashtags at the time). For our 2024 presidential vote share estimates we sample from these U.S. X accounts. Previous efforts to utilize X (formerly Twitter) 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 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 on line. As we pointed out earlier, the "interview" needs to be constructed and administered and is the source of significant measurement error (Krosnick et al., 2009). This is problematic since the accuracy of election polling is very much 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 observe, unobtrusively, digital conversations and infer preferences and opinions from the conversations – they are, for example, instructed to infer vote choice from the digital traces they "digest".

While AI polling is unlikely to suffer from these conventional sources of measurement error, other types of measurement may be prevalent. Of particular concern for our method, from a measurement perspective, is 1) whether individuals are misrepresenting their sincere political preferences; and 2) whether 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. While 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 profile estimate it generates (e.g., the profile's gender, likely vote, etc.).

Weighting Our method of course 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 here – we are not observing the counterfactual identical individuals with each of our socio-political stratification frame profiles who are not X users. These "counterfactual" individuals may not be "missing at random" hence introducing bias into our estimates of vote share (Bailey, 2023, 2024).

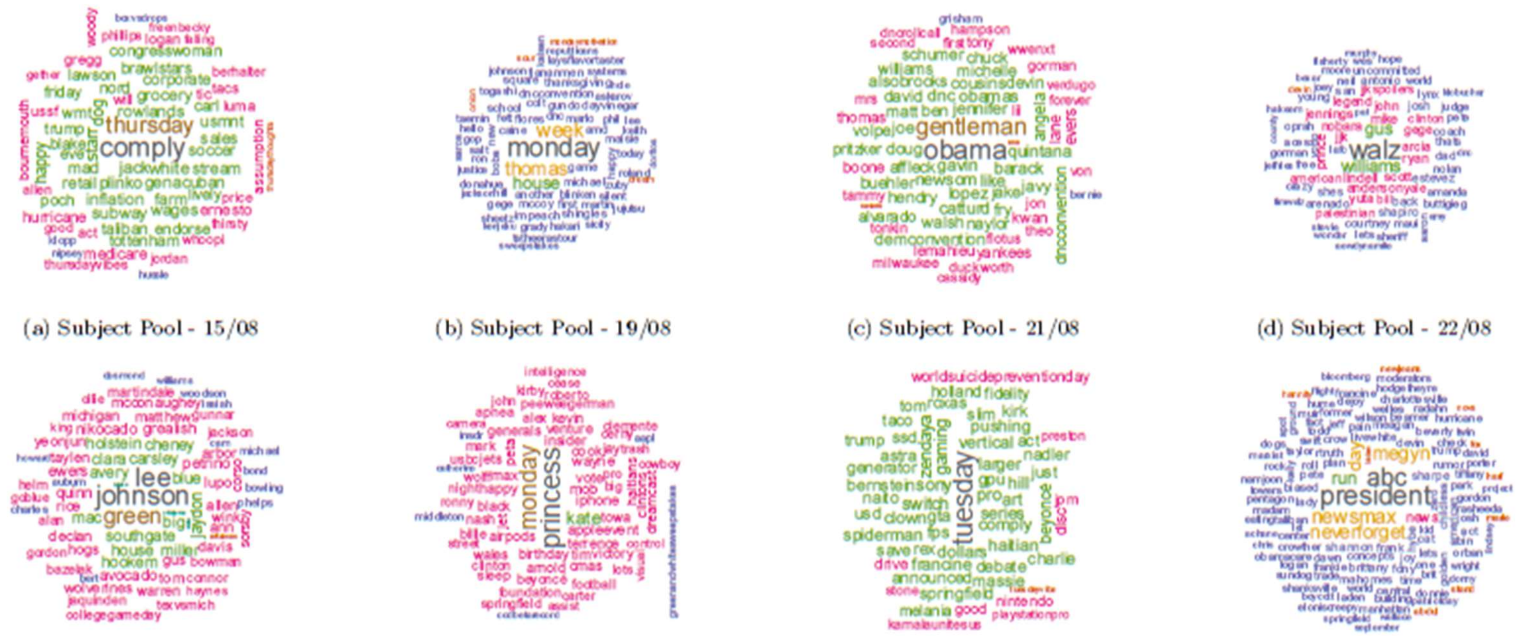
PoSSUM and the 2024 U.S. Presidential Elections: The Method

As with conventional polling, our data collection focuses on sampling and conducting interviews. Our approach is tailored to the X API, which uses the digital trace of X users as the mould for LLM generation. But 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/search API endpoint. Users who have taken part in conversations related to the query over the last 7 days (as per the limits of X 's *Basic* API tier) are gathered to build the digital subject pool. Listing 1 presents an example query for the X API. This sort of query is very likely to yield users who explicitly express opinions about candidates, and will therefore yield highly informative digital traces, that the LLM can annotate with confidence. However selection effects loom large with this sort of query – the kind 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* (available via https://trends24.in/united-states/). Trending topics may still be related to politics, for example during party conventions or televised debates, though they are more likely to be associated with events such as sports, concerts, marketing campaigns, famous people or otherwise *viral* online content. Users engaging with this set of queries are far more likely to be *normies*, who pay relatively little attention to the politics, and can therefore help balance the high-attention selection associated with the query in Listing 1. An illustration of the trending topics associated with users in our digital panel is available in Figure 1.

Listing 1: Search terms for tweets related to candidates involved in the US 2024 presidential election.

```
query <-
  "(
    Kamala OR VP OR KamalaHarris OR                          # Democratic candidate terms
    MAGA OR Trump OR realDonaldTrump OR                      # Republican candidate terms
    Robert Kennedy OR RFK OR RobertKennedyJr OR RFKJr
    OR KennedyShanahan24 OR Kennedy24 OR                                      # RFK terms
    Cornel West OR Dr. West OR CornelWest OR                          # Cornel West terms
    Jill Stein OR DrJillStein OR                                    # Green candidate terms
    ChaseForLiberty                                           # Libertarian candidate terms
    )"
  -from:VP -from:KamalaHarris                            # Don't sample candidate profiles
  -from:realDonaldTrump
  -from:RobertKennedyJr
  -from:CornelWest
  -from:DrJillStein
  -from:ChaseForLiberty
  -is:retweet"
```



(a) Subject Pool - 15/08 (b) Subject Pool - 19/08 (c) Subject Pool - 21/08 (d) Subject Pool - 22/08

(e) Subject Pool - 07/09 (f) Subject Pool - 09/09 (g) Subject Pool - 10/09 (h) Subject Pool - 11/09

Fig 1: Word-cloud presenting words from the 'trending' queries to the 𝕏 API, for PoSSUM polls fielded between 15/08 and 23/08 and 7/09 and 12/09. The words are weighted by the number of users associated with each word.

The digital panel is then further filtered, according to a number of sequential exclusion criteria. This is done for two reasons: First, it contributes to data quality by ensuring that the digital traces belong to real existing users within the population of interest. Second, 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 who have empty self-reported location information and users for whom we have already gathered a digital trace within the last τ days (to avoid over-reliance on frequently-active users). Users who do not represent a real offline person, including accounts for organisations, services or bots, are discarded. Users who reside outside of the U.S. are discarded. Here we rely again 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, according to a stratification frame (see Table 1 for an example). If the user belongs to a cell for which a given representation quota has been filled, the user is discarded.

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 75000	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

Table 1: Example implementation of a stratification frame with quota counter, for a target sample size $\Omega^\star = 1,500$. This is a snapshot taken with 647 respondents still to be collected.

Digital Interview Users who survive the inclusion criteria make up 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 mould to the LLM in order 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 amongst those discussing trending topics are unlikely to frequently generate text associated with political preferences, and as such a larger record of their digital behaviour 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: $m^{\text{tredning}} = \lambda \times m^{\text{politics}}, \ \forall \lambda > 1.$

Listing 2 presents an extract from the feature extraction prompt. A *features-object* (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 the given user. We find 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 between extracted features. But this is extremely inefficient in terms of tokens, given that each prompt has to re-describe the background, the mould 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 sort 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, earlier answers will affect the next-token-probabilities downstream. To minimise the overall effects of auto-regression on the generated survey-object, we can randomise the order of all features in the feature-extraction prompt, so that order effects on the overall sample cancel-out with a large enough number of observations. The auto-regressive nature of the LLM is also the reason we prompt an explanation *before* a given choice is made, as opposed to after – we wish to avoid post-hoc justification of the choice, and instead induce the LLM to pick a choice which follows from a given line of reasoning.

We innovate LLM feature extraction by prompting a *speculation score*. A classic critique of silicon samples is that the data generating process of the LLM is ultimately unknown. More crucially for PoSSUM, it is uncomfortable to be in the dark as to 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 just 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 kinds of knowledge the LLM might leverage. The score has a categorical interpretation, which identifies "highly speculative" imputations at $S > 80$.

Listing 2: Standardised feature extraction operation. The text is followed by a list of features to be extracted, such as those in Listing 3.

```
I will show you a number of categories to which this user may belong to.
The categories are preceded by a title (e.g. "AGE:" or "SEX:" etc.) and a symbol (e.g. "A1", "A2" or "E1" etc.).
Please select, for each title, the most likely category to which this user belongs to.

In your answer present, for each title, the selected symbol.
Write out in full the category associated with the selected symbol.
The chosen symbol / category must be the most likely to accurately represent this user.
You must only select one symbol / category per title.
A title, symbol and category cannot appear more than once in your answer.

For each selected symbol / category, please note the level of Speculation involved in this selection.
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).
Speculation levels should be a direct measure of the amount of useful information available in the user data.
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.
To ensure consistency, use the following guidelines to determine speculation levels:

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)
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)
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)
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)
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)

For each selected category, please explain at length what features of the data contributed to your choice and your speculation level.

Preserve a strictly structured answer to ease parsing of the text.
```

```
Format your output as follows (this is just an example, I do not car e about this specific
    title or symbol / category):

**title: AGE**
**explanation: ...**
**symbol: A1)**
**category: 18-25**
**speculation: 90**

YOU MUST GIVE AN ANSWER FOR EVERY TITLE !

Below is the list of categories to which this user may belong to:

---
```

Listing 3: Example of a “dependent features” object.

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

Model-based Weighting As we have hinted at in earlier paragraphs, some quotas will be difficult to fill given the highly unrepresentative sampling medium (the X platform). The weighting method of choice here is Multilevel Regression with Post-Stratification (MrP) (Gelman and Little, 1997; Lauderdale et al., 2020b; Park et al., 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 can then 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 (U.S. Census Bureau, 2021), is extended according to the MrsP (Leemann and Wasserfallen, 2017) procedure to extend the stratification frame, and include the joint distribution of 2020 Vote Choice as derived from the 2022 Cooperative Election Study (CES) (Schaffner et al., 2023) (as seen in 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 et al., 2021), which can translate to exaggerated relationships between covariates and dependent variables. Adding structured smoothing to the model allows us to correct for this phenomena, 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; state[1] effects are assigned an Intrinsic Conditional Auto-regressive (ICAR) prior (Besag et al., 1991; Donegan, 2022; Morris, 2018); date, income and age effects are given random-walk priors. Separate area-level predictors are created for each dependent variable of interest. Table 2 presents the covariates and parameters used in the model for 2024 vote choice.

[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.

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}	η^P_{tj}	random-walk
/	individual	age_id	a	{1,...,6}	η^A_{aj}	random-walk
/		income_id	h	{1,...,5}	η^H_{hj}	random-walk
/		sex_id	g	{1,2}	γ^G_{gj}	unstructured + shared variance
/		race_id	r	{1,...,6}	γ^R_{rj}	unstructured + shared variance
/		vote20_id	v	{1,...,5}	γ^V_{vj}	unstructured + shared variance
z_1	state	2020 *R* share	/	$\mathbb{R}$	$\beta_{1j=R}$	iid
z_2		On ballot: R.F.K. 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			$\beta_{1j=\text{stay\_home}}$	

Table 2: Model Predictors and Parameters for the 2024 vote-choice model. 'iid' refers to fully independent parameters, or 'fixed' effects Gelman et al. (2013). 'unstructured + shared variance' priors refers to classic random-intercepts. Random-walk and spatial correlation structures are explained in detail below. Note: the Democrat choice "D" is taken as the reference category, hence it has no associated predictor.

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.

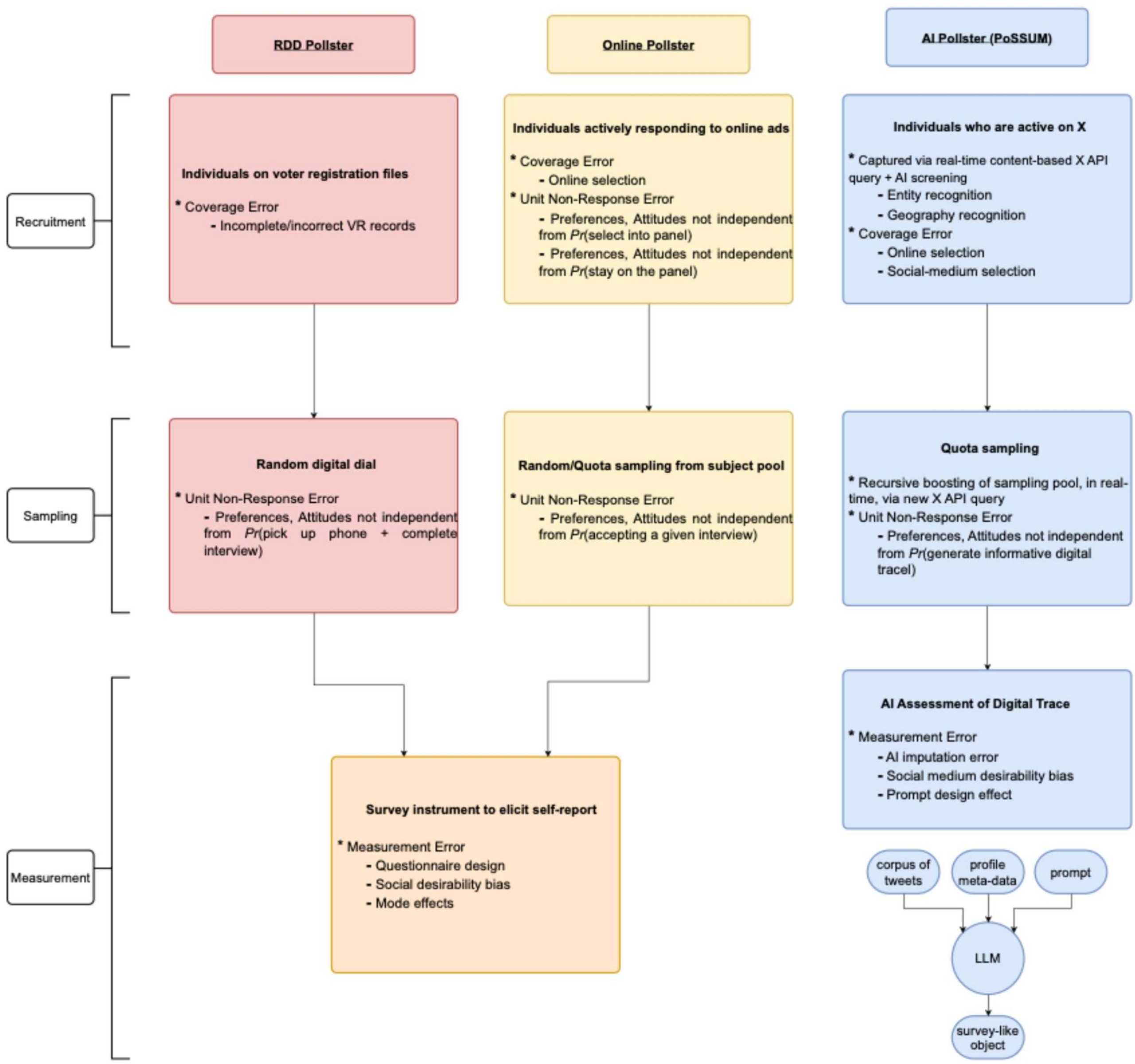


Fig 2: Election Polling: Random Digit Dial, Online, and AI Polling

PoSSUM and the 2024 U.S. Presidential Elections: Results

Over the course of the 2024 U.S. Presidential Election campaign we are publishing bi-weekly vote share estimates for the candidates. These include the national vote share estimates for the Presidential candidates but also the vote share breakouts at the state level along with vote share tables for our key socio-demographic profiles. Our national-level vote share estimates from our August 15-23, 2024 and September 7-12, 2024 AI polls are presented in Table 3. For our first August wave of the PoSSUM we estimated Harris had a national vote share of 46.4% compared to 47.2% for Trump. In the second wave, Harris scored 47.6% while Trump registered 46.8%. Table 4 breaks these estimates out by gender. As most election polling has been suggesting, Harris has a significant lead over Trump with women and Trump leads Harris amongst men. As Table 5 indicates race and ethnic differences between Harris and Trump supporters match those of other polling organizations: 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 national presidential vote share estimates, along with demographic breakouts, align with similar estimates by the leading U.S. polling organizations.

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)

In order to benchmark our estimates against those of other major U.S. Presidential polls we analyze the vote share cross-tabulations produced by these polling organizations. This allows us to benchmark our estimates on a bi-weekly basis. Figure 3 presents the results for our first two polls. Each of the polling estimates includes a 95% confidence intervals. Note that the line in each figure is the overall average for the vote share estimates of all the polling organizations. In the case of the Trump vote share, our PoSSUM MrP share 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].

[2]Note: estimates form the 1^{st} August poll were re-weighted to account for the latest ballot-access information as of

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)

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)

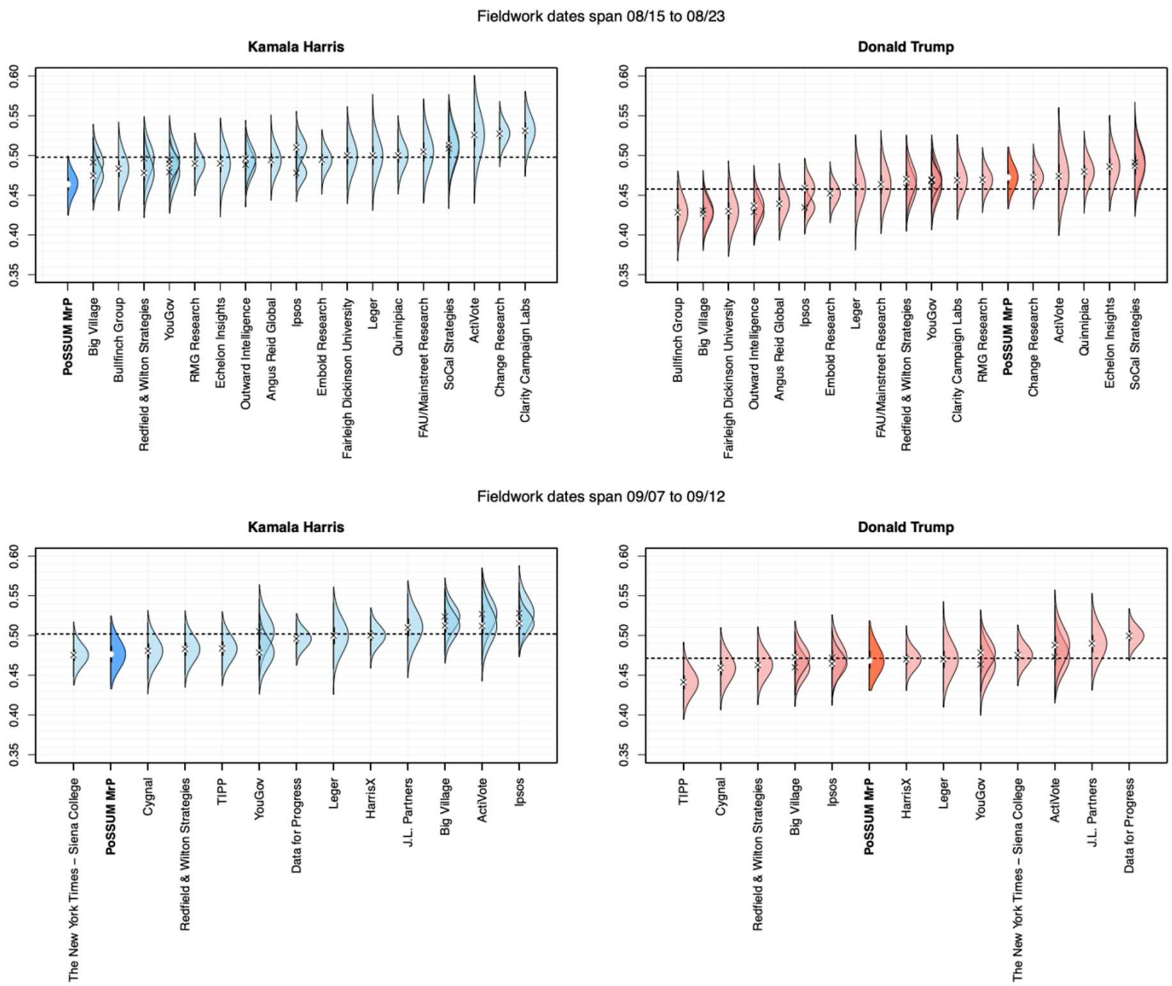


Fig 3: Benchmarking `PoSSUM` 2024 U.S. Presidential Vote Share Estimates with Major Polling Houses. The dotted line represents the simple average of polls for each candidate (excluding `PoSSUM`).

16/09/2024.

As we described earlier, the PoSSUM 2024 Presidential study constructs a national sample of the U.S. voting population. It is feasible though employing 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 very noisy. 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 candidates (Republican vote share minus Democratic vote share). Posterior distributions are shown for states where polls have been fielded in a comparable time period, and are published on the FiveThiryEight state-level polling database. There are some states in which the estimates are implausible – Maine, in particular, though its estimates are based on a total of 4 users across both samples and should as such be discounted. We aim to aggregate samples from our bi-weekly 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 take Arizona, for example, the polling organization average difference between Republicans and Democrats is essentially zero. We are estimating a 2.2 percent lead for the Republicans and a probability of a Republican win of 0.80. While 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.

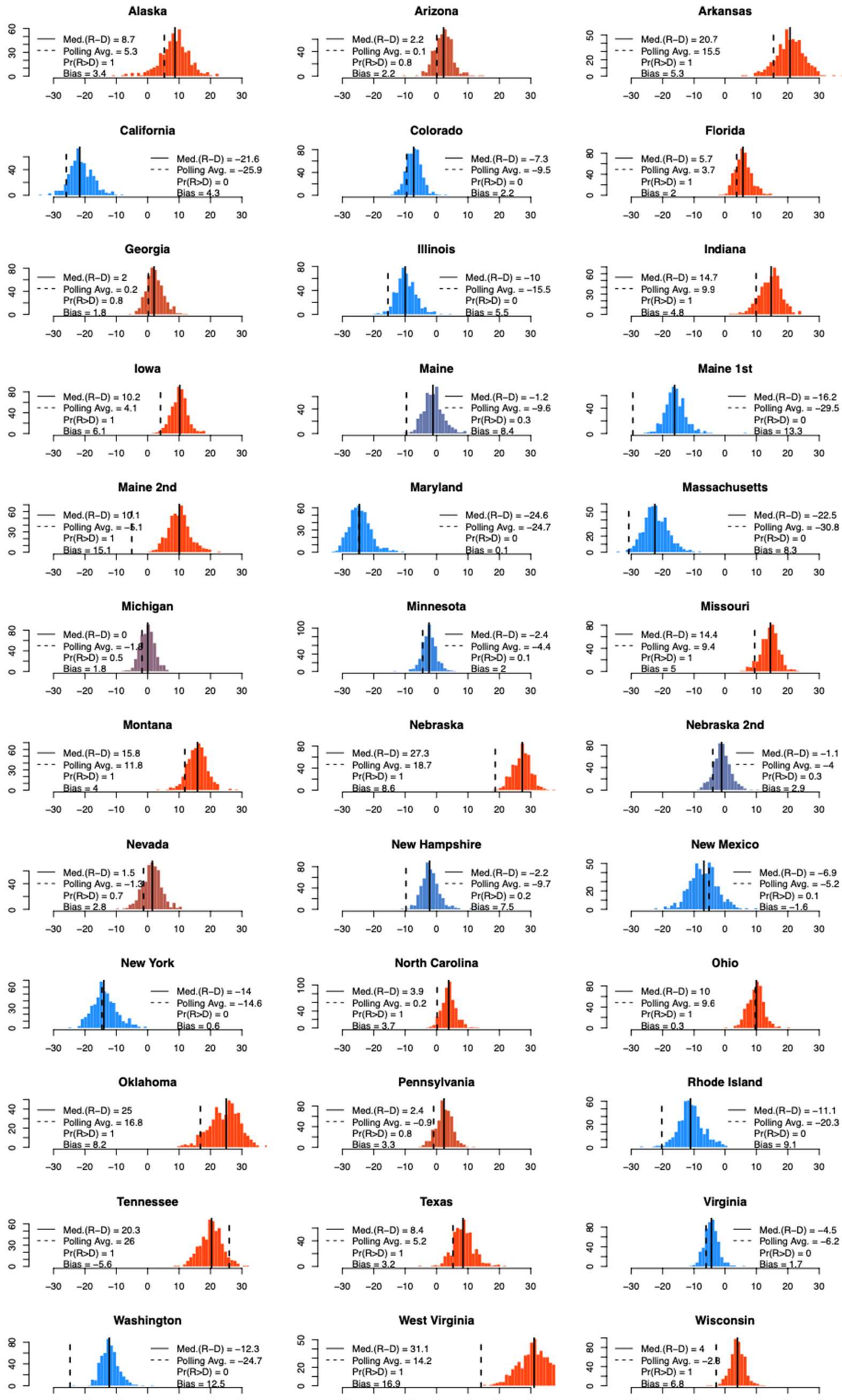


Fig 4: Benchmarking PoSSUM 2024 U.S. Presidential Vote Share Estimates State Breakouts. The dotted line represents the simple polling average for that state. The x-axis presents the Republican lead in the district. States are ordered alphabetically.

Conclusion

The PoSSUM 2024 U.S. 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 U.S. presidential election candidate vote shares. On a bi-weekly basis during the U.S. presidential campaign we publish our vote share estimates at the national and state level. Additionally, we harmonize estimates being generated by other polling organizations and benchmark them against our detailed estimates.

The essay 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 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 observe, collect, and analyze, unobtrusively, human opinions that are expressed by human subjects in social media conversations. Conventional election predictions require a strategy for weighting the data that is generated from increasingly unrepresentative samples. Weighting is accomplished in a transparent fashion by our PoSSUM method because vote probabilities are estimated using MrP with a stratification frame that guides the LLM in creating our digital sample.

The initial predictions presented in the essay 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 bi-weekly vote share estimates for the 2024 U.S. presidential election, and benchmark against those being generated by other polling organizations. Our post-Democratic convention national 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 other national polling being conducted in the U.S. An ultimate test for the PoSSUM polling method will be the final pre-election vote share results that we publish prior to election day November 5, 2024.

Large language models will play an increasingly important role in how we conduct pre-election polling. The methods we have described in this essay, and the open-sourced code being made available to readers, is an important foundation for facilitating the integration of AI into our election polling strategies.

References

Bailey, Michael A., "A New Paradigm for Polling," *Harvard Data Science Review*, jul 27 2023, *5* (3).

— , *Polling at a Crossroads: Rethinking Modern Survey Research* Methodological Tools in the Social Sciences, Cambridge University Press, 2024.

Besag, Julian, Jeremy York, and Annie Mollié, "Bayesian image restoration, with two applications in spatial statistics," *Annals of the institute of statistical mathematics*, 1991, *43* (1), 1–20.

Bradley, Valerie C., Shiro Kuriwaki, Michael Isakov, Dino Sejdinovic, Xiao-Li Meng, and Seth Flaxman, "Unrepresentative big surveys significantly overestimated US vaccine uptake," *Nature*, 2021, *600* (7890), 695–700. Published online: 2021/12/01.

Buuren, Stef Van, *Flexible imputation of missing data*, CRC press, 2018.

Cerina, Roberto and Raymond Duch, "Artificially Intelligent Opinion Polling," 2023.

Choenni, Rochelle, Ekaterina Shutova, and Robert van Rooij, "Stepmothers are mean and academics are pretentious: What do pretrained language models learn about you?," *arXiv preprint arXiv:2109.10052*, 2021.

Claassen, Ryan L. and John Barry Ryan, "Biased polls: investigating the pressures survey respondents feel," *Acta Politica*, 2024. Published online: 2024/07/29.

Clinton, J., J. Cohen, J. Lapinski, and M. Trussler, "Partisan pandemic: How partisanship and public health concerns affect individuals' social mobility during COVID-19," *Science Advances*, 2021, *7* (2), eabd7204.

Donegan, Connor, "Flexible Functions for ICAR, BYM, and BYM2 Models in Stan," *GitHub*, 2022.

Gao, Yuxiang, Lauren Kennedy, Daniel Simpson, and Andrew Gelman, "Improving multilevel regression and poststratification with structured priors," *Bayesian Analysis*, 2021, *16* (3), 719.

Gelman, Andrew, "Struggles with Survey Weighting and Regression Modeling," *Statistical Science*, 2007, *22* (2), 153 – 164.

— **and Thomas C Little**, "Poststratification into many categories using hierarchical logistic regression," 1997.

— **, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin**, *Bayesian data analysis*, Chapman and Hall/CRC, 2013.

Huberty, Mark, "Can we vote with our tweet? On the perennial difficulty of election forecasting with social media," *International Journal of Forecasting*, 2015, *31* (3), 992–1007.

Jackson, N. and Michael Lewis-Beck, *Forecasting the party vote in the 2020 election*, Rowman & Littlefield,

Jennings, Will and Christopher Wlezien, "Election polling errors across time and space," *Nature Human Behaviour*, 2018, *2* (4), 276–283. Published online: 2018/04/01.

Keeter, Scott, Nick Hatley, Courtney Kennedy, and Arnold Lau, "What low response rates mean for telephone surveys," *Pew Research Center*, 2017, *15*, 1–39.

Kennedy, Courtney and Hannah Hartig, "Response rates in telephone surveys have resumed their decline," 2019.

— **, Mark Blumenthal, Scott Clement, Joshua D Clinton, Claire Durand, Charles Franklin, Kyley McGeeney, Lee Miringoff, Kristen Olson, Douglas Rivers et al.**, "An evaluation of the 2016 election polls in the United States," *Public Opinion Quarterly*, 2018, *82* (1), 1–33.

Krosnick, Jon, Stanley Presser, and Art-Sociology Building, "Question and Questionnaire Design," *Handbook of Survey Research*, 03 2009.

Lauderdale, Benjamin E., Delia Bailey, Jack Blumenau, and Douglas Rivers, "Model-based pre-election polling for national and sub-national outcomes in the US and UK," *International Journal of Forecasting*, 2020, *36* (2), 399 – 413.

Lauderdale, Benjamin E, Delia Bailey, Jack Blumenau, and Douglas Rivers, "Model-based pre-election polling for national and sub-national outcomes in the US and UK," *International Journal of Forecasting*, 2020, *36* (2), 399–413.

LeCun, Y, "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" 2023.

Leemann, Lucas and Fabio Wasserfallen, "Extending the use and prediction precision of subnational public opinion estimation," *American journal of political science*, 2017, *61* (4), 1003–1022.

McClain, Colleen, "70% of U.S. social media users never or rarely post or share about political, social issues," May 2019. Pew Research Center.

Mercer, Andrew and Arnold Lau, "Comparing Two Types of Online Survey Samples Opt-in samples are about half as accurate as probability-based panels," 2023.

Morris, Mitzi, "Spatial models in stan: Intrinsic auto-regressive models for areal data," *GitHub repository*, 2018.

Park, David K, Andrew Gelman, and Joseph Bafumi, "Bayesian multilevel estimation with poststratification: State-level estimates from national polls," *Political Analysis*, 2004, *12* (4), 375–385.

Pfeffer, Juergen, Daniel Matter, Kokil Jaidka, Onur Varol, Afra Mashhadi, Jana Lasser, Dennis Assenmacher, Siqi Wu, Diyi Yang, Cornelia Brantner, Daniel M. Romero, Jahna Otterbacher, Carsten Schwemmer, Kenneth Joseph, David Garcia, and Fred Morstatter, "Just Another Day on Twitter: A Complete 24 Hours of Twitter Data," 2023.

Rothschild, Sharad Goel Houshmand Shirani-Mehr David and Andrew Gelman, "Disentangling Bias and Variance in Election Polls," *Journal of the American Statistical Association*, 2018, *113* (522), 607–614.

Schaffner, Brian, Stephen Ansolabehere, and Marissa Shih, "Cooperative Election Study Common Content, 2022," 2023. Dataset.

Sturgis, Patrick, Baker Nick, Callegaro Mario, Fisher Stephen, Green Jane, Will Jennings, Kuha Jouni, Lauderdale Ben, and Smith Patten, “Report of the inquiry into the 2015 British general election opinion polls,” 2016.

U.S. Census Bureau, “American Community Survey, 2021 American Community Survey 5-Year Estimates,” U.S. Census Bureau, American Community Survey (ACS) 2021. Accessed: 2024-08-27.

Wu, Patrick Y., Jonathan Nagler, Joshua A. Tucker, and Solomon Messing, “Large Language Models Can Be Used to Estimate the Latent Positions of Politicians,” 2023.