

PoSSUM: A **P**rotocol for **S**urveying **S**ocial-media **U**sters with **M**ultimodal LLMs

Version 1.0.0 alpha

Roberto Cerina

`r.cerina@uva.nl`

Institute for Logic, Language and Computation
University of Amsterdam

August 30, 2024



Contents

1	Introduction	1
2	Protocol	2
2.1	get_pool	2
2.2	poll_users	4
2.2.1	Prompt Building Framework	6
2.2.2	Inclusion Criteria	8
2.2.3	General Feature Extraction Framework	11
2.2.4	Extracting Response Features	16
2.2.5	Background-informed Features	22
3	2024 US Pre-Election Polling	25
3.1	Subject Pool	25
3.2	Unfilled Quotas	28
3.3	Speculation	30
3.4	Representative Inference	32
3.5	Topline & Crosstabs	35
3.6	Cross-Pollsters Comparison	39

Acknowledgements

The development of this protocol was partially funded by the *Talking to Machines* initiative (<https://talkingtomachines.org>) at Nuffield College, University of Oxford. I am grateful to Ray Duch for believing in this project and supporting my efforts to realise it.

1 Introduction

This research note describes PoSSUM, an open source¹ protocol to poll social-media users unobtrusively and inexpensively using multimodal Large Language Models (LLMs). The protocol seeks to address the skepticism [20] surrounding Artificially Intelligent (AI) polling by establishing a methodology comparable to that used by traditional on-line panels [28]. Concerns around AI polling are summarised aptly by this anonymous review to a related paper [5]: ‘... *The goal of polling is quite simple: TO. LEARN. FROM. PEOPLE. NOW. I believe this model here does not learn, not from people, and not now.*’. Three necessary conditions emerge from this animated critique – to be a credible alternative to random digit dial (rdd) or self-selected online panels, Silicon samples [1] must enable novel *learning* – i.e. must contain more information than the *mould*² on which they are based; must be *human-exchangeable* – i.e. conditional on the same generating process, they must produce a distribution of responses which matches that of humans; must be *timely* – i.e. we must be able to learn about changes in preferences and attitudes over time by studying these samples, and these changes should be reflective of true societal dynamics, rather than artifacts of data engineering.

PoSSUM proposes to poll the public by inferring the attitudes and preferences of real-life social-media users with multimodal LLMs. What I describe is an approach tailored to the X API, which uses the digital trace of X users as the mould for LLM generation, but can be extended to any social-media which allows querying of a user panel via user- and content-level queries. PoSSUM is composed of two principal routines: `get_pool` (Pseudo-code 1) is used to identify a set of potential users of interest based on keywords they have used in their recent tweets; `poll_users` (Pseudo-code 4) is designed to implement a series of inclusion checks on the users in the pool, and infer socio-demographics, attitudes and preferences based on their most recent activity on the platform. For the remainder of the research note I will discuss an application related to the 2024 US presidential election. Throughout I will present pseudo-code to summarise the functions necessary to conduct AI polling according to PoSSUM. The `alpha` version of this protocol is implemented in R, leveraging the `openai` package [26] to call the OpenAI api and prompt the `gpt-4o` model [23, 22].

A few words on the logic of this document. Section 2 presents the PoSSUM protocol. This includes a detailed explanation of `get_pool` and `poll_users`, the two principal algorithmic routines necessary to poll opinion on social media using LLMs. Section 3 presents an application of PoSSUM to the 2024 US Presidential election cycle. This is a “living document”, which I will update on every round of polling between August 15th and November 5th, election day, to show the evolution of the PoSSUM estimates over the campaign, and provide a comparison with more traditional polling. This

¹<https://github.com/robertocerinaprojects/PoSSUM>

²Silicon samples as per Argyle et al. [1] can be elicited from LLMs by using high-quality, real-life survey responses of humans as a *mould*.

document is meant to enhance the transparency and reproducibility of the framework, as well as to chronicle the 2024 US election experiment. My hope that by reviewing the logic behind this protocol, along with its results and alignment with other polling methods, AI polling will be recognized as a legitimate approach to studying public opinion, alongside methods like RDD and self-selected online panels.

2 Protocol

2.1 get_pool

Input:

- q : optimal API search queries
- w : weight of each query

Output:

- Υ : user-data object composed of profile info v and tweets \mathcal{T}

Routine *get_pool*:

```

 $K \leftarrow \text{length}(q)$  ;                               # Get: number of queries
 $\Upsilon \leftarrow \emptyset$  ;                             # Initialize: empty users object
for  $k = 1$  to  $K$  do
   $(\mathcal{T}, v)_{kt} \leftarrow \mathbb{X}(q_k, w_k)$  ;           # Call: sample of tweet-user pairs
   $\Upsilon \leftarrow \Upsilon \cup (\mathcal{T}, v)_{kt}$  ;           # Store: newly observed users
end
```

Algorithm 1: Pseudo-code for the *get_pool* routine.

I wish to sample a representative set of US adults amongst \mathbb{X} users. Given the timely nature of the inference I wish to make (what are people’s attitudes **today** ?) I further wish to look at activity on the platform on the same day. Given prohibitively expensive access to the *enterprise* or *pro* tiers of the \mathbb{X} api, I will assume users of PoSSUM have access to the *basic* tier, and hence have no access to a simple-random-sample of Tweets from a given day³. I therefore propose to use a combination of search queries for the `tweets/search/recent` endpoint⁴, and obtain a series of users who have engaged in language associated with the search terms on the platform, up to seven days prior to query-time. I implement a set of functions⁵ reminiscent of the now-defunct `rtweet` [15] and `academictwitterR` [2], in order to specify an appropriate set of queries.

³I use the *basic* tier \mathbb{X} api, meaning I pay \$ 100 for downloading 10k tweets per month. I can pay this price multiple times a month, and each payment allows another 10k tweets. The allowance is reset to the original 10k at the end of every month.

⁴See <https://developer.x.com/en/docs/twitter-api/tweets/search/introduction> for more details.

⁵Available in the GitHub repository, in the file `X.api.v2_function.R`

I need the content produced by the selected users to be informative of their political beliefs and attitudes. One way to ensure this is to use political search terms in the \mathbb{X} query. To perform US 2024 pre-election polling we could use a query such as that in Listing 1. Notice this is a joint query for all the candidates. This is preferable to independent queries per candidate, as these would yield estimates of support subject to selection on the dependent variable. The independent approach ignores the distribution of the search terms across tweets, and over-samples supporters of each candidate, distorting the distribution of support in favour of smaller parties⁶. I assign a *weight* (number of tweets extracted) to this query of size w .

⁶On the other hand, this sampling is very efficient per party – if you have access to selection-correction terms in the style of King & Zeng [16, 5], this approach would allow for the most sampling-efficient analysis.

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"

```

Individuals who talk about politics on \mathbb{X} are still unlike their counterparts in the general population. In particular, these are high political attention individuals, who are significantly more likely to vote than their population counterparts. To alleviate selection on political-attention I rely on a second set of queries, which are more likely to sample *normies*. I extract a random sample of L trending topics in the US (obtainable via <https://trends24.in/united-states/>), and produce a separate query for each topic. Each trending query is assigned a weight of $\frac{n}{L}$, such that the queries seeking to capture high-attention individuals and those capturing *normies* are assigned the same weight. Note this is arbitrary – I noticed this worked well in the US, but in general n and L are hyper-parameters that need tuning. We end up with a set of queries \mathbf{q} , which is an object of size $K = L + 1$, and a corresponding set of weights $\mathbf{w} = \{n, \frac{n}{L}, \dots, \frac{n}{L}\}$.

We execute each of these queries in a loop, and for each we obtain a tweet-user object $(\mathcal{T}, \mathbf{v})_{kt}$ containing at most w_k tweets and the user profiles responsible for generating them. We are uninterested in the tweets at this stage, and temporarily store the unique user-ids responsible for the tweets. The result of `get_pool` is a dated user-object Υ with contains the profile information about the user (e.g. self-reported description, location, profile picture, etc.), the date on which this user was collected, and the search-query used to capture them, and the set of tweets (typically of size 1) which the user is responsible for.

2.2 poll_users

For each real-life social media user in a given pool, we wish to infer responses to survey questions via harnessing LLMs’ intelligence. In what follows I describe PoSSUM’s LLM prompting protocol, and the full sequence of operations which define the `poll_users` routine. Pseudo-code 4 presents an overview of the `poll_users` routine.

Input:

- $\tau(t)$: temporal inclusion criteria
- Υ : users object database
- \mathcal{P}^E : entity inclusion prompt
- E : list of acceptable entities
- \mathcal{P}^G : geographic inclusion operation
- G : list of acceptable geographies
- $\mathcal{F} \leftarrow (\mathcal{F}^x, \mathcal{F}^y)$: list of independent and dependent features
- \mathcal{P}^ϕ : feature extraction prompt
- $(X^\mathcal{Q}, \omega^*, \omega' = 0)$: acceptable features, expected frequency, and sample counter
- m : number of tweets per user

Output:

- Z : survey object with extracted features

Routine *poll_users*:

```

 $\Upsilon^* \leftarrow \Upsilon[\tau = \text{TRUE} \vee \mathcal{G} \neq \emptyset]$  ;      # Filter: recent + valid location
 $N \leftarrow \text{length}(\Upsilon^*)$  ;                      # Get: number of valid users
 $Z \leftarrow \emptyset$  ;                          # Initialize: empty survey object

for  $i = 1$  to  $N$  do
     $e_i \leftarrow \text{GPT} \{ \mathcal{P}^E(\Upsilon_i) \}$  ;      # Call: GPT entity inclusion
    if  $e_i \in E$  then
         $g_i \leftarrow \text{GPT} \{ \mathcal{P}^G(\Upsilon_i) \}$  ;      # Call: GPT geographic inclusion
        if  $g_i \in G$  then
             $X_i \leftarrow \text{GPT} \{ \mathcal{P}^\phi(\Upsilon_i, \mathcal{F}^x) \}$  ;      # Call: GPT quota inclusion
            if  $X_i \in X^\mathcal{Q} \cup \omega'_i < \omega_i^*$  then
                 $\mathcal{T}_i^+ \leftarrow \mathbb{X}(\Upsilon_i)$  ;      # Call: sample last  $m$  tweets
                 $z_i \leftarrow \text{GPT} \{ \mathcal{P}^\phi(\Upsilon_i, \mathcal{T}_i^+, \mathcal{F}) \}$  ;      # Call: GPT extraction
                 $Z \leftarrow Z \cup z_i$  ;      # Store: survey object
                 $\omega'_i \leftarrow \omega'_i + 1$  ;      # Update: sample quota counter
            end
        end
    end
end

```

Algorithm 2: Pseudo-code for the *poll_users* routine.

2.2.1 Prompt Building Framework

In Pseudo-code 4 I use compressed notation for PoSSUM’s prompting strategy. \mathcal{P} can be described as a function that converts raw data into a *message object*, which is then passed on to `gpt-4o` for analysis. In the context of the OpenAI API, the message object is composed of *content sub-objects*. I use *image-type content* to provide the LLM with a public url for the users’ profile image, which it can then analyse; *text-type content* is used to pass all other text-based information, which I describe below. Figure 1 presents a stylised image of prompt building under PoSSUM.

Text-type content objects are assigned a standard structure composed of three elements: i. *background information* \mathcal{B} ; ii. a *mould* based on the available user data \mathcal{M} ; iii. a *feature extraction operation* \mathcal{O} . Pseudo algorithm 3 describes the building of the prompt. The *background* element includes any necessary background information which could help the LLM improve its decision making for a given task. For example, when prompting a LLM to deduce “2020 past vote” for a given user, we might wish to provide it with the election results by candidate for the state of residency of the user. The *mould* describes a given user in terms of the information which we have derived via the \mathbb{X} api – i.e. the users’ profile picture, name, username, self-reported description, and the full set of time-stamped tweets written by the user which are available to us at this stage. It is helpful to think of the mould as a function of the user data, $\mathcal{M}(\Upsilon)$, which returns a natural language description of the user, structured in some useful manner. The *feature extraction operation* element describes a feature extraction task. It is again useful to think of an operation as a function which takes a list of features as input, $\mathcal{O}(\mathcal{F})$, and returns a structured natural language task, whereby these features are embedded. For example, an entity recognition operation would be roughly described in terms of the following words: “*is this user a person ?*”; the feature of interest in this example would then be a dichotomous “personhood” variable. Pseudo-code 3 presents the steps involved in the prompt-building routine.

Input:

- \mathcal{B} : background information
- \mathcal{F} : features identified for extraction
- \mathcal{M} : mould
- \mathcal{O} : operation
- Υ : user data

Output:

- \mathcal{P} : unified prompt

Routine *build_prompt*:

| $\mathcal{P} = \mathcal{B} \parallel \mathcal{M}(\Upsilon) \parallel \mathcal{O}(\mathcal{F})$; # **Concatenate:** to form unified prompt

Algorithm 3: Pseudo-code for the *build_prompt* routine.

7

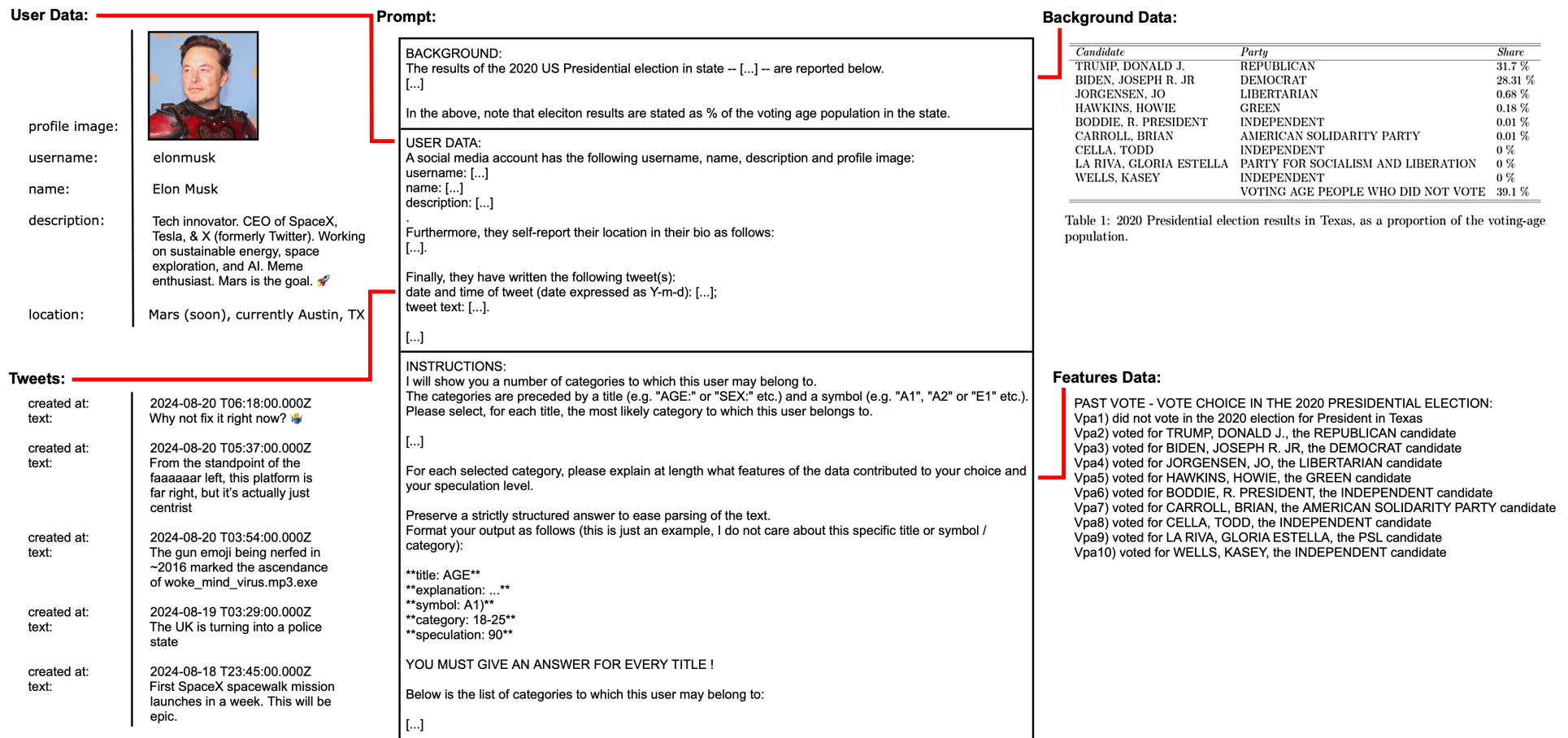


Figure 1: Toy example showing the composition of a prompt under the PoSSUM framework. The ‘[...]’ indicate areas where components are slotted in. The above toy example showcases a single feature (2020 vote choice), though multiple features can generally be extracted simultaneously. Not every prompt contains all of the elements indicated in this Figure. Instructions have been reduced to aid readability. Consult the GitHub repository for more details.

2.2.2 Inclusion Criteria

Much like in a traditional survey, several checks to ensure data quality and representation must be put in place.

Temporal Inclusion Criteria – τ : A rule limiting the number of survey-like responses that we wish to obtain from a single user within a given time-frame. For example, in the context of pre-election polling, we may wish to collect new data on a given user only once every 30 days – if the digital fieldwork is spread over a full month – despite their more frequent content creation. The routine to implement the temporal exclusion criteria involves: i. tallying the users that have been processed up to now; ii. identifying which of those have been processed within the last 30 days (or whatever the exclusion criteria); iii. removing those users from the *fresh* pool generated by the `get_pool` routine.

Input:

- τ : temporal inclusion criteria
- Υ : user data

Output:

- Υ^* : filtered user data including only recent users

Routine *temporal_inclusion*:

| $\Upsilon^* \leftarrow \Upsilon[\tau(t) = \text{TRUE}]$; # **Filter**: exclude users based on criteria

Algorithm 4: Pseudo-code for the `temporal_inclusion` routine.

Null Geography Exclusion Criteria – $\mathcal{G} \neq \emptyset$: This is a relatively simple data-quality check. Geography is a fundamental part of pre-election opinion polling – we must be able to place individuals within the given geographic boundary we wish to make inference for. If the user has no self-reported location, we exclude the user a-priori. Intelligent geographic filtering is in principle possible when an explicit location field is absent, by prompting the LLM to infer a location from other content generated by the user – this tends to be less accurate and more expensive due to the larger amounts of input-tokens necessary.

Input:

- Υ : user data
- \mathcal{G} : geographic information associated with users

Output:

- Υ^* : filtered user data with non-null geographic information

Routine *exclude_null_geography*:

| $\Upsilon^* \leftarrow \Upsilon[\mathcal{G} \neq \emptyset]$; **# Filter:** exclude users with null geography

Algorithm 5: Pseudo-code for the *exclude_null_geography* routine.

Entity Inclusion Criteria – E : Consists of defining the kinds of social-media profiles we want to include in our analysis. For pre-election opinion polling, we would wish to exclude \mathbb{X} accounts related to organisations (e.g. news outlets, NGOs), bots, and focus solely on real-life persons. This is done in PoSSUM via the feature extraction operation in Listing 2.

Listing 2: Example of an entity extraction operation.

1	Is this the account of a real-life existing Person, or of another kind of entity ?
2	Respond either with "P" for Person or "O" for Other.

Geographic Inclusion Criteria – \mathcal{G} : The geographic inclusion criteria helps us filter-out users who are unlikely to reside in the Level 1 geography. Level 1 geography constitutes the broadest boundary within which individuals belonging to the population of interest fall. For US pre-election polling we set this to the “United States of America”. Level 2 geography is then intended to be the “State”, and Level 3 is the relevant “Congressional District”. It is efficient to use a prompt which allows rejection of users who fail the Level 1 inclusion criteria, and simultaneously extracts the Level 2 information. Listing 3 presents an implementation of the geographic extraction prompt. We reject users who are “*Not from a state in the USA*”. The great advantage of using \mathbb{X} relative to other platforms is the relatively high rate of available self-reported locations, which makes geographically-bound polling possible.

Listing 3: Example of a geography extraction operation. Users that are not inferred to be living in the US by the LLM are excluded from the analysis.

```

1 Which state of the USA do they live in?
2 If they do not specify a state, but are still from the United States, write "USA".
3 If they are not from a state in the USA, write "Not from a state in the USA".
4 Write out just the full name of the state.
5 If they are from the District of Columbia, also known as Washington D.C., write "District of
  Columbia".

```

Quota Inclusion Criteria – \mathcal{Q} : The population of \mathbb{X} users is notoriously unrepresentative of the US population [14]. It is nevertheless a very large pool of US residents, accounting for around 22% of the US population. Whilst some categories – namely higher educated and higher income individuals – are extraordinarily over-represented, the pool is “deep enough” that we could expect to eventually find a number of representatives for most relevant socio-demographic groups in the population. It follows that implementing quotas is liable to make sampling more efficient.

PoSSUM implements quota sampling as follows: i. define a stratification frame (e.g. Table 1) which describes the number of individuals ω_c^* from each “cell” $c \in \{1, \dots, C\}$, which we would expect to capture in a random sample of target size Ω^* users – we could for instance set $\Omega^* = 1,500$ to produce polls of a somewhat traditional sample size; ii. a feature extraction operation is deployed to infer the values of the relevant variables for the user at hand. At this stage the LLM does not make use of any background information, and it utilises the same user-level information as the other intelligent inclusion criteria; iii. if the user belongs to a cell in the stratification frame for which the number of sampled users ω'_c is smaller than the number of desired users ω_c^* , I retain the user and update the quota counter – otherwise I exclude the user from the analysis. Pseudo-code 4 contains a symbolic description of the quota exclusion criteria implemented here.

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
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
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 $\Omega^* = 1,500$. This is a snapshot taken with 647 respondents still to be collected.

2.2.3 General Feature Extraction Framework

Implementing the quota inclusion criteria requires a defined set of features of interest \mathcal{F} . These features should be extracted from the user profiles and should have direct counterparts in the stratification frame, such that we are able to stop sampling from a given cell once it becomes fully populated.

It can be helpful at this stage to categorise features of interest in two distinct sets: those whose distribution in the population is known, and is directly available to us via auxiliary surveys, census data, election results, etc., we consider *independent* variables; those whose distribution in the population is unknown, we consider *dependent* variables. When ascertaining if a given feature belongs to either group, a simple rule of thumb is used: could I weight the poll by the marginal distribution of this variable? if so, this is an independent variable; if not, I consider it dependent. For the purpose of defining the quota inclusion criteria, I will discuss the independent features; the dependent variables will be discussed in Section 2.2.4.

Listing 4 presents a comprehensive list of independent features. Notice this is a more expansive list when compared to those in Table 1. This is deliberate, as we wish to retain the option of further weighting the sample after the data collection process has concluded, and we may wish to target a wider spectrum of marginal/joint distributions than those of the variables used to determine the quotas. A “features object” is assigned 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*. Categories in the features object can seem needlessly verbose – there are two reasons for this: i. detailed descriptions of the categories can help reduce the “neutrality bias” of the LLM – namely the tendency for the LLM to systematically prefer a more “neutral” or “wide-net” option under uncertainty; ii. associating each category with a unique text string helps ensure the unique parsing of the LLM output, especially when dealing with long prompts which include multiple questions sharing the same answer-text.

Having defined the set of features \mathcal{F} to extract, we need to define a feature extraction operation O to complete our prompt. Listing 5 presents PoSSUM’s standardised feature extraction operation. The chosen $O(\mathcal{F})$ endows the prompt with a set of standard instructions, and an example output which enforces a desirable structure to the LLM generation.

An important innovation here is the introduction of a *speculation score*, which allows the LLM to self-report the degree to which it is making an assessment based on data it has direct access to, such as the text of the tweets or the profile image, as opposed to via other means. I will describe its uses in Section 3.3.

The feature extraction operation considers all features jointly, 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 the order in which features are presented in the prompt. The auto-regressive nature of LLMs [18], 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.

Listing 4: Example of an “independent features” object. Categories are shortened for readability, see the [GitHub repository](#) for details.

```

1 ind.features <- c(
2   ETHNICITY:
3   E1) white - individuals with origins in any of the original peoples of europe...
4   E2) black or african american - individuals with origins in any of the black racial ...
5   E3) hispanic or latino - includes individuals of mexican, puerto rican, salvadoran, cuban...
6   E4) asian - individuals with origins in any of the original peoples of central or east ...
7   E5) american indian or alaskan native or native hawaiian or pacific islander - ...
8   E6) multiracial - individuals who identify explicitly as belonging to more than one of ...
9
10  ',
11  'AGE:
12  A1) under 18 years old
13  A2) 18 to 24 years old
14  A3) 25 to 34 years old
15  A4) 35 to 44 years old
16  A5) 45 to 54 years old
17  A6) 55 to 64 years old
18  A7) 65 or older
19
20  ',
21  'SEX:
22  S1) masculine sex - male
23  S2) feminine sex - female
24
25  ',
26  'INTEREST IN POLITICS:
27  I1) not interested at all in politics
28  I2) slightly interested in politics
29  I3) moderately interested in politics
30  I4) highly interested in politics
31
32  ',
33  'MARITAL STATUS:
34  M1) married - currently legally married and living with a spouse
35  M2) single - never married, including those who are legally separated
36  M3) divorced - legally divorced and not currently remarried
37  M4) widowed - spouse has passed away and not currently remarried
38
39  ',
40  "HIGHEST EDUCATIONAL QUALIFICATION:
41  Q1) completed education up to and including high school - high school diploma, ...
42  Q2) completed education at the college or university level - bachelor's degree, ...
43
44  ",
45  'HOUSEHOLD INCOME BRACKET:
46  H1) up to 25000 USD per year
47  H2) between 25000 and 50000 USD per year
48  H3) between 50000 and 75000 USD per year
49  H4) between 75000 and 100000 USD per year
50  H5) more than 100000 USD per year
51
52  ',
53  'GENERAL TRUST IN OTHER PEOPLE:
54  Tru1) always trust other people
55  Tru2) most of the time trust other people
56  Tru3) about half of the time trust other people
57  Tru4) some of the time trust other people
58  Tru5) never trust other people
59
60  ',
61  'PAYING ATTENTION TO THE 2024 PRESIDENTIAL ELECTION:
62  Att1) not paying attention at all to the 2024 Presidential election in the US

```

```

63 Att2) paying only a little attention to the 2024 Presidential election in the US
64 Att3) paying some attention to the 2024 Presidential election in the US
65 Att4) paying a lot of attention to the 2024 Presidential election in the US
66
67 ',
68 'PARTISAN LOYALTIES:
69 Pid1) strongly identifies with Democrats
70 Pid2) weakly identifies with Democrats
71 Pid3) independent closer to the Democrats
72 Pid4) independent not closer to either party
73 Pid5) independent closer to the Republicans
74 Pid6) weakly identifies with Republicans
75 Pid7) strongly identifies with Republicans
76
77 ',
78 "IDEOLOGICALLY, THIS PERSON APPEARS TO BE:
79 Ide2) very ideologically liberal
80 Ide3) somewhat ideologically liberal
81 Ide4) moderate in ideological orientation
82 Ide5) somewhat ideologically conservative
83 Ide6) very ideologically conservative
84
85 ",
86 'PAST VOTE - TURNOUT IN THE 2020 PRESIDENTIAL ELECTION:
87 Tpa1) no chance this individual turned out to vote - Probability: 0 ...
88 Tpa2) highly unlikely this individual turned out to vote - Probability: 0.15 ...
89 Tpa3) unlikely this individual turned out to vote - Probability: 0.3 ...
90 Tpa4) 50-50 likelihood that this individual voted - Probability: 0.5 ...
91 Tpa5) likely this individual turned out to vote - Probability: 0.7 ...
92 Tpa6) highly likely this individual turned out to vote - Probability: 0.85 ...
93 Tpa7) certain this individual turned out to vote - Probability: 1 ...
94
95 ',
96 'PAST VOTE - TURNOUT IN THE 2022 HOUSE OF REPRESENTATIVES ELECTION:
97 Thpa1) no chance this individual turned out to vote - Probability: 0 ...
98 Thpa2) highly unlikely this individual turned out to vote - Probability: 0.15 ...
99 Thpa3) unlikely this individual turned out to vote - Probability: 0.3 ...
100 Thpa4) 50-50 likelihood that this individual voted - Probability: 0.5 ...
101 Thpa5) likely this individual turned out to vote - Probability: 0.7 ...
102 Thpa6) highly likely this individual turned out to vote - Probability: 0.85 ...
103 Thpa7) certain this individual turned out to vote - Probability: 1 ...
104
105 ',
106 'PAST VOTE - VOTE CHOICE IN THE 2020 PRESIDENTIAL ELECTION:
107 Vpa1) did not vote in the 2020 election for President in their state
108 Vpa2) voted for Donald Trump, the Republican Party candidate, ...
109 Vpa3) voted for Joe Biden, the Democratic Party candidate, ...
110 Vpa4) voted for Jo Jorgensen, the Libertarian Party candidate, ...
111 Vpa5) voted for Howie Hawkins, the Green Party candidate, ...
112 Vpa6) voted for a candidate other than the Republican, Democrat, Libertarian, or Green ...
113
114 ',
115 'PAST VOTE - VOTE CHOICE IN THE 2022 HOUSE OF REPRESENTATIVES ELECTION:
116 Vhpa1) did not vote in the 2022 elections for the House of Representatives in their ...
117 Vhpa2) voted for a Republican Party candidate, ...
118 Vhpa3) voted for a Democratic Party candidate, ...
119 Vhpa4) voted for a Libertarian Party candidate, ...
120 Vhpa5) voted for a Green Party candidate, ...
121 Vhpa6) voted for a candidate other than the Republican, Democrat, Libertarian, or Green ...
122
123 ')
124 }

```


Listing 5: Standardised feature extraction operation. The text is followed by a list of features to be extracted, such as those in Listing 4.

```

1 tsk <-
2   paste0(
3     'I will show you a number of categories to which this user may belong to.
4     The categories are preceded by a title (e.g. "AGE:" or "SEX:" etc.) and a symbol (e.g. "A1",
5     "A2" or "E1" etc.).
6     Please select, for each title, the most likely category to which this user belongs to.
7     In your answer present, for each title, the selected symbol.
8     Write out in full the category associated with the selected symbol.
9     The chosen symbol / category must be the most likely to accurately represent this user.
10    You must only select one symbol / category per title.
11    A title, symbol and category cannot appear more than once in your answer.
12
13    For each selected symbol / category, please note the level of Speculation involved in this
14    selection.
15    Present the Speculation level for each selection on a scale from 0 (not speculative at all,
16    every single element of the user data was useful in the selection) to 100 (fully
17    speculative, there is no information related to this title in the user data).
18    Speculation levels should be a direct measure of the amount of useful information available
19    in the user data.
20    Speculation levels pertain only to the information available in the user data -- namely the
21    username, name, description, location, profile picture and tweets from this user -- and
22    should not be affected by additional information available to you from any other source.
23    To ensure consistency, use the following guidelines to determine speculation levels:
24
25    0-20 (Low speculation): The user data provides clear and direct information relevant to the
26    title. (e.g., explicit mention in the profile or tweets)
27    21-40 (Moderate-low speculation): The user data provides indirect but strong indicators
28    relevant to the title. (e.g., context from multiple sources within the profile or tweets
29    )
30    41-60 (Moderate speculation): The user data provides some hints or partial information
31    relevant to the title. (e.g., inferred from user interests or indirect references)
32    61-80 (Moderate-high speculation): The user data provides limited and weak indicators
33    relevant to the title. (e.g., very subtle hints or minimal context)
34    81-100 (High speculation): The user data provides no or almost no information relevant to
35    the title. (e.g., assumptions based on very general information)
36
37    For each selected category, please explain at length what features of the data contributed
38    to your choice and your speculation level.
39
40    Preserve a strictly structured answer to ease parsing of the text.
41    Format your output as follows (this is just an example, I do not care about this specific
42    title or symbol / category):
43
44    **title: AGE**
45    **explanation: ...**
46    **symbol: A1**
47    **category: 18-25**
48    **speculation: 90**
49
50    YOU MUST GIVE AN ANSWER FOR EVERY TITLE !
51
52    Below is the list of categories to which this user may belong to:
53
54    ',
55      # randomise feature order
56      paste0(
57        sample(ind.features),
58        collapse = '\n'
59      )
60    )

```

2.2.4 Extracting Response Features

Surviving user profiles can be reasonably assessed as sufficiently information-rich, representing a real-life person in the Level 1 geography. Their latest digital trace is recent, at most 1 week old from the moment the `get_pool` routine is initiated. It is then worthwhile to expend resources to “survey” these profiles. We do so by prompting the LLM under the general feature extraction framework described in Section 2.2.3, with two important differences: i. we expand the digital trace available for each user by querying their timeline for their last m tweets, and generate an expanded user mould; ii. we impute a complete set of *independent* and *dependent* characteristics. The full list of dependent characteristics is available in Listing 6.

With respect to the expansion of the profile to include a greater number of recent tweets per user, it is worth distinguishing between users captured via *trending topics*, as opposed to *political talk*, queries. A very small number of tweets is necessary to estimate the preferences of those who talk politics on \mathbb{X} . Conversely, users discussing trending topics on \mathbb{X} can be totally enigmatic with respect to their politics – their last m tweets could never mention anything remotely useful to indicate political preferences. As a results we set two distinct values of m for these two sets of subjects: $m^{trending} = \lambda \times m^{politics}$, $\forall \lambda > 1$. I use $\lambda = 2$ and $m = 20$, but this is open to further tuning. What is generally true is that, where resources permit, “more is better” in terms of information used to generate a user’s mould.

Listing 6: Example of a “dependent features” object. Categories are shortened for readability, see the [GitHub repository](#) for details.

```

1 dep.features <- c(
2   "UNDECIDEDNESS AROUND 2024 US PRESIDENTIAL ELECTION VOTE CHOICE - ...
3   Und1) no chance this individual will change their mind - Probability: 0 - ...
4   Und2) highly unlikely this individual will change their mind - Probability: 0.15 - ...
5   Und3) unlikely this individual will change their mind - Probability: 0.3 - ...
6   Und4) 50-50 likelihood that this individual will change their mind - Probability: 0.5 - ...
7   Und5) likely this individual will change their mind - Probability: 0.7 - ...
8   Und6) highly likely this individual will change their mind - Probability: 0.85 - ...
9   Und7) certain this individual will change their mind - Probability: 1 - ...
10
11  ",
12  "JOE BIDEN PRESIDENTIAL JOB APPROVAL:
13  Bap1) Strongly Approves of the way Joe Biden is handling his job as president
14  Bap2) Somewhat Approves of the way Joe Biden is handling his job as president
15  Bap3) Somewhat Disapproves of the way Joe Biden is handling his job as president
16  Bap4) Strongly Disapproves of the way Joe Biden is handling his job as president
17
18  ",
19  "FAVOURABILITY OF US POLITICIAN JOE BIDEN:
20  Bfa1) Very favourable view of Joe Biden
21  Bfa2) Somewhat favourable view of Joe Biden
22  Bfa3) Somewhat unfavourable view of Joe Biden
23  Bfa4) Very unfavourable view of Joe Biden
24  Bfa5) This person is unlikely to know who Joe Biden is, so they cannot have an opinion ...
25
26  ",
27  "FAVOURABILITY OF US POLITICIAN KAMALA HARRIS:
28  Kfa1) Very favourable view of Kamala Harris
29  Kfa2) Somewhat favourable view of Kamala Harris
30  Kfa3) Somewhat unfavourable view of Kamala Harris
31  Kfa4) Very unfavourable view of Kamala Harris
32  Kfa5) This person is unlikely to know who Kamala Harris is, so they cannot have an ...
33
34  ",
35  "FAVOURABILITY OF US POLITICIAN TIM WALZ:
36  Twa1) Very favourable view of Tim Walz
37  Twa2) Somewhat favourable view of Tim Walz
38  Twa3) Somewhat unfavourable view of Tim Walz
39  Twa4) Very unfavourable view of Tim Walz
40  Twa5) This person is unlikely to know who Tim Walz is, so they cannot have an opinion ...
41
42  ",
43  "FAVOURABILITY OF US POLITICIAN DONALD TRUMP:
44  Tfa1) Very favourable view of Donald Trump
45  Tfa2) Somewhat favourable view of Donald Trump
46  Tfa3) Somewhat unfavourable view of Donald Trump
47  Tfa4) Very unfavourable view of Donald Trump
48  Tfa5) This person is unlikely to know who Donald Trump is, so they cannot have an ...
49
50  ",
51  "FAVOURABILITY OF US POLITICIAN JD VANCE:
52  Jfa1) Very favourable view of JD Vance
53  Jfa2) Somewhat favourable view of JD Vance
54  Jfa3) Somewhat unfavourable view of JD Vance
55  Jfa4) Very unfavourable view of JD Vance
56  Jfa5) This person is unlikely to know who JD Vance is, so they cannot have an opinion ...
57
58  ",
59  "FAVOURABILITY OF US POLITICIAN ROBERT F. KENNEDY JR.:
60  Rfa1) Very favourable view of Robert F. Kennedy Jr.
61  Rfa2) Somewhat favourable view of Robert F. Kennedy Jr.
62  Rfa3) Somewhat unfavourable view of Robert F. Kennedy Jr.

```

63 Rfa4) Very unfavourable view of Robert F. Kennedy Jr.
64 Rfa5) This person is unlikely to know who Robert F. Kennedy Jr. is, so they cannot have ...
65
66 ',
67 'FAVOURABILITY OF US POLITICIAN CORNEL WEST:
68 Wfa1) Very favourable view of Cornel West
69 Wfa2) Somewhat favourable view of Cornel West
70 Wfa3) Somewhat unfavourable view of Cornel West
71 Wfa4) Very unfavourable view of Cornel West
72 Wfa5) This person is unlikely to know who Cornel West is, so they cannot have an opinion ...
73
74 ',
75 'FAVOURABILITY OF US POLITICIAN JILL STEIN:
76 Sfa1) Very favourable view of Jill Stein
77 Sfa2) Somewhat favourable view of Jill Stein
78 Sfa3) Somewhat unfavourable view of Jill Stein
79 Sfa4) Very unfavourable view of Jill Stein
80 Sfa5) This person is unlikely to know who Jill Stein is, so they cannot have an opinion ...
81
82 ',
83 'FAVOURABILITY OF US POLITICIAN CHASE OLIVER:
84 Ofa1) Very favourable view of Chase Oliver
85 Ofa2) Somewhat favourable view of Chase Oliver
86 Ofa3) Somewhat unfavourable view of Chase Oliver
87 Ofa4) Very unfavourable view of Chase Oliver
88 Ofa5) This person is unlikely to know who Chase Oliver is, so they cannot have an ...
89
90 ',
91 'BELIEF IN MOST IMPORTANT ISSUE TODAY:
92 Moi01) jobs and the economy are the most important issue for this user today
93 Moi02) immigration is the most important issue for this user today
94 Moi03) climate change and the environment are the most important issue for this user today
95 Moi04) foreign policy is the most important issue for this user today
96 Moi05) national security is the most important issue for this user today
97 Moi06) education is the most important issue for this user today
98 Moi07) healthcare is the most important issue for this user today
99 Moi08) taxes and government spending are the most important issue for this user today
100 Moi09) abortion is the most important issue for this user today
101 Moi10) civil rights and racism are the most important issue for this user today
102 Moi11) guns are the most important issue for this user today
103 Moi12) crime is the most important issue for this user today
104 Moi13) criminal justice reform and over-incarceration are the most important issue for ...
105 Moi14) inflation and the cost of living is the most important issue for this user today
106
107 ',
108 'HAPPINESS LEVEL:
109 Hap1) very happy - consistently feels a high level of joy and satisfaction with ...
110 Hap2) rather happy - generally feels happy and satisfied with their life - ...
111 Hap3) not very happy - experiences some level of dissatisfaction or lack of joy in their ...
112 Hap4) not happy at all - consistently feels unhappy and dissatisfied with their life ...
113
114 ',
115
116 'BIG 5 PERSONALITY TRAIT -- OPENNESS TO EXPERIENCE:
117 Ope1) very open to experience - highly curious, imaginative, and open to exploring ...
118 Ope2) somewhat open to experience - open to new experiences and ideas, but they may ...
119 Ope3) neither open nor closed to experience - shows a balanced approach, being neither ...
120 Ope4) somewhat closed to experience - tends to prefer familiar routines and may be ...
121 Ope5) very closed to experience - prefers a highly structured and predictable ...
122
123 ',
124 'BIG 5 PERSONALITY TRAIT -- CONSCIENTIOUSNESS:
125 Con1) very conscientious - highly organized, responsible, and dependable - they ...
126 Con2) somewhat conscientious - generally reliable and organized but may occasionally ...
127 Con3) neither conscientious nor unconscientious - demonstrates an average level of ...

```

128 Con4) somewhat unconscientious - tends to be less organized and may struggle with ...
129 Con5) very unconscientious - often lacks organization and discipline - they may be ...
130
131 ',
132 'BIG 5 PERSONALITY TRAIT -- EXTRAVERSION:
133 Ext1) very extraverted - highly outgoing, energetic, and enjoys social interactions - ...
134 Ext2) somewhat extraverted - enjoys social interactions and can be outgoing, but ...
135 Ext3) neither extraverted nor introverted - displays a balanced mix of extraverted ...
136 Ext4) somewhat introverted - tends to be more reserved and prefers quieter, less ...
137 Ext5) very introverted - highly reserved, enjoys solitude, and prefers minimal social ...
138
139 ',
140 "BIG 5 PERSONALITY TRAIT -- AGREEABLENESS:
141 Agr1) very agreeable - highly cooperative, compassionate, and trusting - they ...
142 Agr2) somewhat agreeable - generally kind and cooperative but may occasionally ...
143 Agr3) neither agreeable nor disagreeable - generally kind and cooperative but ...
144 Agr4) somewhat disagreeable - tends to be more competitive and less concerned with ...
145 Agr5) very disagreeable - often uncooperative and less empathetic - they prioritize ...
146
147 ",
148 'BIG 5 PERSONALITY TRAIT -- NEUROTICISM:
149 Neu1) very high in neuroticism - frequently experiences intense emotions such ...
150 Neu2) somewhat high in neuroticism - tends to experience negative emotions more ...
151 Neu3) neither high nor low in neuroticism - has an average level of emotional stability ...
152 Neu4) somewhat low in neuroticism - generally calm and emotionally stable - they ...
153 Neu5) very low in neuroticism - exceptionally calm, resilient, and emotionally stable - ...
154
155 ',
156 'CURRENT VOTING PREFERENCES - TURNOUT IN THE 2024 PRESIDENTIAL ELECTION IF THE ELECTION
    WERE HELD ON THE DATE OF THEIR MOST RECENT TWEET:
157 Tcu1) no chance this individual would turn-out to vote - Probability: 0 - in the ...
158 Tcu2) highly unlikely this individual would turn-out to vote - Probability: 0.15 - ...
159 Tcu3) unlikely this individual would turn-out to vote - Probability: 0.3 - in the ...
160 Tcu4) 50-50 likelihood that this individual would vote - Probability: 0.5 - in the ...
161 Tcu5) likely this individual would turn-out to vote - Probability: 0.7 - in the ...
162 Tcu6) highly likely this individual would turn-out to vote - Probability: 0.85 - in the ...
163 Tcu7) certain this individual would turn-out to vote - Probability: 1 - in the ...
164
165 ',
166 'CURRENT VOTING PREFERENCES - VOTE CHOICE IN THE 2024 PRESIDENTIAL ELECTION IF THE
    ELECTION WERE HELD ON THE DATE OF THEIR MOST RECENT TWEET:
167 Vcu1) would not vote in the 2024 elections for President in their state
168 Vcu2) would vote for Donald Trump, the Republican Party candidate ...
169 Vcu3) would vote for Kamala Harris, the Democratic Party candidate ...
170 Vcu4) would vote for Robert F. Kennedy Jr., who is not affiliated with any ...
171 Vcu5) would vote for Jill Stein, the Green Party candidate ...
172 Vcu6) would vote for Chase Oliver, the Libertarian Party candidate ...
173 Vcu7) would vote for Dr. Cornel West, who is not affiliated with any political party, ...
174
175 )

```

Listing 7: Example of LLM output in response to an independent-features extraction propt.

```

1 **title: HIGHEST EDUCATIONAL QUALIFICATION**
2 **explanation: The user mentions being a former instructor at University, which strongly
   implies a high level of education.**
3 **symbol: Q2**
4 **category: completed education at the college or university level - bachelor's degree,
   master's degree, doctorate**
5 **speculation: 15**
6
7 **title: ETHNICITY**
8 **explanation: The user includes the Italian flag in their bio, indicating a likely Italian
   heritage, which falls under the white category.**
9 **symbol: E1**
10 **category: white**
11 **speculation: 5**
12
13 **title: HOUSEHOLD INCOME BRACKET**
14 **explanation: The user's professional background as a journalist, essayist, poet, and
   former instructor at prestigious universities suggests a higher income bracket, although
   this is moderately speculative.**
15 **symbol: H5**
16 **category: more than 100000 USD per year**
17 **speculation: 50**
18
19 **title: INTEREST IN POLITICS**
20 **explanation: The user identifies as a Democrat and engages in political commentary on
   Twitter, indicating a high interest in politics.**
21 **symbol: I4**
22 **category: highly interested in politics**
23 **speculation: 10**
24
25 **title: SEX**
26 **explanation: The user's name "Karen" and profile picture indicate a feminine sex.**
27 **symbol: S2**
28 **category: feminine sex - female**
29 **speculation: 5**
30
31 **title: IDEOLOGICALLY, THIS PERSON APPEARS TO BE**
32 **explanation: The user identifies as a Democrat and expresses strong opinions about
   political figures, suggesting a strong ideological leaning.**
33 **symbol: Ide2**
34 **category: very ideologically liberal**
35 **speculation: 20**
36
37 **title: MARITAL STATUS**
38 **explanation: There is no direct information about marital status in the user data, making
   this highly speculative.**
39 **symbol: M2**
40 **category: single**
41 **speculation: 85**
42
43 **title: PAST VOTE - TURNOUT IN THE 2020 PRESIDENTIAL ELECTION**
44 **explanation: The user's strong political engagement and identification as a Democrat make
   it highly likely they turned out to vote.**
45 **symbol: Tpa7**
46 **category: certain this individual turned out to vote**
47 **speculation: 20**
48
49 **title: PAST VOTE - VOTE CHOICE IN THE 2020 PRESIDENTIAL ELECTION**
50 **explanation: The user identifies as a Democrat, making it likely they voted for Joe Biden.
   **
51 **symbol: Vpa3**
52 **category: voted for Joe Biden, the Democratic Party candidate, in the 2020 election for

```

```

53     President in their state**
54 **speculation: 10**
55 **title: GENERAL TRUST IN OTHER PEOPLE**
56 **explanation: There is no information directly indicating this user's level of trust in
57     other people, making this highly speculative.**
58 **symbol: Tru3**
59 **category: about half of the time trust other people**
60 **speculation: 80**
61 **title: PAYING ATTENTION TO THE 2024 PRESIDENTIAL ELECTION**
62 **explanation: The user's active political engagement and commentary indicate they are
63     paying a lot of attention to the upcoming election.**
64 **symbol: Att4**
65 **category: paying a lot of attention to the 2024 Presidential election in the US**
66 **speculation: 20**
67 **title: PAST VOTE - VOTE CHOICE IN THE 2022 HOUSE OF REPRESENTATIVES ELECTION**
68 **explanation: As a self-identified Democrat, it is likely they voted for a Democratic Party
69     candidate in the 2022 elections.**
70 **symbol: Vhpa3**
71 **category: voted for a Democratic Party candidate in the 2022 elections for the House of
72     Representatives in their congressional district**
73 **speculation: 30**
74 **title: AGE**
75 **explanation: Based on the professional experience detailed in the bio, including 20 years
76     writing about jazz and previous hard news experience, the user is likely in an older age
77     bracket.**
78 **symbol: A6**
79 **category: 55 to 64 years old**
80 **speculation: 30**
81 **title: PARTISAN LOYALTIES**
82 **explanation: The user identifies as a Democrat, suggesting strong partisan loyalty.**
83 **symbol: Pid1**
84 **category: strongly identifies with Democrats**
85 **speculation: 10**
86 **title: PAST VOTE - TURNOUT IN THE 2022 HOUSE OF REPRESENTATIVES ELECTION**
87 **explanation: Given the user's high political engagement, it is highly likely they turned
88     out to vote in the 2022 elections.**
89 **symbol: Thpa7**
90 **category: certain this individual turned out to vote**
91 **speculation: 20**

```

2.2.5 Background-informed Features

Level 2 (2020 and 2024 Presidential election vote at the state-level) and Level 3 (2022 House of Representatives election vote at the district-level) voting behaviour prompts could be improved by providing the LLM with a realistic list of candidate choices which were / are available to the user. For instance, the Presidential ticket headlined by Dr. Cornel West does not have ballot access for the 2024 election in the state of California, according to Wikipedia (https://en.wikipedia.org/wiki/Ballot_access_in_the_2024_United_States_presidential_election). It stands to reason that removing Dr. West from the choices available to the user would improve the chances of generating a realistic estimate of vote choice.

A background-informed feature extraction prompt, tailored to extract background-informed features, can be operationalised by a set of sequential sub-prompts. The first prompt is used to generate the the relevant features-object, conditional on the background information. In the example above this would be a set of candidate choices, with associated unique symbols, which is updated to reflect the latest ballot access data. The second prompt is used to extract the background-informed features for the user at hand. This would be the specific choice preferred by the user, out of those available in the background-informed prompt generated by the LLM. Pseudo-algorithm 6 describes this prompting strategy; Listing 8 presents an example of the first-step prompt \mathcal{P}_1 .

Input:

- \mathcal{B} : background information
- Υ : user data
- \mathcal{O}_1 : operation to generate background-informed features
- \mathcal{F}_1 : empty features object
- \mathcal{O}_2 : operation to extract features
- \mathcal{M} : mould structure

Output:

- \mathcal{F}_2 : background-informed features generated by GPT
- \mathbf{X}_i : extracted features for the user data

Routine *background_informed_feature_extraction*:

$\mathcal{P}_1 = \mathcal{B} \parallel \mathcal{O}_1(\mathcal{F}_1) ;$	# Concatenate: to form 1 st prompt
$\mathcal{F}_2 \leftarrow \text{GPT}(\mathcal{P}_1) ;$	# Call: GPT to generate bg-informed features
$\mathcal{P}_2 = \mathcal{B} \parallel \mathcal{M}(\Upsilon) \parallel \mathcal{O}_2(\mathcal{F}_2) ;$	# Concatenate: to form 2 nd prompt
$\mathbf{X}_i \leftarrow \text{GPT}(\mathcal{P}_2) ;$	# Call: GPT to extract features

Algorithm 6: Pseudo-code for the background-informed feature extraction routine.

Note that this strategy can be extended further to multiple rounds of sequential prompting. Take the example of Level 3 background informed vote choice extraction. Here we have three layers of prompting which feed into each other.

A first layer pertains the allocation of the \mathbb{X} user into a plausible Level 3 geography (Congressional District), conditional on the previously-imputed Level 2 geography (State). To perform this operation, the LLM needs access to information about the latest boundaries for each Congressional District is necessary – including the distribution of socio-demographic groups, the relevant economic and cultural centres, its political history in the most recent elections, etc. This information can be compiled separately and passed to the LLM as background. It might be tempting to assume the LLM’s inherent knowledge of district boundaries and characteristics is up to date – somewhat surprisingly this was not the case for `gpt-4o`. Despite the model being trained with information up to and including October 2023, its reflexive recall of the boundaries of congressional districts appears to be prior to 2022 redistricting (as of July 2024).

Upon obtaining a background-informed Level 3 geography imputation, we wish to know who the user voted for amongst the candidates available to them in 2022. To do this, a second layer of prompting is implemented, now similar to that shown in Listing 8, where data from the MIT Election Lab⁷ is used to describe the number of votes obtain by each candidate in the 2022 congressional, in the imputed District. Finally a feature extraction prompt is used to extract the 2022 past vote of the given user, given the choice-set available to them.

⁷<https://electionlab.mit.edu>

Listing 8: Example of a full background-informed, feature-building prompt for the 2024 US Presidential election. The prompt is completed by the LLM, and it is subsequently passed on as a “features object” for a feature-extraction operation.

```

1 BACKGROUND INFORMATION:
2 The candidates running in the 2024 US Presidential election in state -- Massachusetts -- are
   reported below.
3
4 Kamala Harris, the Democratic Party candidate
5 Robert F. Kennedy Jr., who is not affiliated with any political party
6 Donald Trump, the Republican Party candidate'
7
8 INSTRUCTIONS:
9 Based on what you know of the candidates in the 2024 Presidential election held in this
   state on November 5, 2024, please complete the following set of questions and their
   options.
10 If there are no candidates for the given party, remove the option related to the given party
   entirely -- do not present that party's option at all.
11 If there is more than one candidate for a single party, write out each option in two
   separate lines, and assign a different symbol for the identifier to each.
12 Below is the set of questions and options for you to complete - your job is to replace the
   instances wrapped in <...> with the correct knowledge for this state.
13 Do not produce any other text beyond the completed set of questions.
14
15 CURRENT VOTING PREFERENCES - TURNOUT IN THE 2024 PRESIDENTIAL ELECTION IF THE ELECTION WERE
   HELD ON THE DATE OF THEIR MOST RECENT TWEET:
16 Tcu1) no chance this individual will turn-out to vote - Probability: 0 - in the 2024
   election for President in <INSERT_STATE_NAME_HERE>
17 Tcu2) highly unlikely this individual will turn-out to vote - Probability: 0.15 - in the
   2024 election for President in <INSERT_STATE_NAME_HERE>
18 Tcu3) unlikely this individual will turn-out to vote - Probability: 0.3 - in the 2024
   election for President in <INSERT_STATE_NAME_HERE>
19 Tcu4) 50-50 likelihood that this individual will vote - Probability: 0.5 - in the 2024
   election for President in <INSERT_STATE_NAME_HERE>
20 Tcu5) likely this individual will turn-out to vote - Probability: 0.7 - in the 2024 election
   for President in <INSERT_STATE_NAME_HERE>
21 Tcu6) highly likely this individual will turn-out to vote - Probability: 0.85 - in the 2024
   election for President in <INSERT_STATE_NAME_HERE>
22 Tcu7) certain this individual will turn-out to vote - Probability: 1 - in the 2024 election
   for President in <INSERT_STATE_NAME_HERE>
23
24 CURRENT VOTING PREFERENCES - VOTE CHOICE IN THE 2024 PRESIDENTIAL ELECTION IF THE ELECTION
   WERE HELD ON THE DATE OF THEIR MOST RECENT TWEET:
25 Vcu1) would not vote in the 2024 election for President in <INSERT_STATE_NAME_HERE>
26 Vcu<INSERT_OPTION_NUMBER_HERE>) would vote for <INSERT_REPUBLICAN_CANDIDATE_NAME_HERE>, the
   Republican Party candidate, in the 2024 election for President in <INSERT_STATE_NAME_
   HERE>
27 Vcu<INSERT_OPTION_NUMBER_HERE>) would vote for <INSERT_DEMOCRATIC_CANDIDATE_NAME_HERE>, the
   Democratic Party candidate, in the 2024 election for President in <INSERT_STATE_NAME_
   HERE>
28 Vcu<INSERT_OPTION_NUMBER_HERE>) would vote for <INSERT_LIBERTARIAN_CANDIDATE_NAME_HERE>, the
   Libertarian Party candidate, in the 2024 election for President in <INSERT_STATE_NAME_
   HERE>
29 Vcu<INSERT_OPTION_NUMBER_HERE>) would vote for <INSERT_GREEN_CANDIDATE_NAME_HERE>, the Green
   Party candidate, in the 2024 election for President in <INSERT_STATE_NAME_HERE>
30 Vcu<INSERT_OPTION_NUMBER_HERE>) would vote for <INSERT_INDEPENDENT_CANDIDATE_NAME_HERE>, a
   candidate who is not affiliated with any political party, in the 2024 election for
   President in <INSERT_STATE_NAME_HERE>"

```

3 2024 US Pre-Election Polling

I field a PoSSUM poll starting on August 15th and ending on August 23rd. The quotas were tuned to target a representative sample of 1,500 individuals, though the protocol was stopped when the expected number of new acceptable users in a “fresh” pool dropped to cost-ineffective levels (see Table 2). During this time in the field I obtain valid digital traces for 945 unique users. The anonymised individual-level data is freely available on <https://github.com/robertocerinaprojects/PoSSUM>. The costs of the exercise can be broken down as follows: \$145 were spent on \mathbb{X} API calls to generate a subject pool via the `get_pool` routine; \$287 (upper-bound) were spent on querying user-specific timelines to get their latest tweets; \$471 (upper bound) were allocated to LLM (OpenAI API) processing costs. The total cost of the poll was \$885; the cost-per-accepted-user was \$0.94⁸.

3.1 Subject Pool

The `get_pool` routine is executed on four occasions, in order to replenish the subject pool after a given set of subjects is exhausted, either because they do not meet an inclusion criteria, or because they have already been analysed. Table 2 presents the total number of candidate users by pool-date, broken down by the trending and political search-terms. The same table also presents the number of users amongst those added to the pool who are accepted for the final survey-object. Decreasing returns are evident: holding the subject pool size roughly constant, the first pool adds 489 new subjects to the poll; the last only adds 80, making subsequent “recruitment” relatively unaffordable. We can further explore the reasons for user exclusion for each new subject-pool. Table 3 presents this breakdown. Complementing the information in Table 2, the overall acceptance rate drops from 13.3% in the subject pool obtained on 15/08, to a mere 2.3% on 22/08, again pointing to decreasing returns to further sampling. The main reason for rejecting users at any stage of sampling is the absence of a self-reported location, accounting for the rejection of around 40% of the users in our subject pool. Just over 10% of users are rejected for failing the entity inclusion criteria (`not_a_person`), and another 10% fail the geographic inclusion criteria (`not_from_usa`), whilst 5% of users are typically inferred to have characteristics too rare to be included in the sample (`no_quota_match`). The prevalence of these rejection reasons tends to be stable over the fieldwork period. On the other hand, the `quota_is_full` reason takes up a larger share of rejections in subject pools collected towards the end of digital fieldwork: on the 15/08 subject pool, 14.3% of users are rejected due to filled quotas, but this proportion increases to 32% on 21/08, and 26.9% on 22/08 respectively.

⁸It is worth noting that these costs are liable to drop substantially with further optimisation. One way to cut costs is to gain access to the Pro-tier of the \mathbb{X} API, which cut in half the cost-per-tweet – so the cost-per-accepted-user falls to \$0.72 under this tier. Further note that the LLM API processing costs are coming down exponentially (see: <https://llmpricecheck.com>).

	2024-08-15	2024-08-19	2024-08-21	2024-08-22
“Fresh” Users in the Subject Pool				
Total	3,670	3,721	3,615	3,520
Politics	1,960	1,944	1,955	1,867
Trends	1,710	1,777	1,660	1,653
Accepted Users				
Total	489	258	129	80
Politics	190	95	56	33
Trends	299	163	73	47

Table 2: Breakdown of users by subject-pool type and inclusion in the survey-object, for each instance of the `get_pool` routine.

	2024-08-15	2024-08-19	2024-08-21	2024-08-22
Total				
accept	489 (13.3%)	258 (6.9%)	129 (3.6%)	80 (2.3%)
no_quota_match	144 (3.9%)	173 (4.6%)	214 (5.9%)	178 (5.1%)
not_a_person	457 (12.5%)	469 (12.6%)	400 (11.1%)	410 (11.6%)
not_from_USA	478 (13.0%)	488 (13.1%)	270 (7.5%)	401 (11.4%)
null_location	1,576 (42.9%)	1,512 (40.6%)	1,444 (39.9%)	1,504 (42.7%)
quota_is_full	526 (14.3%)	821 (22.1%)	1158 (32.0%)	947 (26.9%)
Politics				
accept	190 (9.7%)	95 (4.9%)	56 (2.9%)	33 (1.8%)
no_quota_match	83 (4.2%)	88 (4.5%)	102 (5.2%)	98 (5.2%)
not_a_person	215 (11.0%)	193 (9.9%)	189 (9.7%)	186 (10.0%)
not_from_USA	225 (11.5%)	192 (9.9%)	154 (7.9%)	184 (9.9%)
null_location	854 (43.6%)	881 (45.3%)	857 (43.8%)	845 (45.3%)
quota_is_full	393 (20.1%)	495 (25.5%)	597 (30.5%)	521 (27.9%)
Trends				
accept	299 (17.5%)	163 (9.2%)	73 (4.4%)	47 (2.8%)
no_quota_match	61 (3.6%)	85 (4.8%)	112 (6.7%)	80 (4.8%)
not_a_person	242 (14.2%)	276 (15.5%)	211 (12.7%)	224 (13.6%)
not_from_USA	253 (14.8%)	296 (16.7%)	116 (7.0%)	217 (13.1%)
null_location	722 (42.2%)	631 (35.5%)	587 (35.4%)	659 (39.9%)
quota_is_full	133 (7.8%)	326 (18.3%)	561 (33.8%)	426 (25.8%)

Table 3: Breakdown of reasons for rejecting users from being included in the sample. Note that each \mathbb{X} API query is processed sequentially, meaning multiple ‘trending’ queries can be processed before the large ‘politics’ query is attended to. The processing order of the queries is random.



Figure 2: Word-cloud presenting words from the ‘trending’ queries to the \mathbb{X} API. The words are weighted by the number of users associated with each word. The uppermost row presents the words in trending queries associated with users in the broader subject pool, whilst the bottom row presents those associated with users who pass the inclusion criteria and are accepted in the survey-object.

3.2 Unfilled Quotas

I use a stratification frame defined by the following variables: `sex` \in {male,female}; `age` \in {25-34, 35-44, 45-54, 55-64, 18-24, 65+}; `hh_income` \in {[0, 25k), [25k, 50k), [50k, 75k), [75k, 100k), [100k, ∞]} ; `race_ethnicity` \in {black, white, hispanic, asian, multiracial, native }; `vote2020` \in {D, G, L,R, stayed home}. The sample frame (an extract of which is available in Table 1) is composed of 435 mutually exclusive cells.

Table 3 prompts a review of the level of granularity encoded in the stratification frame used to generate quotas. Figure 3 presents the marginal distributions associated with the filled quotas, against those of the unfilled quotas. Note that the filled-quota marginals may differ slightly from the marginals of the final survey-object, as final survey-object inference is augmented with the latest m tweets for each user, whilst the original quota-level guess is based on a single tweet (see Section 2.2.2). Comparing the unfilled subjects with the quota targets, the missing sample is more female (Female +12.5% relative to target); closer to the upper and lower tails of the age distribution ([18-24] +7%, [65, ∞) +11.9%); substantially more non-white (White -19%); significantly poorer (i<25k +12.8%); and finally substantially more likely to have not voted in the 2020 election (+13.3%).

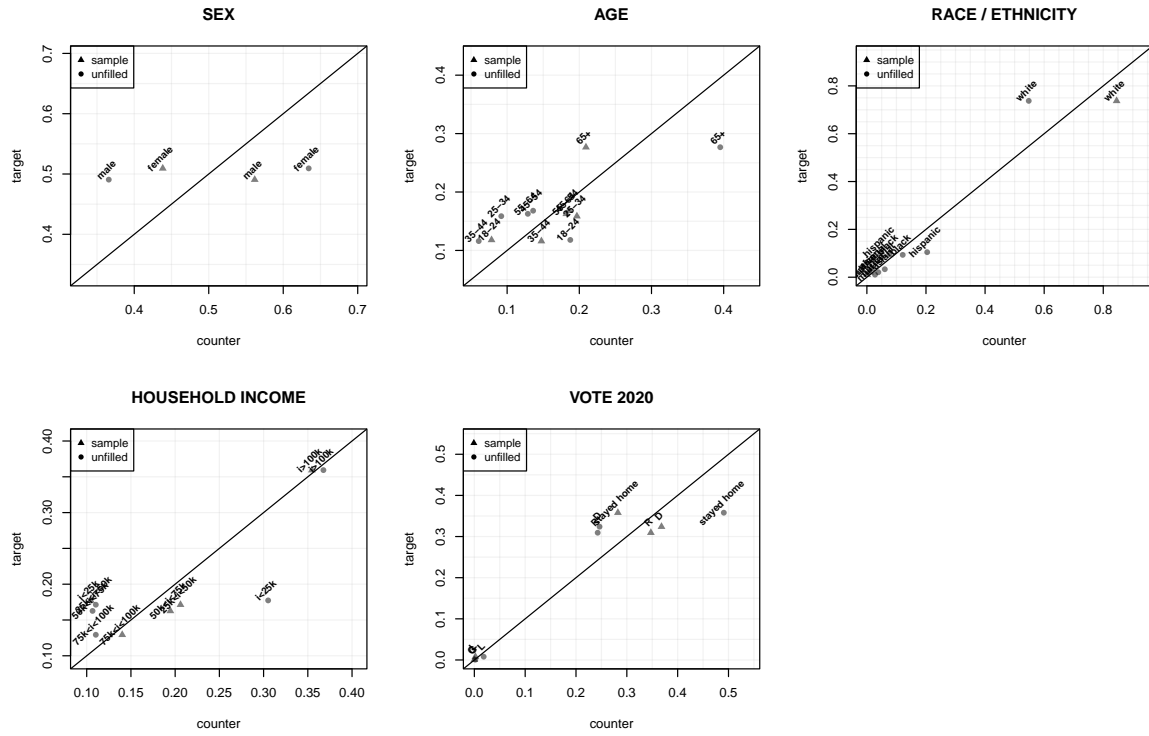


Figure 3: Comparing the filled and unfilled quotas (x-axis) against the sample-level stratification frame (y-axis).

3.3 Speculation

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 is acquired during its training phase, is responsible for a given estimate, and how much is just evident in the \mathbb{X} profile and tweets. Generally it is preferable for the LLM to not “speculate” too much, though it is not immediately clear that less speculation is always better. To some degree, we want the LLM to make decisions about user attributes under uncertainty.

I have provided the LLM with instructions (Listing 4) 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 which is directly useful to the imputation task, and distinguished this from other kinds of knowledge the LLM might leverage. The score has a categorical interpretation, which identifies “highly speculative” imputations at $S > 80$. I use this score to test the sensitivity of PoSSUM’s estimates to the dropping-out of highly-speculative ($S > 80$) records. Figure 4 presents the general distribution of speculation levels across covariates and dependent variables. Speculation levels tend to rise for “default” categories (e.g. “White” ethnicity, which is often imputed as “it is statistically the most likely”), categories which are synonymous with high-uncertainty (e.g. “No Opinion” on the favourability ratings), and those that are genuinely difficult to discern based on the data at hand (e.g. “Mixed-race” ethnicity, the “stay home” vote-choice option, or the distinguishing of fine-grained income brackets).

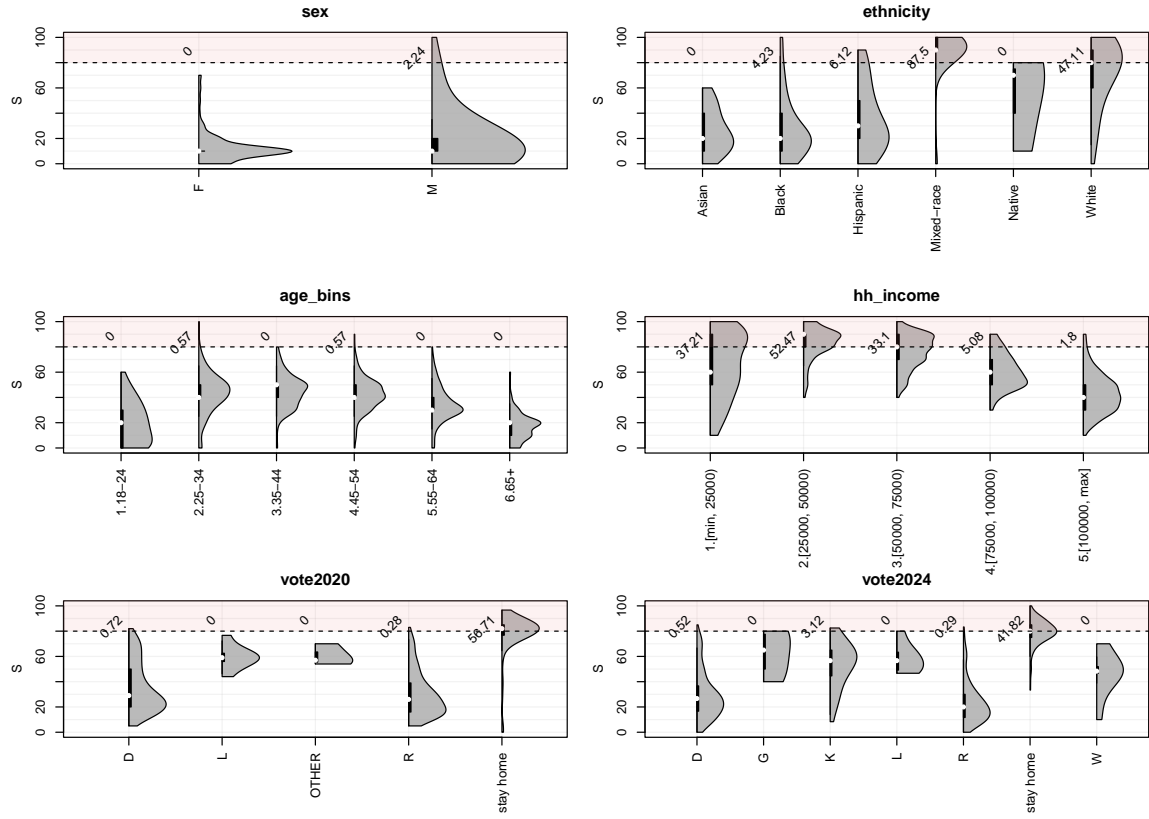


Figure 4: Distribution of the speculation score within variables used for modeling. Each distribution is accompanied by a number on the top-left, representing the % of the speculation scores which are greater than the “highly speculative” threshold.

3.4 Representative Inference

As demonstrated by Figure 3, the sample is unbalanced in important ways. The weighting method of choice here is Multilevel Regression with Post-Stratification (MrP) [10, 24, 17]. I 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 [30], 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 topline than traditional adjustments. This approach could fail if the relationship between the characteristics of the incomplete and the dependent variables differs from those of the complete cells, in which case the dependent variable estimates for the unfilled quotas will be biased.

Extending the Stratification Frame: To improve the MrP estimates I use a modified MrsP [19] procedure (Smooth MrsP). The goal of this procedure is to extend the stratification frame, which is derived from the 2021 American Community Survey [29], to include 2020 Vote Choice. It differs from traditional MrsP in that it doesn’t use the actual survey crosstabs to augment the frame, but rather it fits a model to smooth the crosstabs first, and then project these onto the existing frame. This approach can help generate more plausible estimates for ‘noisy’ cells, when the number of cells in the frame is large and the sample-size-per-cell relatively small. I use the 2022 Cooperative Election Study (CES) [27] to get estimates of 2020 recall vote⁹. I fit a deep-MrP [12, 13] model using Stan [4] to generate estimates of past-vote which leverage interactions between demographics as much as possible, in order to avoid attenuation bias in the estimated cell-level distribution. The likelihood of the model is categorical, and SoftMax is used as the link-function. The “depth” of the Bayesian Hierarchical model is given by the inclusion of marginal effects of sex, age, ethnicity, education, household income and state, as well as all two- and three- way interactions. All effects are estimated as random effects under non-centered parametrisation and recommended weakly-informative priors [9]. The Stan code for this model is available in the GitHub repository under name “`model_ai.survey_SmoothMrsP.stan`”. The resulting frame is then raked to the known state-level distribution of demographics and past vote, using the `anesrake` procedure [25]. The quota-frames used for the quota-inclusion criteria are samples from this “mother-frame”, where a new “daughter-frame” is sampled to generate targets for a new poll.

Structured Priors: The final Hierarchical Model used to generate smoothed estimates

⁹I use this dataset for the following reasons: a. it is a large sample of 60k subjects, affording greater scope for estimating interaction effects between demographic attributes; b. the alternative (ANES) was much too biased in favour of the Democratic candidate in 2020; c. it allows me flexibility to extend the frame further by 2022 vote, using the same dataset, if it is reasonable to do so at a later stage.

of the dependent variable of interest is a simple MrP with structured priors [8]. The “structure” of the model plays an important role here, as it can help smooth the learned effects of a model trained on AI generated data in a sensible way. LLMs can leverage stereotypes in making their imputations [6], 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.

I 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; state¹⁰ effects are assigned an Intrinsic Conditional Auto-regressive (ICAR) prior [7, 21, 3]; income and age effects are given random-walk priors. Separate area-level predictors are created for each dependent variable of interest. Table 4 presents the covariates and parameters used in the model for 2024 vote choice.

<i>predictor</i>	<i>level</i>	<i>description</i>	<i>index</i>	<i>domain</i>	<i>parameter</i>	<i>prior correlation structure</i>
1	global	/	/	/	α_j	iid
/	state	state_id	l	$\{1, \dots, 54\}$	λ_{sj}	spatial (BYM2)
/	individual	age_id	a	$\{1, \dots, 6\}$	η_{aj}^A	random-walk
/		income_id	h	$\{1, \dots, 5\}$	η_{hj}^H	random-walk
/		sex_id	g	$\{1, 2\}$	γ_{gj}^G	unstructured + shared variance
/		race_id	r	$\{1, \dots, 6\}$	γ_{rj}^R	unstructured + shared variance
/		vote20_id	v	$\{1, \dots, 5\}$	γ_{vj}^V	unstructured + shared variance
z_1	state	2020 R share	/	\mathbb{R}	$\beta_{1j=R}$	iid
z_2		$R - D$ 2020 margin			$\beta_{2j=R}$	
z_3		On ballot: R.F.K. Jr.			$\beta_{1j=K}$	
z_4		On ballot: Jill Stein			$\beta_{1j=G}$	
z_5		2020 G share			$\beta_{2j=G}$	
z_6		On ballot: Chase Oliver			$\beta_{1j=K}$	
z_7		2020 L share			$\beta_{2j=R}$	
z_8		On ballot: Cornel West			$\beta_{1j=W}$	
z_9		2020 “stay home” share			$\beta_{1j=\text{stay_home}}$	

Table 4: Model Predictors and Parameters for the 2024 vote-choice model. ‘iid’ refers to fully independent parameters, or ‘fixed’ effects [11]. ‘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.

The covariates used for other models involving different dependent variables are very similar, and available in the GitHub repository. Equations 1 to 16 present the setup of the Hierarchical Bayesian model – see [5] for a more attentive explanation of each model component. I describe the generation of given choice $j \in \{1, \dots, J\}$, made by a sampled user $i \in \{1, \dots, n\}$, as follows:

¹⁰Because 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.

$$y_{ij} \sim \text{Categorical}(\pi_{i1}, \dots, \pi_{iJ}) \quad \text{likelihood} \quad (1)$$

$$\pi_{ij} = \frac{\exp(\mu_{ij})}{\sum_j \exp(\mu_{i,j})}; \quad \text{softmax link} \quad (2)$$

$$\mu_{ij} = \alpha_j + \quad (3)$$

$$\lambda_{\text{state_id}[i],j} + \eta_{\text{age_id}[i],j}^A + \eta_{\text{income_id}[i],j}^H + \quad (4)$$

$$\gamma_{\text{sex_id}[i],j}^G + \gamma_{\text{race_id}[i],j}^R + \gamma_{\text{vote20_id}[i],j}^V + \quad (5)$$

$$\beta_j \mathbf{Z}^j; \quad \text{linear predictor} \quad (6)$$

$$\alpha_j \sim N(0, 1); \quad \text{intercept} \quad (7)$$

$$\lambda_{sj} = \sigma_j^\lambda \left(\phi_{sj} \sqrt{(1 - \xi_j)} + \psi_{sj} \sqrt{(\xi_j / \epsilon)} \right); \quad \text{BYM2 effects} \quad (8)$$

$$\phi_{sj} \sim N(0, 1); \quad \text{unstructured effect} \quad (9)$$

$$\psi_{sj} \mid \psi_{s'j} \sim N \left(\frac{\sum_{l \neq j} \psi_{s'l}}{\nu_l}, \frac{1}{\sqrt{\nu_l}} \right); \quad \text{conditional auto-regressive effects} \quad (10)$$

$$\xi_j \sim \text{Beta} \left(\frac{1}{2}, \frac{1}{2} \right); \quad \text{mixing weights} \quad (11)$$

$$\sigma_j^\lambda \sim N^+(0, 1); \quad \text{state-level scale} \quad (12)$$

$$\gamma_{uj}^U \mid \gamma_{u-1j}^U \dots \gamma_{1j}^U \sim N(\gamma_{u-1j}^U, \sigma_j^U), \quad \forall U \in \{A, H\}; \quad \text{random walk effects} \quad (13)$$

$$\gamma_{uj}^U \sim N(0, \sigma_j^U), \quad \forall U \in \{G, R, V\}; \quad \text{unstructured effects} \quad (14)$$

$$\sigma_j^U \sim N^+(0, 1); \quad \text{random effect scales} \quad (15)$$

$$\beta_j \sim N^+(0, 1). \quad \text{fixed effects} \quad (16)$$

3.5 Topline & Crosstabs

In this section I present the PoSSUM estimates for this poll. Results for the distribution of votes across candidate choices are given as a share of the likely voter population, whilst shares of those who will turn-out or abstain are provided as a fraction of the adult population of the US, according to the 2021 ACS.

Table 5 presents the topline, national vote share estimates. Crosstabs by Region (Table 6), Sex (Table 7), Race/Ethnicity (Table 11), Education (Table 8), Household Income (Table 9) and 2020 Vote Choice (Table 12) are available below.

Table 5: Topline estimates of 2024 Vote Choice.

Population	Vote2024	Topline
Voters	Kamala Harris	45.8 (43.8 , 47.8)
Voters	Donald Trump (R)	45.6 (43.3 , 47.2)
Voters	Robert F. Kennedy Jr. (Ind.)	5.5 (3.8 , 7.5)
Voters	Jill Stein (G)	0.9 (0.5 , 2.6)
Voters	Cornel West (Ind.)	0.8 (0.4 , 1.8)
Voters	Chase Oliver (L)	1 (0.4 , 1.9)
Adults	Stay Home	28.3 (26 , 30.5)
Adults	Turnout	71.7 (69.5 , 74)

Table 6: Estimates of 2024 Vote Choice by Region

Population	Vote2024	Midwest	Northeast	South	West
Voters	Kamala Harris (D)	43.6 (41.4 , 45.6)	51.2 (48.4 , 53.4)	42 (40 , 44.6)	49.8 (46 , 52.4)
Voters	Donald Trump (R)	48.8 (46.6 , 50.8)	42.8 (39.9 , 45.7)	49.2 (46.5 , 51.6)	39.4 (36.4 , 41.6)
Voters	Robert F. Kennedy Jr. (Ind.)	5.5 (3.2 , 7.9)	3.3 (1.9 , 5.2)	5.6 (4 , 7.9)	6.9 (4.3 , 10.3)
Voters	Jill Stein (G)	0.4 (0.2 , 1.3)	0.9 (0.4 , 2.5)	0.8 (0.3 , 2.1)	1.6 (0.7 , 4.7)
Voters	Cornel West (Ind.)	0.8 (0.3 , 1.8)	1 (0.4 , 2.2)	0.9 (0.3 , 1.7)	0.7 (0.2 , 1.5)
Voters	Chase Oliver (L)	0.8 (0.3 , 1.9)	0.5 (0.2 , 1.2)	1 (0.4 , 2)	1.3 (0.5 , 3)
Adults	Stay Home	28 (25 , 30.8)	30.4 (26.1 , 34.5)	29.9 (27.2 , 32.7)	24 (19.5 , 27.8)
Adults	Turnout	72 (69.2 , 75)	69.6 (65.5 , 73.9)	70.1 (67.3 , 72.8)	76 (72.2 , 80.5)

Table 7: Estimates of 2024 Vote Choice by Sex

Population	Vote2024	F	M
Voters	Kamala Harris (D)	50.6 (47.8 , 53)	40.3 (38 , 43)
Voters	Donald Trump (R)	42.1 (39.6 , 44.4)	49.6 (47.3 , 52.2)
Voters	Robert F. Kennedy Jr. (Ind.)	4.7 (2.9 , 7.4)	6 (4 , 8.7)
Voters	Jill Stein (G)	1 (0.4 , 2.6)	0.9 (0.3 , 2.6)
Voters	Cornel West (Ind.)	0.6 (0.2 , 1.7)	1 (0.4 , 2)
Voters	Chase Oliver (L)	0.6 (0 , 1.6)	1.4 (0.6 , 2.7)
Adults	Stay Home	26 (22 , 29.6)	30.8 (28.1 , 33.7)
Adults	Turnout	74 (70.4 , 78)	69.2 (66.3 , 71.9)

Table 8: Estimates of 2024 Vote Choice by College Grad.

Population	Vote2024	Does not Have a College Degree	Has a College Degree
Voters	Kamala Harris (D)	41.5 (39.2 , 44.3)	53.1 (50.9 , 54.9)
Voters	Donald Trump (R)	49.2 (46.6 , 52)	39.2 (37.4 , 40.9)
Voters	Robert F. Kennedy Jr. (Ind.)	5.6 (3.8 , 7.9)	5.2 (3.8 , 7)
Voters	Jill Stein (G)	1 (0.4 , 2.9)	0.9 (0.4 , 2)
Voters	Cornel West (Ind.)	1 (0.4 , 2)	0.6 (0.2 , 1.4)
Voters	Chase Oliver (L)	1.1 (0.4 , 2.1)	0.8 (0.3 , 1.5)
Adults	Stay Home	33.5 (31.1 , 36.3)	17 (15 , 18.9)
Adults	Turnout	66.5 (63.7 , 68.9)	83 (81.1 , 85)

Table 9: Estimates of 2024 Vote Choice by Houshold Income

Population	Vote2024	[0, 25k)	[25k, 50k)	[50k, 75k)	[75k, 100k)	[100k, +)
Voters	Kamala Harris (D)	47.6 (41.2 , 52.8)	43.5 (40.1 , 47.5)	43.1 (40.4 , 46.1)	43.6 (40.3 , 46.8)	48.1 (45.1 , 50.3)
Voters	Donald Trump (R)	41.8 (34.2 , 47.7)	48.5 (44.1 , 52.2)	49.8 (46.6 , 52.9)	49.1 (45.9 , 52.7)	43.1 (40.3 , 45.2)
Voters	Robert F. Kennedy Jr. (Ind.)	5 (1.9 , 9.3)	4.9 (2.9 , 7.5)	4.3 (2.4 , 6.7)	4.6 (2.9 , 7.3)	6.6 (4.1 , 9.4)
Voters	Jill Stein (G)	1.6 (0.4 , 7)	0.9 (0.3 , 2.8)	0.6 (0.1 , 1.9)	0.6 (0.1 , 1.9)	0.8 (0.3 , 2.4)
Voters	Cornel West (Ind.)	1.6 (0.5 , 3.6)	0.9 (0.4 , 1.8)	0.8 (0.3 , 1.6)	0.7 (0.2 , 1.5)	0.6 (0.2 , 1.5)
Voters	Chase Oliver (L)	1.1 (0.3 , 3.4)	0.9 (0.3 , 2.3)	0.8 (0.3 , 1.8)	0.9 (0.4 , 2.1)	0.9 (0.4 , 1.9)
Adults	Stay Home	47.2 (40.7 , 53.2)	34.9 (31.3 , 38.9)	27 (23.7 , 30.5)	20.7 (16.4 , 23.8)	19.5 (16.5 , 22)
Adults	Turnout	52.8 (46.8 , 59.3)	65.1 (61.1 , 68.7)	73 (69.5 , 76.3)	79.3 (76.2 , 83.6)	80.5 (78 , 83.5)

Table 10: Estimates of 2024 Vote Choice by Age

Population	Vote2024	18-24	25-34	35-44	45-54	55-64	65+
Voters	Kamala Harris (D)	54.5 (44.6 , 60.6)	48.7 (43.9 , 52.1)	49.2 (45.2 , 52.5)	45 (41 , 48.2)	43.8 (41.2 , 46.8)	43.2 (40.5 , 45.2)
Voters	Donald Trump (R)	28.3 (21.9 , 34.5)	35.8 (31.6 , 40.2)	37.9 (34.5 , 41)	47.2 (43.8 , 51.2)	50.6 (47.5 , 52.9)	53.3 (50.2 , 55.7)
Voters	Robert F. Kennedy Jr. (Ind.)	9.4 (3.4 , 17.7)	11.6 (7.6 , 17)	9 (6.4 , 14)	5.1 (2.7 , 7.7)	3.1 (1.6 , 5.3)	1.6 (0.5 , 3.8)
Voters	Jill Stein (G)	1.1 (0.2 , 4.8)	1 (0.2 , 3.1)	1 (0.4 , 2.8)	0.9 (0.4 , 2.5)	0.9 (0.3 , 2.1)	0.8 (0.2 , 2.9)
Voters	Cornel West (Ind.)	4.5 (1.1 , 12)	1.4 (0.5 , 3)	0.9 (0.4 , 2.7)	0.3 (0 , 1.3)	0.3 (0.1 , 0.8)	0.2 (0 , 0.8)
Voters	Chase Oliver (L)	1.1 (0.3 , 3.6)	1.1 (0.4 , 2.4)	0.9 (0.4 , 2.2)	1.1 (0.4 , 2.7)	0.9 (0.3 , 1.8)	0.7 (0.1 , 1.6)
Adults	Stay Home	49.8 (40.3 , 59.9)	40.2 (35.6 , 44.7)	30 (26.3 , 33.8)	23.7 (17.6 , 27.7)	21 (17.4 , 24.7)	20.5 (15.7 , 24.3)
Adults	Turnout	50.2 (40.1 , 59.7)	59.8 (55.3 , 64.4)	70 (66.2 , 73.7)	76.3 (72.3 , 82.4)	79 (75.3 , 82.6)	79.5 (75.7 , 84.3)

Table 11: Estimates of 2024 Vote Choice by Race/Ethnicity

Population	Vote2024	Asian	Black	Hispanic	Mixed-Race	Native	White
Voters	Kamala Harris (D)	61.4 (50.7 , 68.5)	76.9 (70.4 , 83)	57.3 (50.7 , 62.3)	41.1 (34.9 , 47)	39.9 (29.7 , 50.1)	40.1 (38.1 , 42.3)
Voters	Donald Trump (R)	26.7 (22.5 , 34)	15.5 (10.1 , 20.5)	32.4 (27.1 , 37.9)	44.5 (36 , 54.3)	43 (30.3 , 55.6)	52 (49.7 , 54.2)
Voters	Robert F. Kennedy Jr. (Ind.)	5.2 (0.6 , 10)	2.7 (0.4 , 7.5)	6 (1.9 , 12.5)	8.4 (3.1 , 17.3)	7.8 (2.7 , 32)	5.4 (3.7 , 7.5)
Voters	Jill Stein (G)	2.2 (0.7 , 12.6)	1.1 (0.3 , 4.9)	1.3 (0.2 , 7.1)	0.9 (0.2 , 3.1)	1.4 (0.3 , 8.8)	0.7 (0.3 , 1.6)
Voters	Cornel West (Ind.)	0.7 (0.2 , 2.5)	2 (0.6 , 4.7)	0.9 (0.2 , 2.2)	1.6 (0.4 , 4.2)	1 (0.2 , 3.4)	0.6 (0.3 , 1.5)
Voters	Chase Oliver (L)	1.1 (0.3 , 3.8)	1.1 (0.3 , 3.2)	1.2 (0.3 , 4)	1.2 (0.3 , 4.4)	1.6 (0.3 , 5.3)	0.8 (0.4 , 1.7)
Adults	Stay Home	24 (13.4 , 32.4)	35.9 (31.3 , 40.7)	35.3 (28.6 , 42.3)	29 (23 , 37.1)	33.4 (19.7 , 41.9)	26.4 (24.1 , 28.4)
Adults	Turnout	76 (67.6 , 86.6)	64.1 (59.3 , 68.7)	64.7 (57.7 , 71.4)	71 (62.9 , 77)	66.6 (58.1 , 80.3)	73.6 (71.6 , 75.9)

Table 12: Estimates of 2024 Vote Choice by 2020 Vote Choice

Population	Vote2024	D	R	L	G	OTHER	Stay Home
Voters	Kamala Harris (D)	93.2 (90.4 , 95.1)	0.6 (0.2 , 1.7)	8.2 (2.4 , 25.2)	11.6 (1.8 , 32.4)	22.4 (6.5 , 42.5)	33.9 (21.2 , 49)
Voters	Donald Trump (R)	1 (0.3 , 2.3)	95.4 (91.9 , 97)	1.8 (0.1 , 17.2)	16.3 (0.1 , 90.5)	8.5 (0.2 , 34.4)	39 (23.6 , 49.6)
Voters	Robert F. Kennedy Jr. (Ind.)	2.6 (1.4 , 4.6)	3.1 (1.7 , 6)	55.3 (29 , 77.9)	19.3 (0.4 , 85.1)	27.7 (8.3 , 53.4)	18.9 (10.9 , 30.1)
Voters	Jill Stein (G)	1.3 (0.5 , 3.2)	0 (0 , 0.4)	0.4 (0 , 5.4)	0.9 (0 , 21.7)	1.2 (0.1 , 14.1)	2.3 (0.4 , 11.3)
Voters	Cornel West (Ind.)	1 (0.4 , 2.5)	0.3 (0 , 1.1)	1.2 (0 , 8.4)	1.3 (0.1 , 19.9)	14.9 (4.3 , 30.2)	1 (0.1 , 3.5)
Voters	Chase Oliver (L)	0.2 (0 , 0.9)	0.3 (0 , 0.9)	24.7 (8.1 , 46.3)	3.4 (0 , 68.9)	13.1 (2.5 , 37.9)	2.8 (0.5 , 10.9)
Adults	Stay Home	7.4 (4.8 , 10.1)	2.6 (1.2 , 5.1)	5.2 (0.3 , 26.3)	13 (0.7 , 71.8)	12.5 (0.6 , 48.7)	75.6 (69.9 , 81.8)
Adults	Turnout	92.6 (89.9 , 95.2)	97.4 (94.9 , 98.8)	94.8 (73.7 , 99.7)	87 (28.2 , 99.3)	87.5 (51.3 , 99.4)	24.4 (18.2 , 30.1)

3.6 Cross-Pollsters Comparison

I compare the topline and crosstabs derived from PoSSUM against those of other pollsters were in the field around the same time. The inclusion rule for a poll is that it shares at least 1 fieldwork day with PoSSUM. The main PoSSUM estimate is PoSSUM MrP. PoSSUM Raw represents the unweighted estimate, and is a useful comparison point to understand the effects of the weighting procedure. PoSSUM MrP - Spec. Mod. and PoSSUM Raw - Spec. Mod represent versions of the MrP and unweighted estimates respectively, under the exclusion of highly speculative records¹¹. Note here this leads to a reduction in sample size of over 50%, leading these estimates to be extraordinarily noisy for a single poll.

The empirical distribution of weighted PoSSUM estimates is available as a direct result of the MrP, in the form of 104 posterior simulations. Published crosstabs from typically do not provide much information beyond point-estimates, though some do provide counts at the crosstab level. We can leverage these to fit a simple conjugate Dirichlet-Multinomial model (Equation 17), independently for each poll and crosstab, and recover posterior simulations for the crosstab-level vote distribution. Note here the n for each poll is typically the weighted- n for the given crosstab, though in some cases, pollsters only make the unweighted- n available at the crosstab-level.

$$y \sim \text{Multinomial}(\boldsymbol{\pi}, n); \quad (17)$$

$$\boldsymbol{\pi} \sim \text{Dirichlet}\left(\frac{\mathbf{1}}{2}\right) \quad (18)$$

Unfortunately, there is large heterogeneity in pollsters' methodologies, making direct comparison to PoSSUM challenging. There is variation in the number of candidates which are presented as options to the respondents; the underlying population of respondents; the granularity of the crosstabs; etc. One option would be to model these differences explicitly, and project each poll onto the same methodological space as PoSSUM, to create a fair comparison. However the number of polls is not large, and any modeling risks denaturing the published estimates, which is ultimately what the public consumes. I therefore perform only a minimal distortion to align estimates – namely I remove all ‘would not vote’ options for those pollsters who have them, so that I can model likely voters in each poll. I further aggregate all vote choice options with the exception of the two main parties and R.F.K. Jr. into an ‘Other’ category.

Figure 5 presents the topline comparison across pollsters. Figures 6 to 23 present comparisons for each crosstab available for a large-enough number of pollsters, and with a large-enough n -per-pollster, to make for useful comparisons. The headline finding here is that, on the whole, PoSSUM MrP estimates are not unusual and broadly fit with patterns shown by other pollsters.

¹¹A record is considered highly speculative if at least 1 amongst covariates and the dependent variable of interest is highly speculative ($S > 80$).

Like other pollsters, PoSSUM MrP has some marked house-effects: i. lower-than-average (4 % points on the point estimate) share of Harris voters in the topline, largely driven by a lower share amongst white voters (3 % points on the point estimate); ii. greater-than-average R.F.K. Jr. support across most crosstabs (1.5 % points on the topline point estimate); iii. sharper age-gradient, with Harris performing better-than-average amongst younger voters, and worse-than-average amongst older age-groups; iv. lower-than-average % of ‘switchers’ between the two major parties, relative to the 2020 vote.

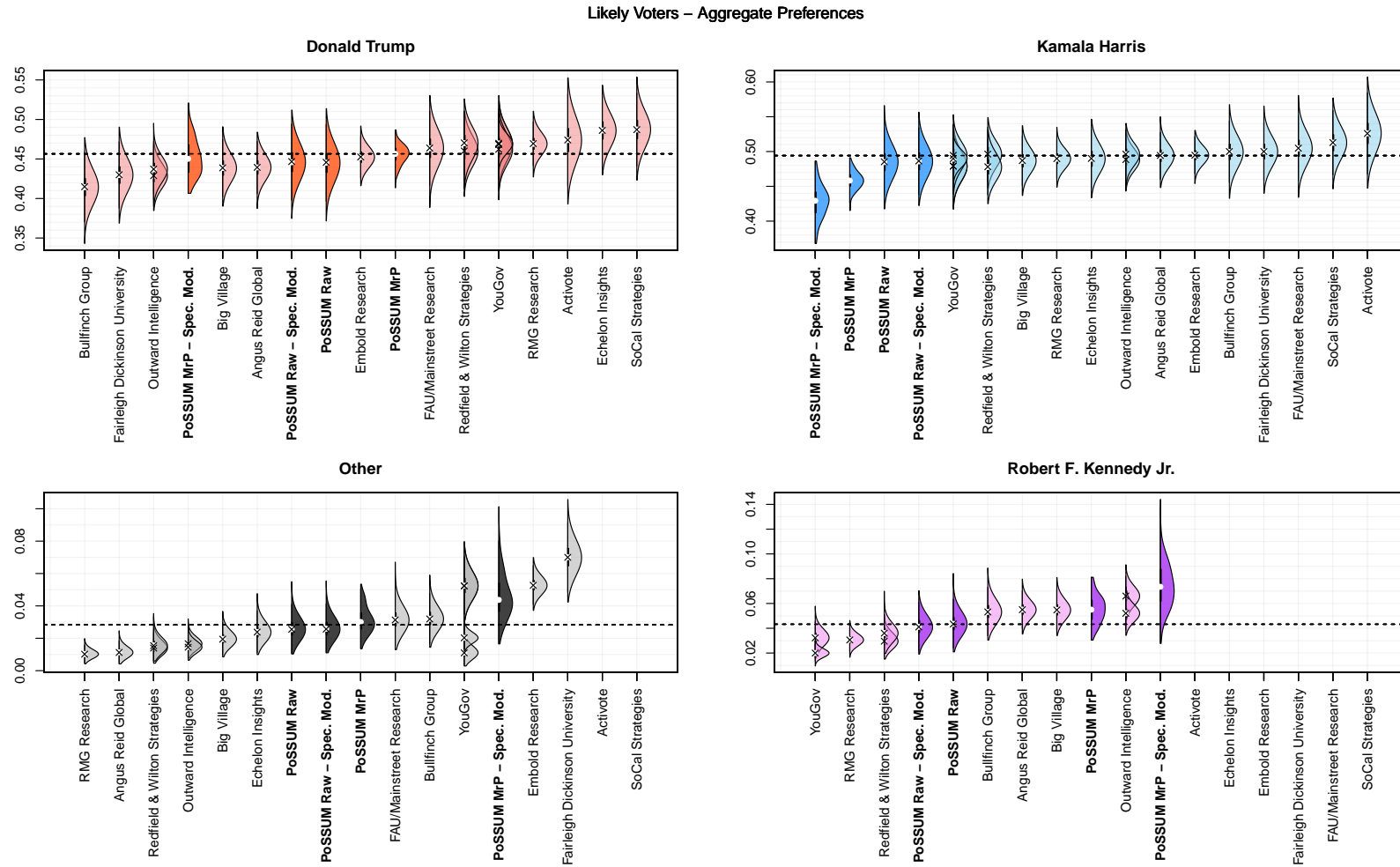


Figure 5: A comparison of the topline distribution of 2024 vote choice across pollsters. The dotted line represents the simple arithmetic average of all polls. PoSSUM MrP is the primary estimate of interest.

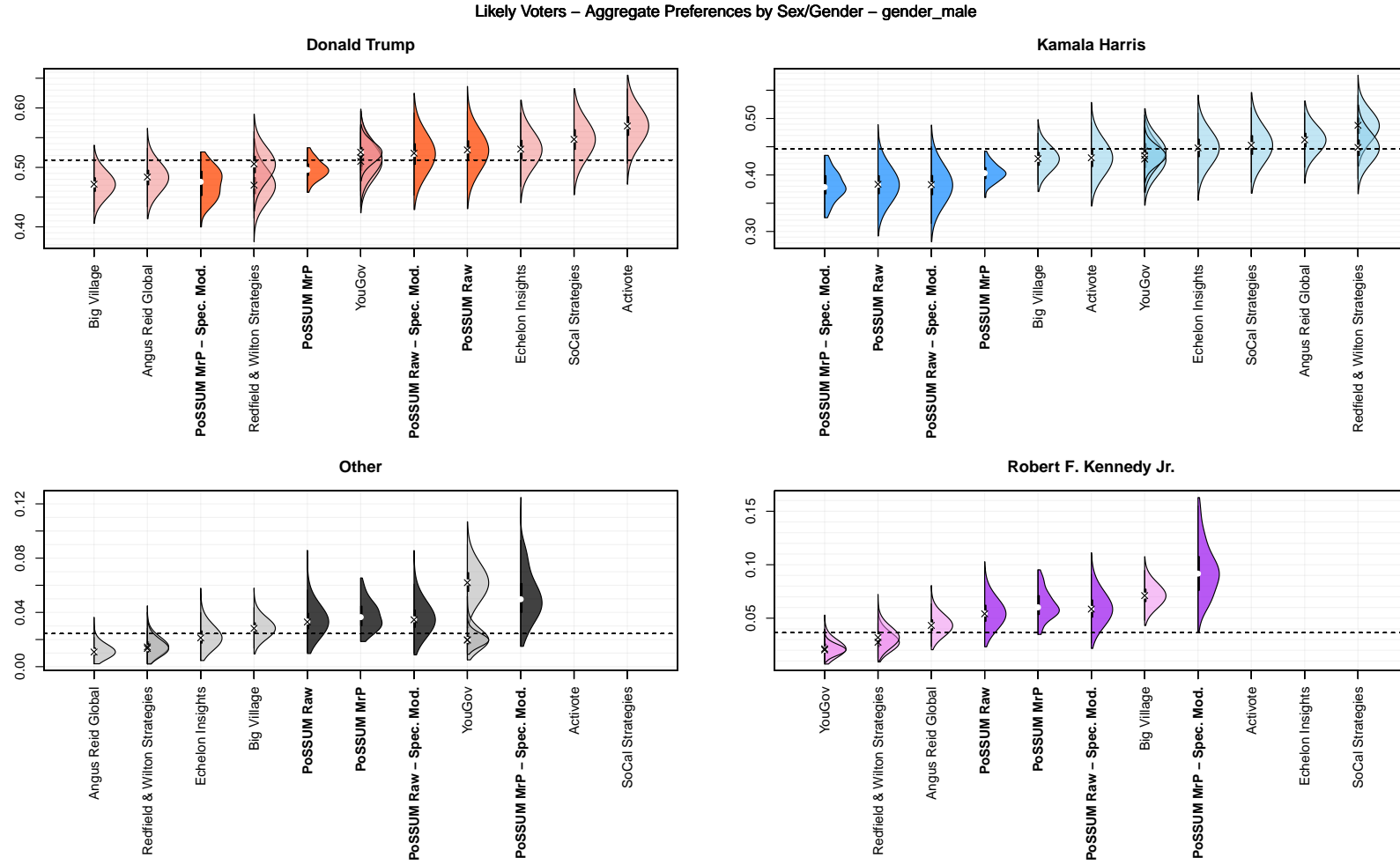


Figure 6: A comparison of the Male distribution of 2024 vote choice across pollsters. The dotted line represents the simple arithmetic average of all polls. PoSSUM MrP is the primary estimate of interest. Large intervals for a given pollster correspond to small n for the given crosstab.

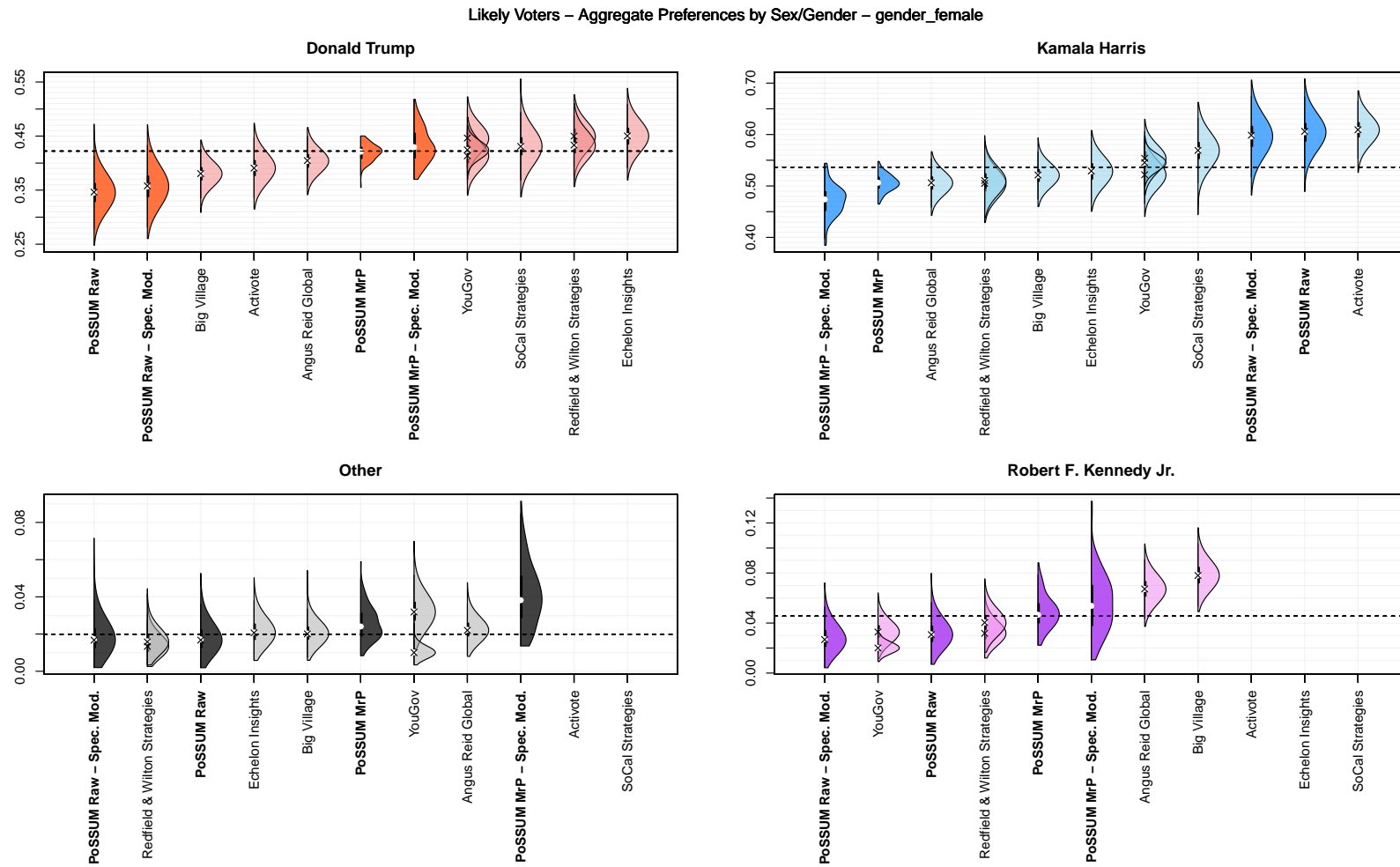


Figure 7: A comparison of the Female distribution of 2024 vote choice across pollsters. The dotted line represents the simple arithmetic average of all polls. PoSSUM MrP is the primary estimate of interest. PoSSUM MrP is the primary estimate of interest. Large intervals for a given pollster correspond to small n for the given crosstab

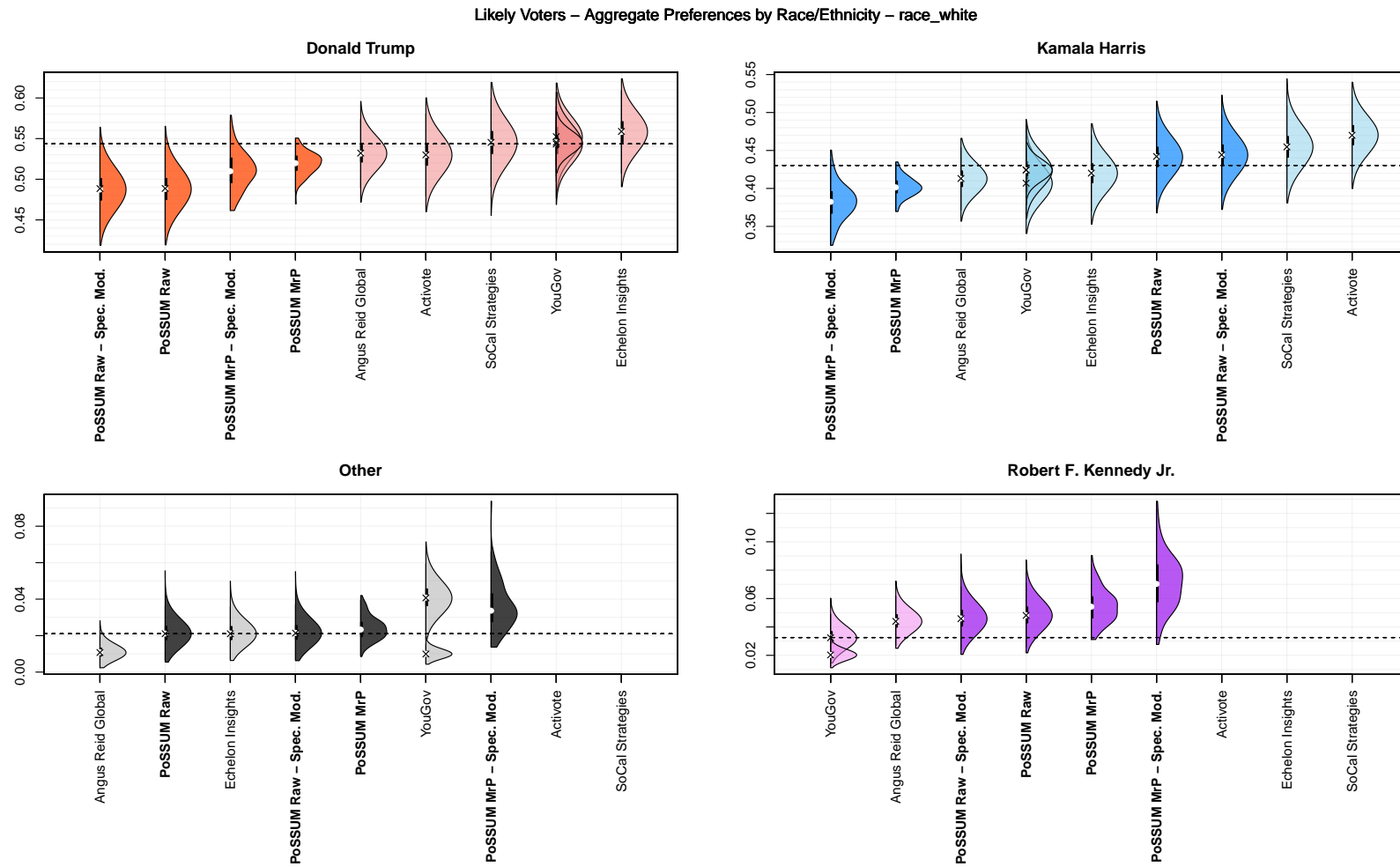


Figure 8: A comparison of the White distribution of 2024 vote choice across pollsters. The dotted line represents the simple arithmetic average of all polls. PoSSUM MrP is the primary estimate of interest. Large intervals for a given pollster correspond to small n for the given crosstab

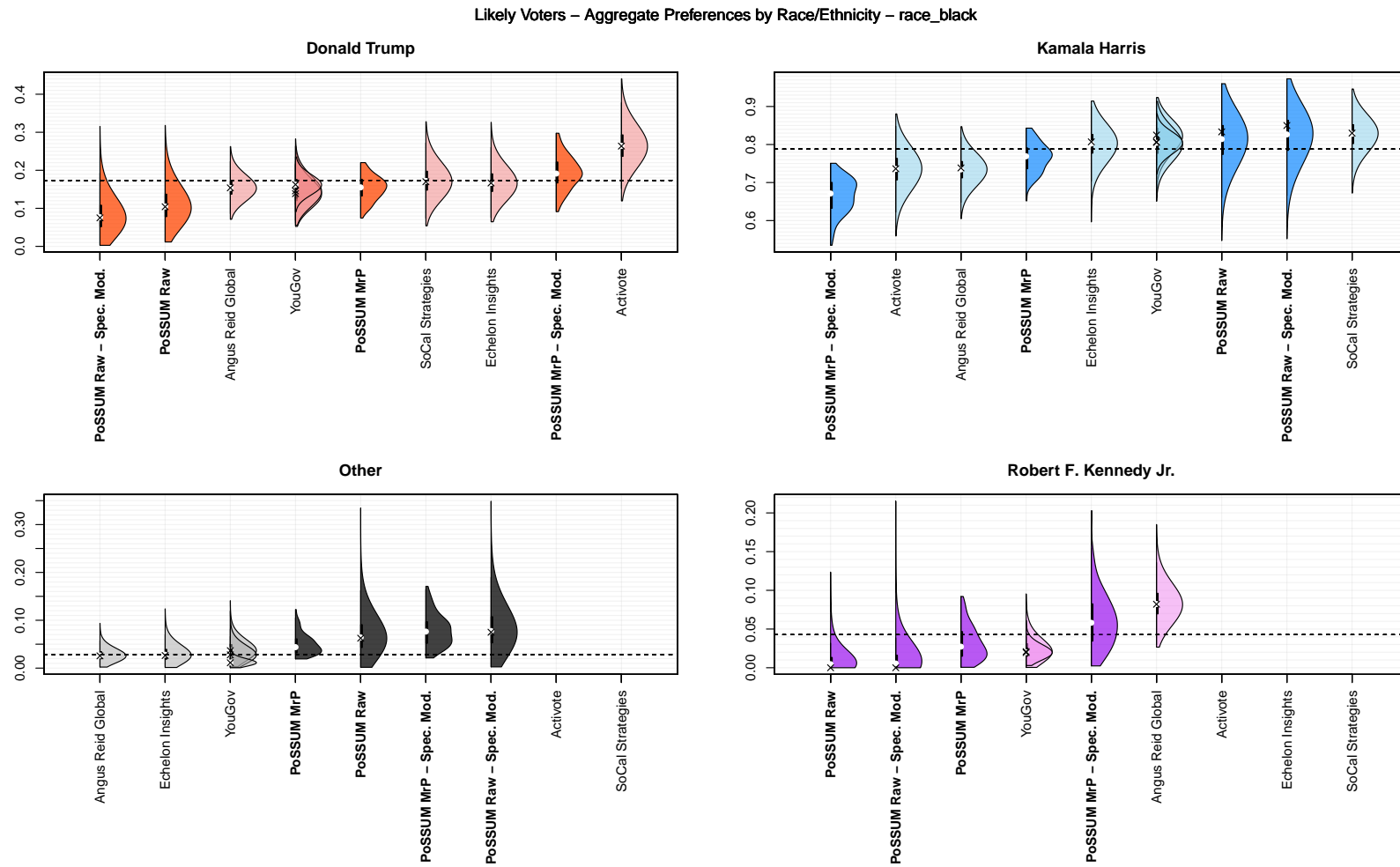


Figure 9: A comparison of the Black distribution of 2024 vote choice across pollsters. The dotted line represents the simple arithmetic average of all polls. PoSSUM MrP is the primary estimate of interest. Large intervals for a given pollster correspond to small n for the given crosstab

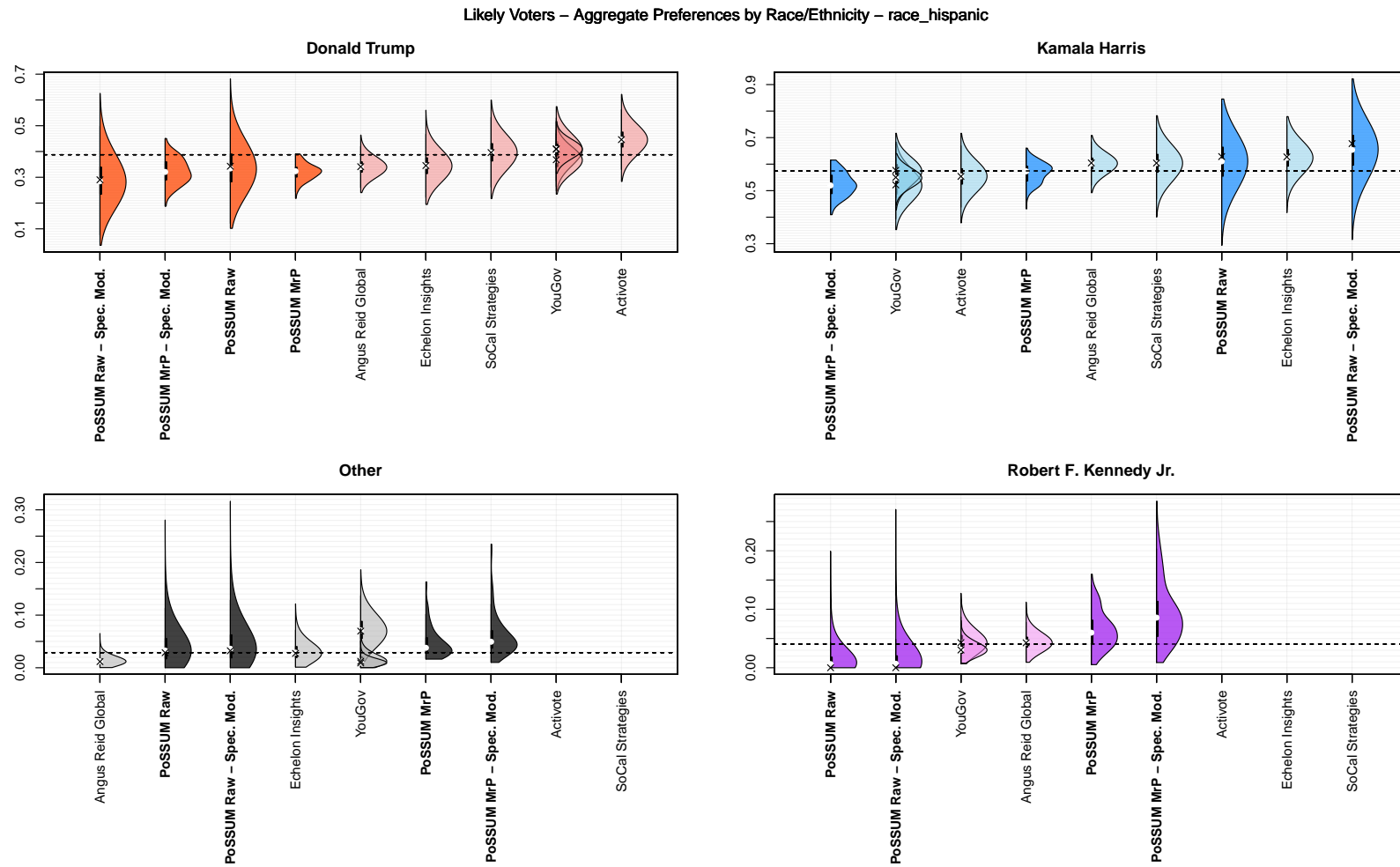


Figure 10: A comparison of the Hispanic distribution of 2024 vote choice across pollsters. The dotted line represents the simple arithmetic average of all polls. PoSSUM MrP is the primary estimate of interest. Large intervals for a given pollster correspond to small n for the given crosstab.

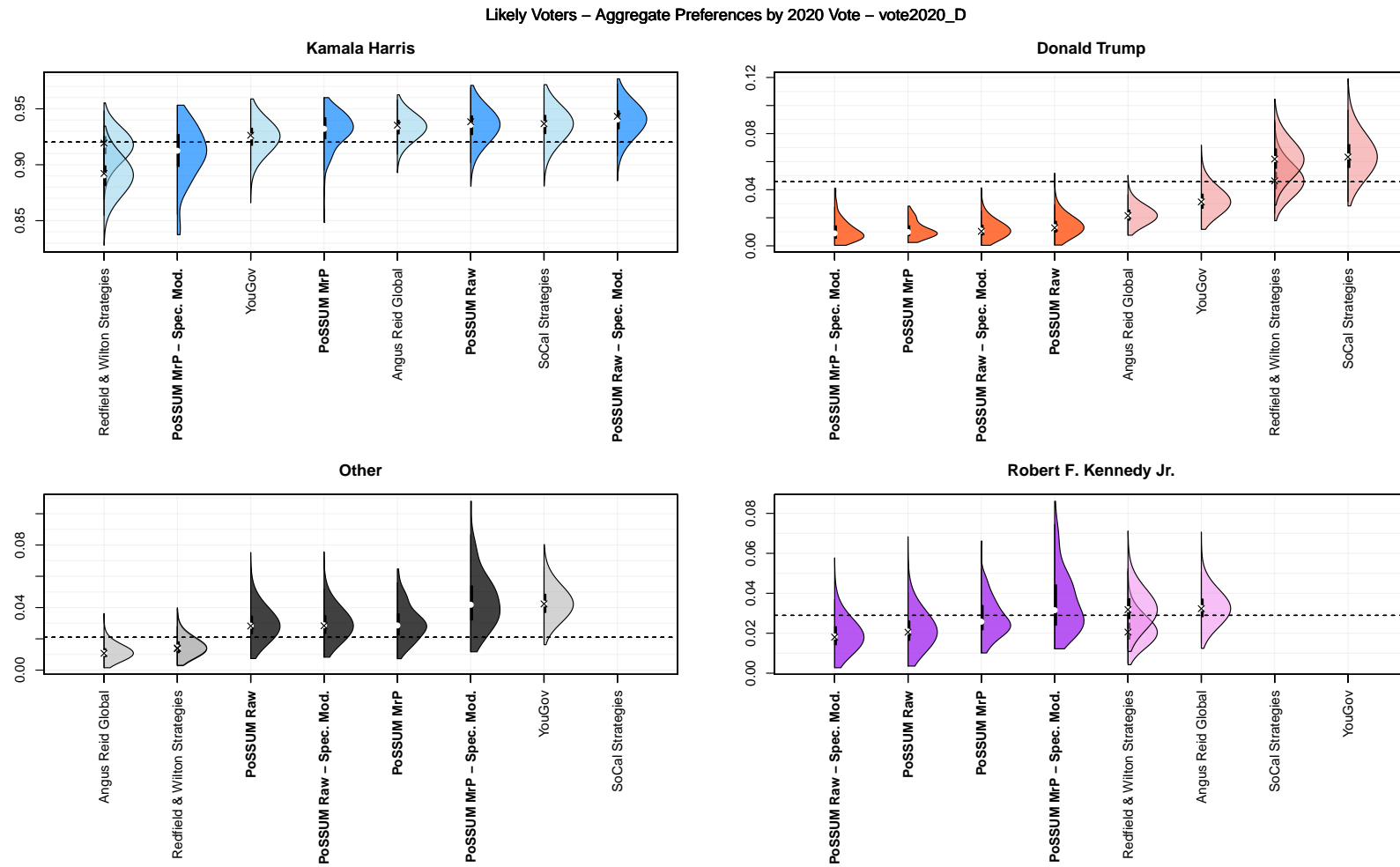


Figure 11: A comparison of the distribution of 2024 vote choice across pollsters, for those who voted for Joe Biden (D) in 2020. The dotted line represents the simple arithmetic average of all polls. PoSSUM MrP is the primary estimate of interest. Large intervals for a given pollster correspond to small n for the given crosstab.

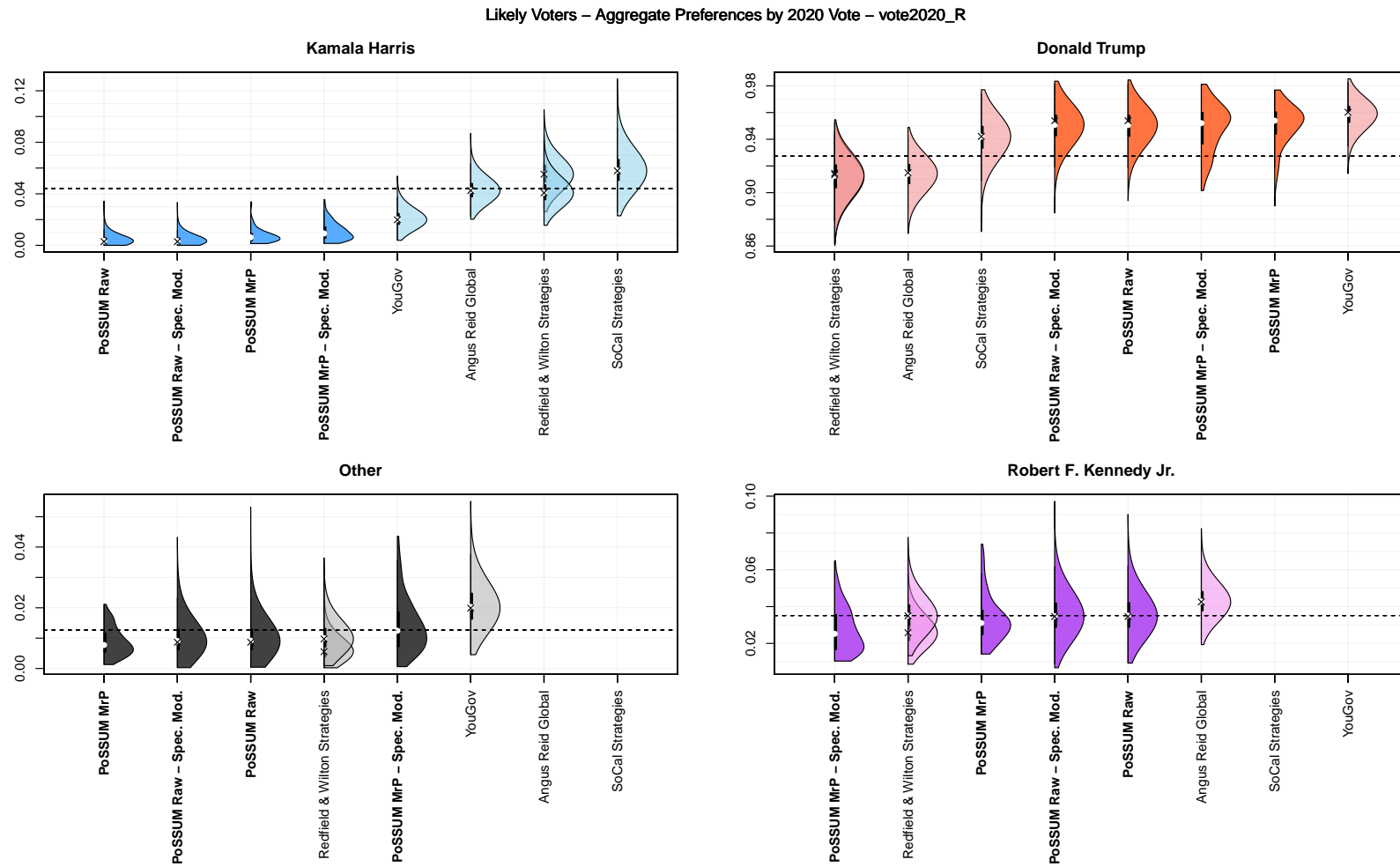


Figure 12: A comparison of the distribution of 2024 vote choice across pollsters, for those who voted for Donald Trump (R) in 2020. The dotted line represents the simple arithmetic average of all polls. PoSSUM MrP is the primary estimate of interest. Large intervals for a given pollster correspond to small n for the given crosstab.

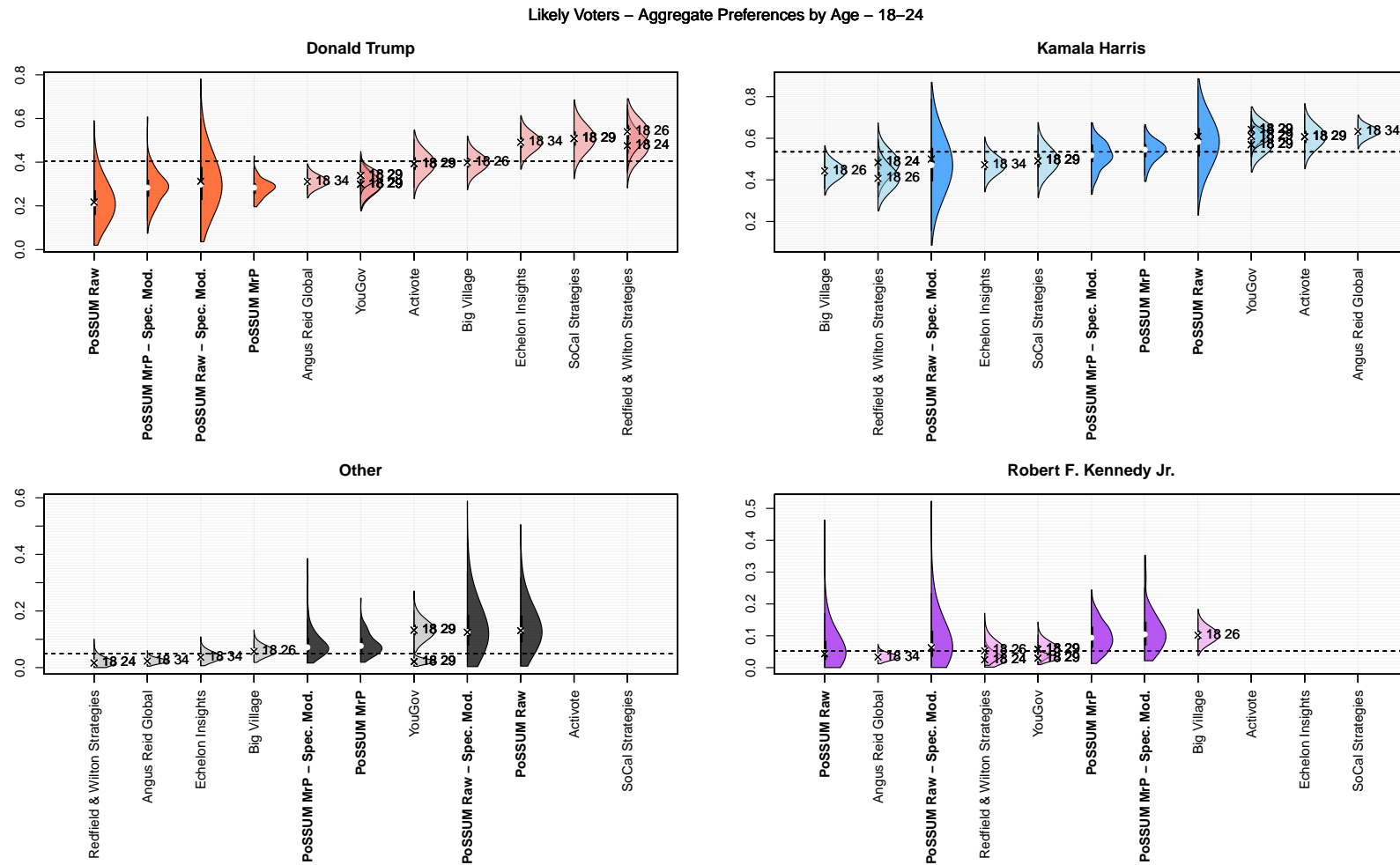


Figure 13: A comparison of the 18-24 years old distribution of 2024 vote choice across pollsters. The dotted line represents the simple arithmetic average of all polls. PoSSUM MrP is the primary estimate of interest. Large intervals for a given pollster correspond to small n for the given crosstab. To enable a comparison with multiple pollsters, and account for heterogeneity in crosstab granularity, I present the distribution of all age-groups which overlap with the 18-24 category.

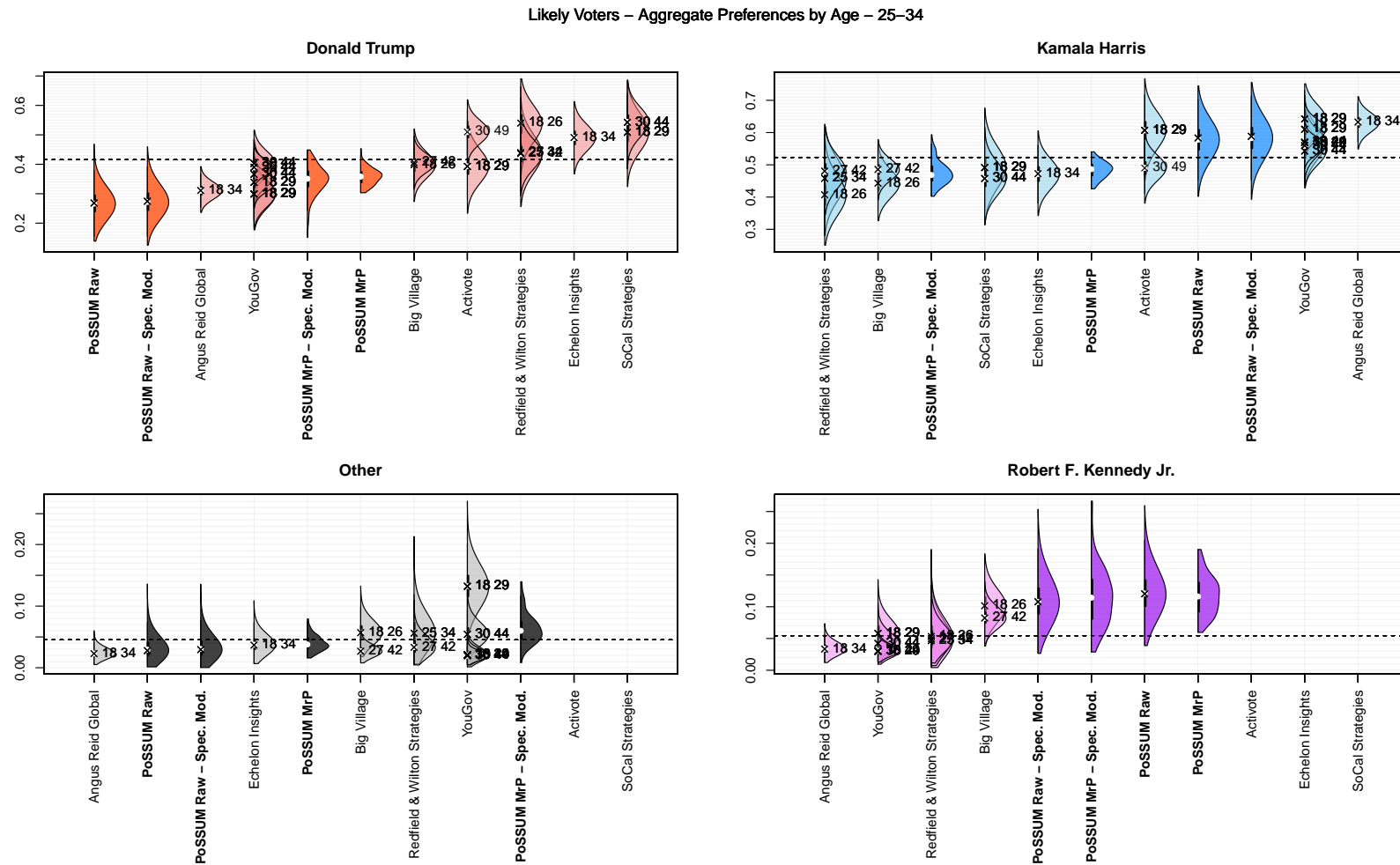


Figure 14: A comparison of the 25-34 years old distribution of 2024 vote choice across pollsters. The dotted line represents the simple arithmetic average of all polls. PoSSUM MrP is the primary estimate of interest. Large intervals for a given pollster correspond to small n for the given crosstab. To enable a comparison with multiple pollsters, and account for heterogeneity in crosstab granularity, I present the distribution of all age-groups which overlap with the 25-34 category.

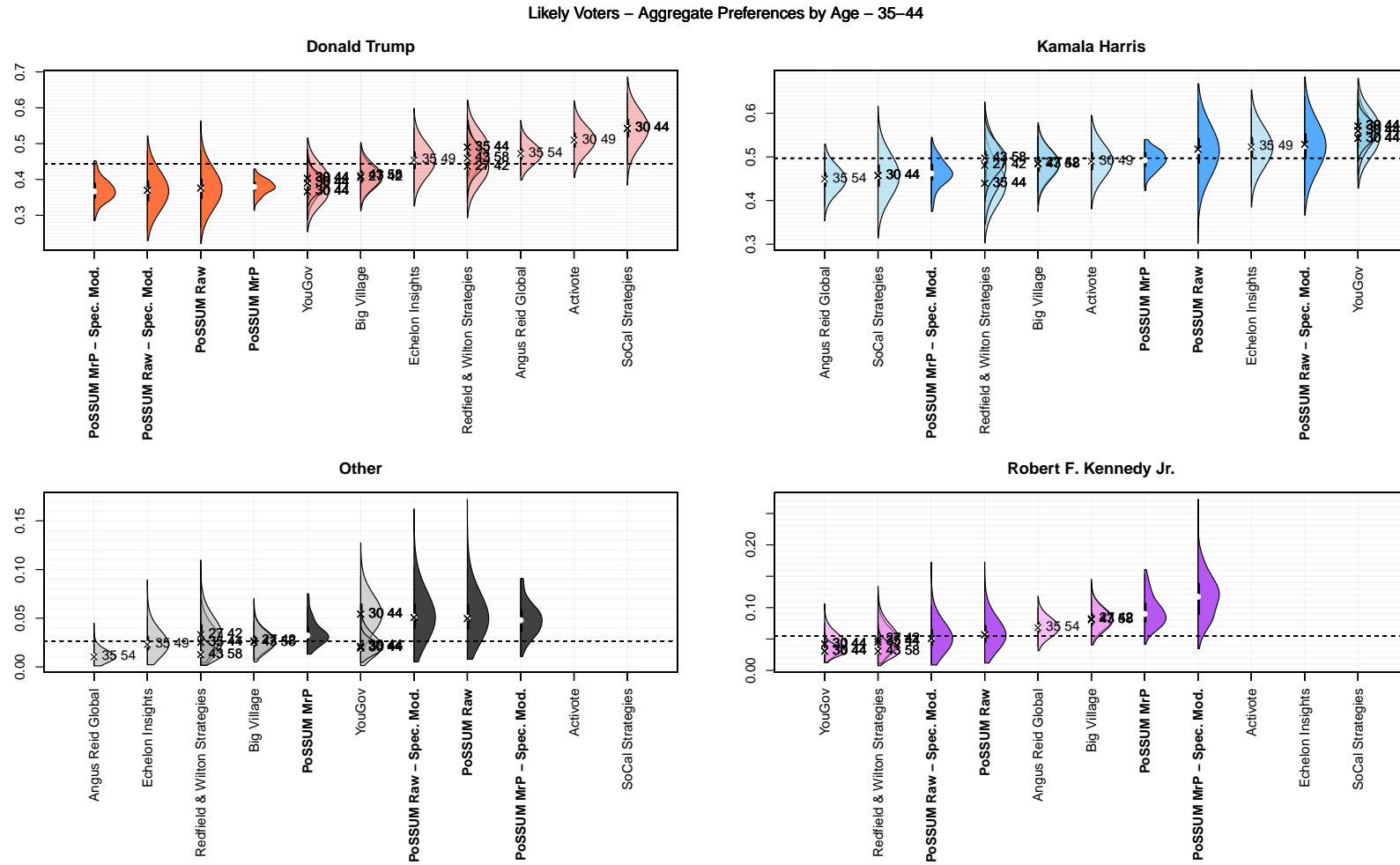


Figure 15: A comparison of the 35-44 years old distribution of 2024 vote choice across pollsters. The dotted line represents the simple arithmetic average of all polls. PoSSUM MrP is the primary estimate of interest. Large intervals for a given pollster correspond to small n for the given crosstab. To enable a comparison with multiple pollsters, and account for heterogeneity in crosstab granularity, I present the distribution of all age-groups which overlap with the 35-44 category.

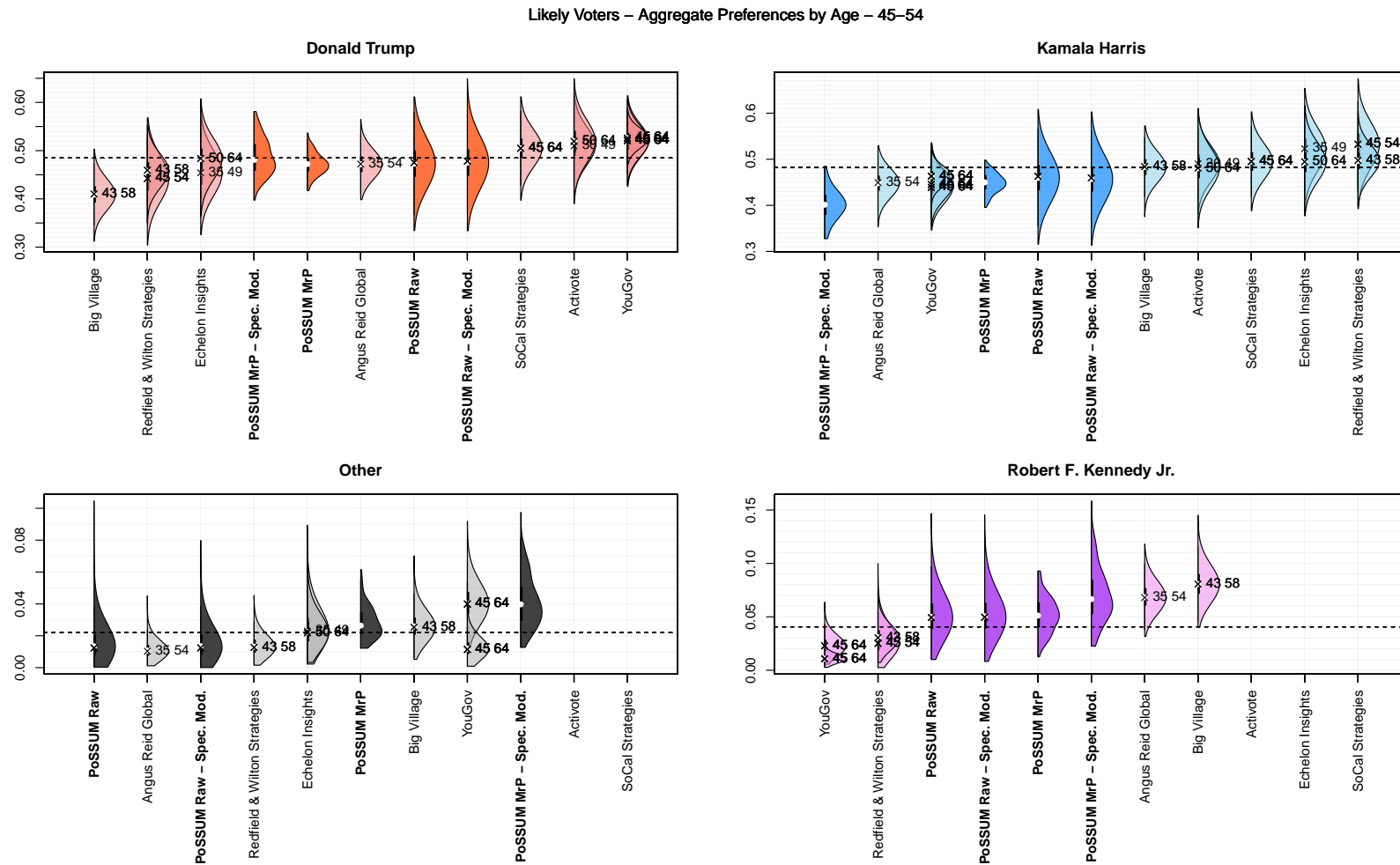


Figure 16: A comparison of the 45-54 years old distribution of 2024 vote choice across pollsters. The dotted line represents the simple arithmetic average of all polls. PoSSUM MrP is the primary estimate of interest. Large intervals for a given pollster correspond to small n for the given crosstab. To enable a comparison with multiple pollsters, and account for heterogeneity in crosstab granularity, I present the distribution of all age-groups which overlap with the 45-54 category.

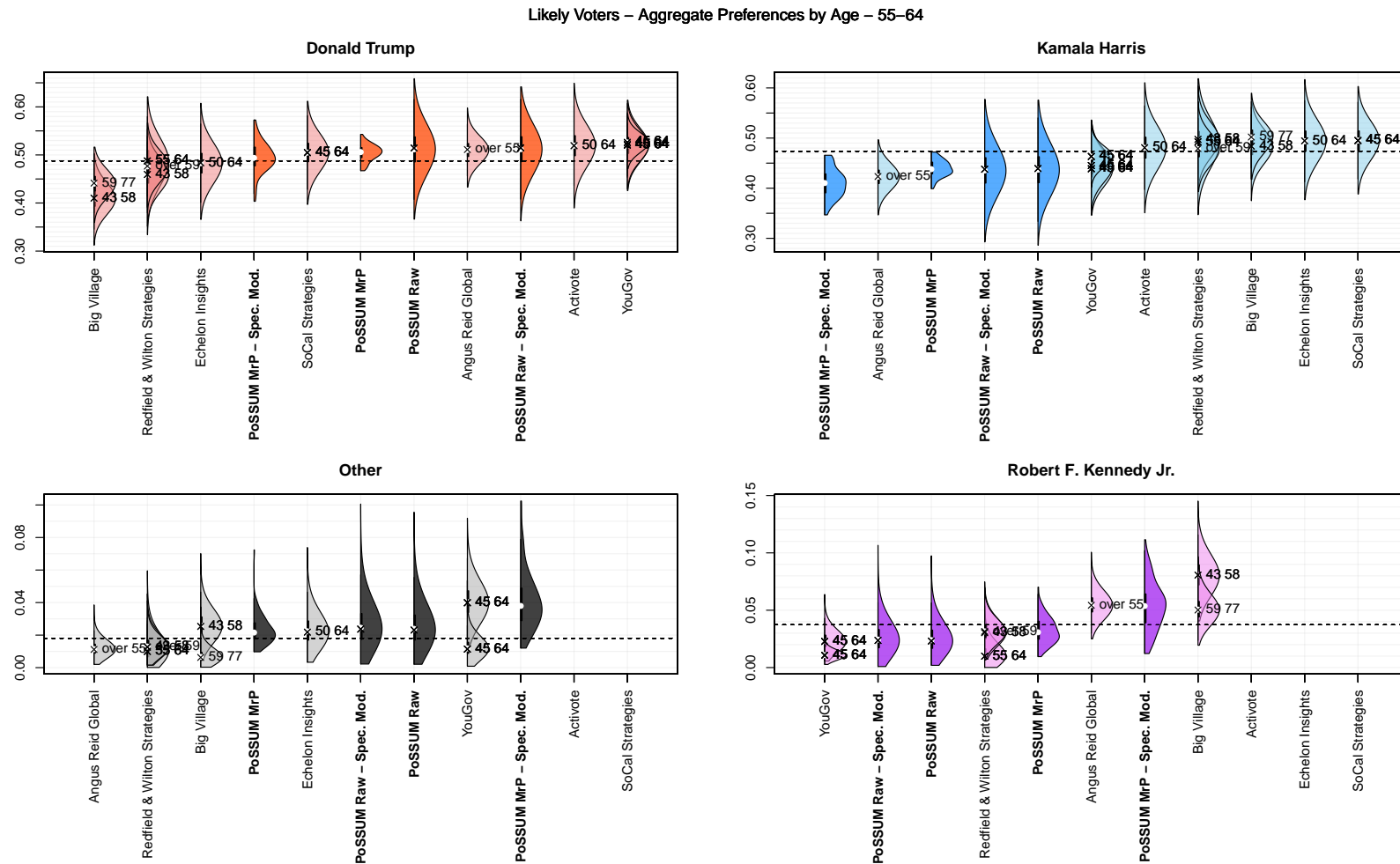


Figure 17: A comparison of the 55-64 years old distribution of 2024 vote choice across pollsters. The dotted line represents the simple arithmetic average of all polls. PoSSUM MrP is the primary estimate of interest. Large intervals for a given pollster correspond to small n for the given crosstab. To enable a comparison with multiple pollsters, and account for heterogeneity in crosstab granularity, I present the distribution of all age-groups which overlap with the 55-64 category.

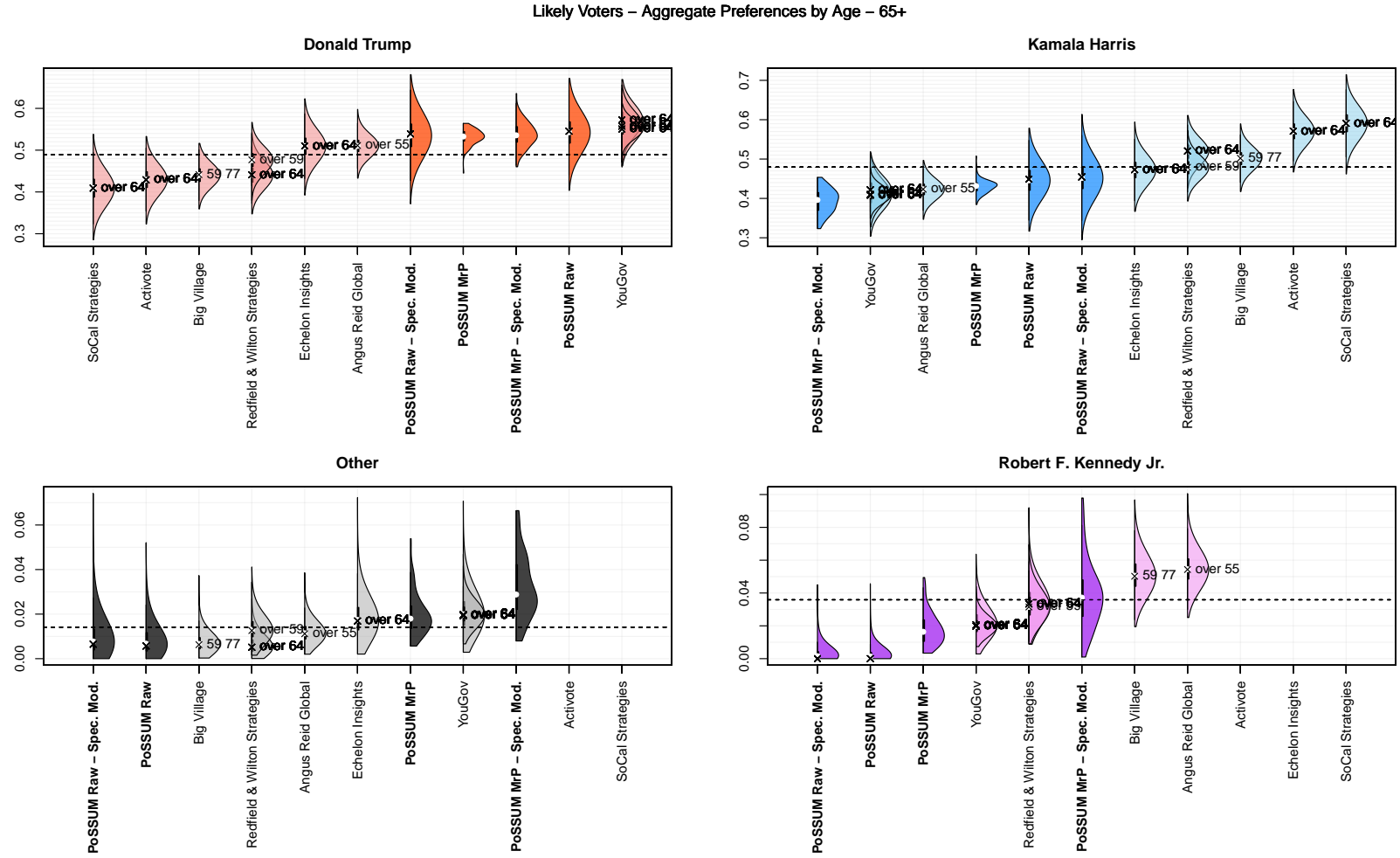


Figure 18: A comparison of the 65 years old and older distribution of 2024 vote choice across pollsters. The dotted line represents the simple arithmetic average of all polls. PoSSUM MrP is the primary estimate of interest. Large intervals for a given pollster correspond to small n for the given crosstab. To enable a comparison with multiple pollsters, and account for heterogeneity in crosstab granularity, I present the distribution of all age-groups which overlap with the 65-and-older category.

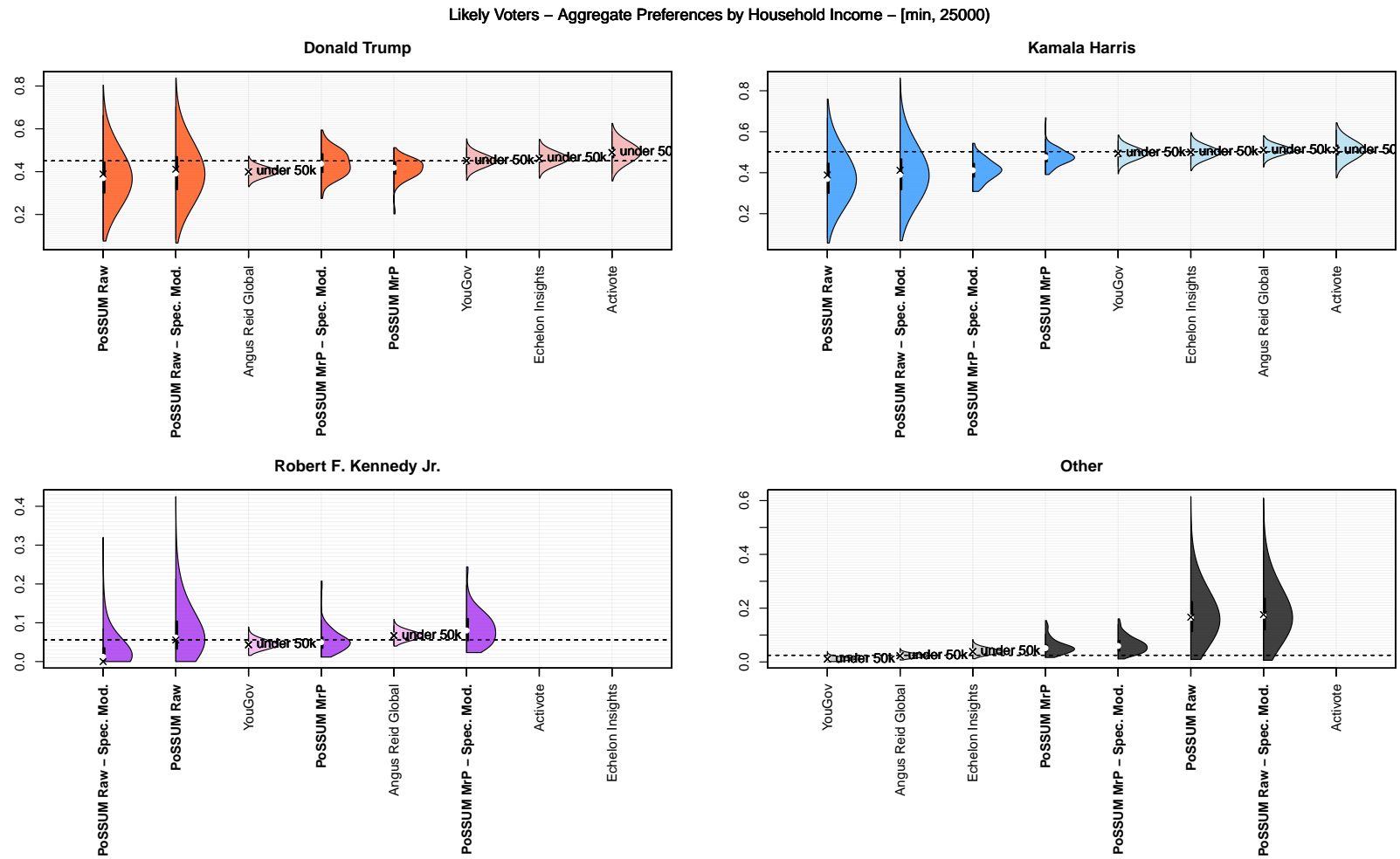


Figure 19: A comparison of the 2024 vote choice across pollsters, for voters with household income below 25k. The dotted line represents the simple arithmetic average of all polls. PoSSUM MrP is the primary estimate of interest. Large intervals for a given pollster correspond to small n for the given crosstab. To enable a comparison with multiple pollsters, and account for heterogeneity in crosstab granularity, I present the distribution of all income-groups which overlap with the under 25k category.

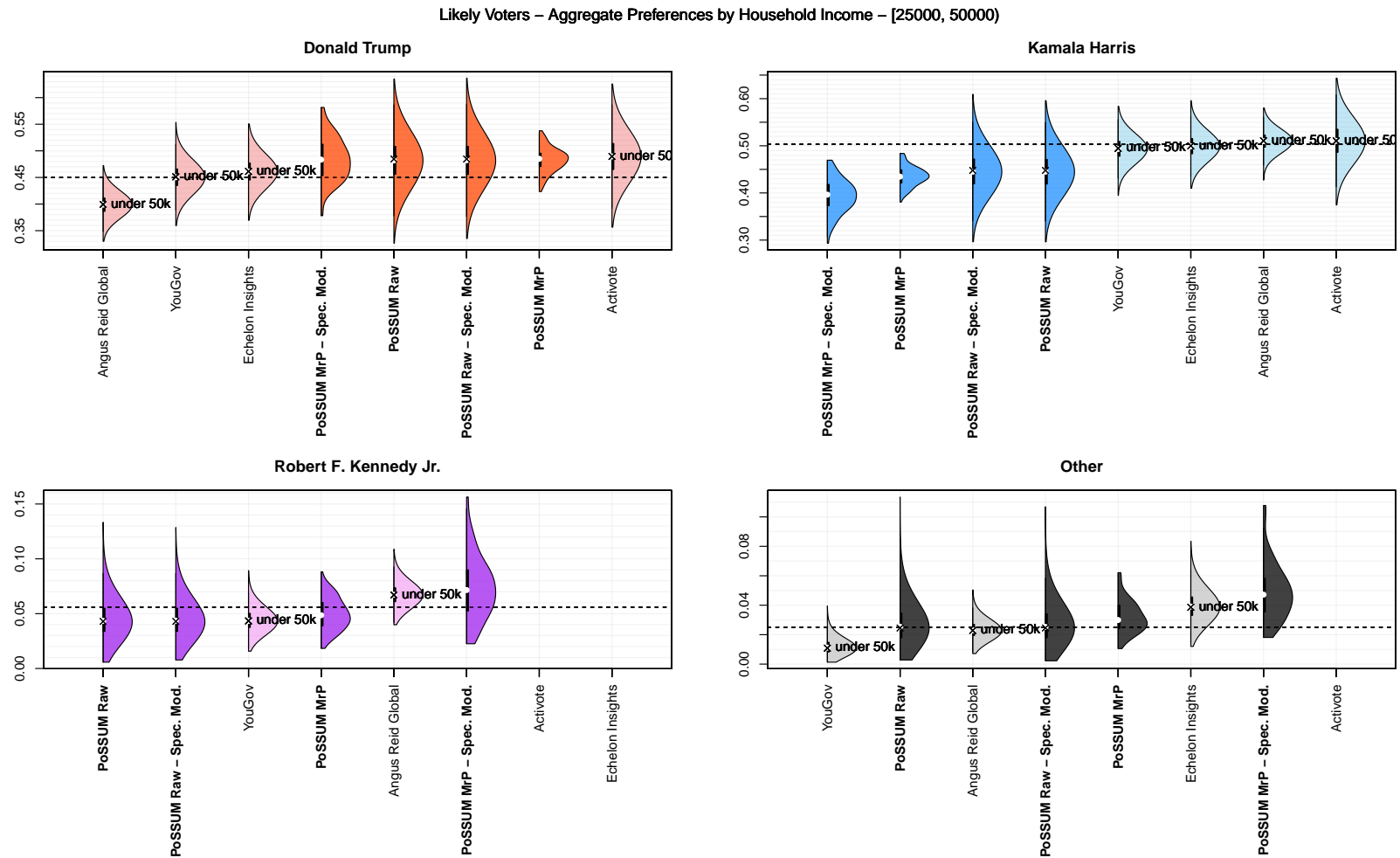


Figure 20: A comparison of the 2024 vote choice across pollsters, for voters with household income between 25k and 50k. The dotted line represents the simple arithmetic average of all polls. PoSSUM MrP is the primary estimate of interest. Large intervals for a given pollster correspond to small n for the given crosstab. To enable a comparison with multiple pollsters, and account for heterogeneity in crosstab granularity, I present the distribution of all income-groups which overlap with the between 25k and 50k category.

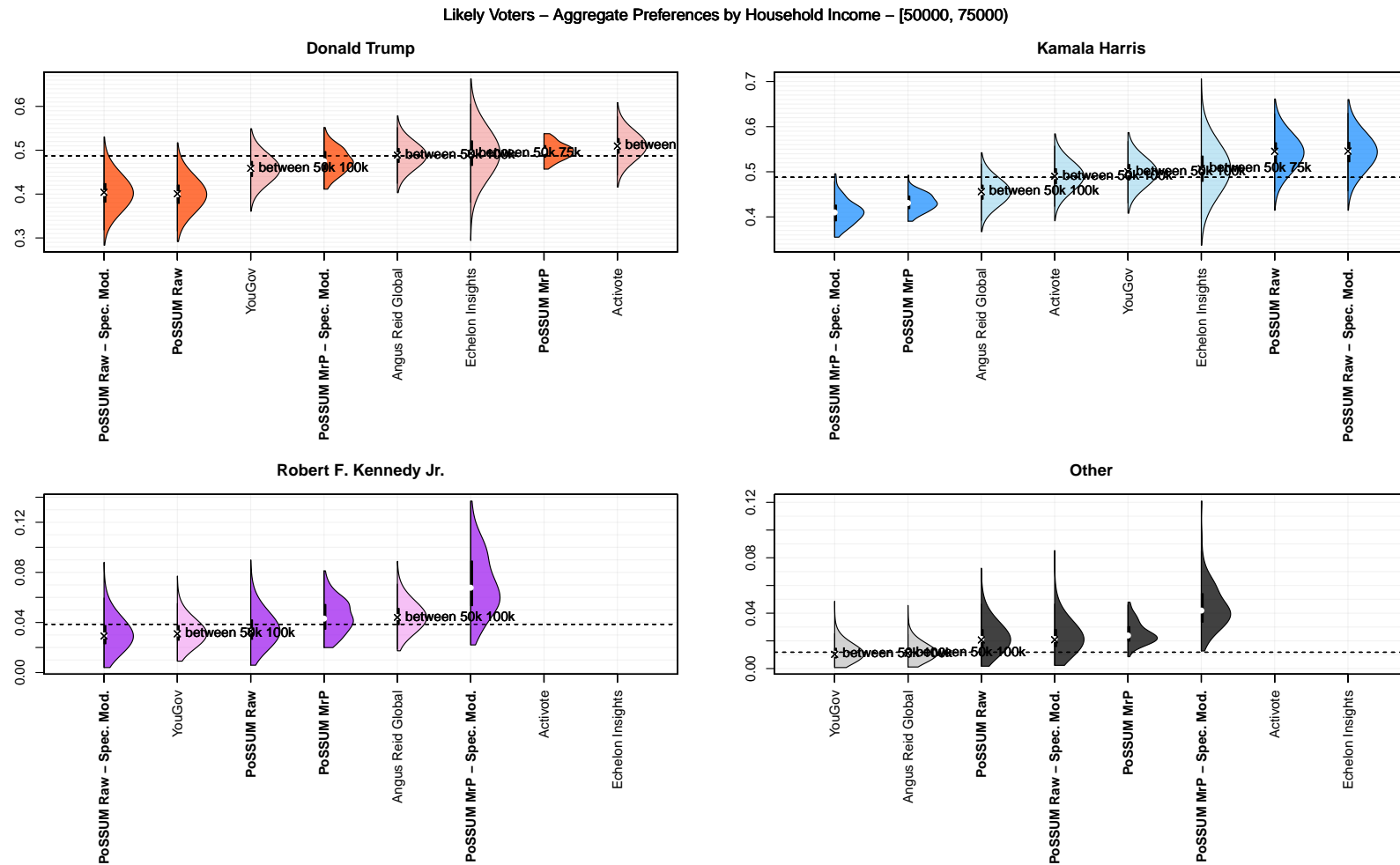


Figure 21: A comparison of the 2024 vote choice across pollsters, for voters with household income between 50k and 75k. The dotted line represents the simple arithmetic average of all polls. PoSSUM MrP is the primary estimate of interest. Large intervals for a given pollster correspond to small n for the given crosstab. To enable a comparison with multiple pollsters, and account for heterogeneity in crosstab granularity, I present the distribution of all income-groups which overlap with the between 50k and 75k category.

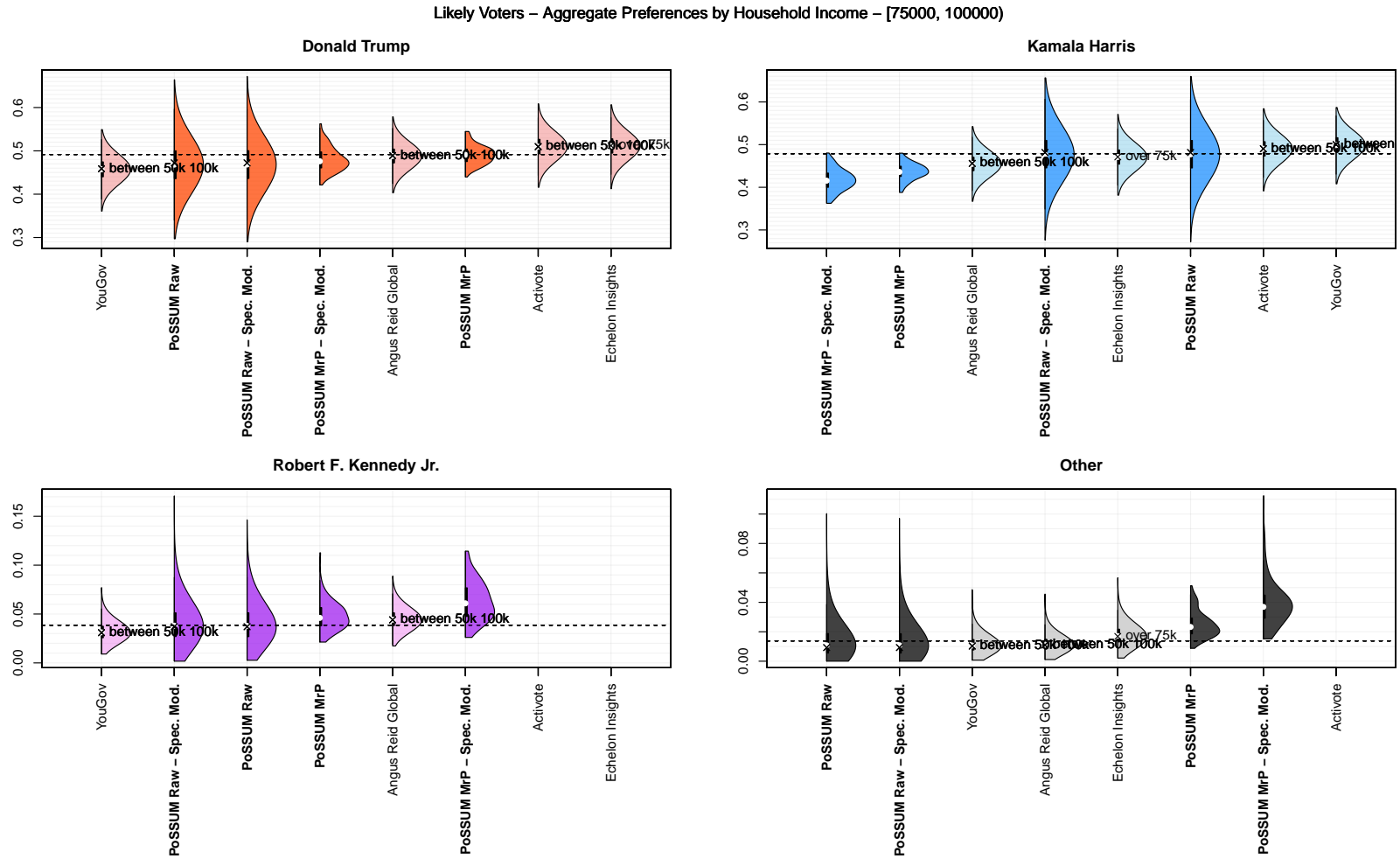


Figure 22: A comparison of the 2024 vote choice across pollsters, for voters with household income between 75k and 100k. The dotted line represents the simple arithmetic average of all polls. PoSSUM MrP is the primary estimate of interest. Large intervals for a given pollster correspond to small n for the given crosstab. To enable a comparison with multiple pollsters, and account for heterogeneity in crosstab granularity, I present the distribution of all income-groups which overlap with the between 75k and 100k category.

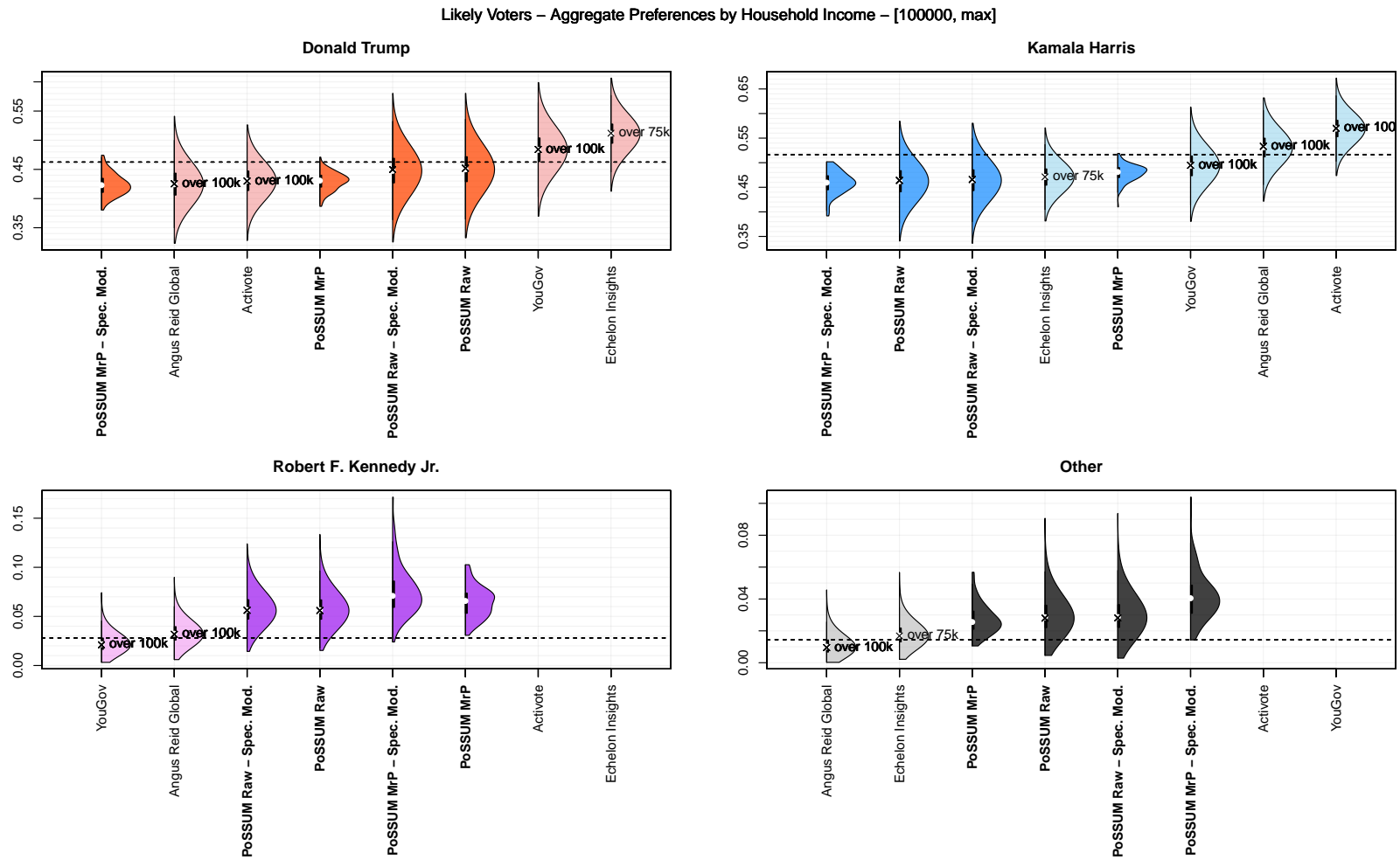


Figure 23: A comparison of the 2024 vote choice across pollsters, for voters with household income above 100k. The dotted line represents the simple arithmetic average of all polls. PoSSUM MrP is the primary estimate of interest. Large intervals for a given pollster correspond to small n for the given crosstab. To enable a comparison with multiple pollsters, and account for heterogeneity in crosstab granularity, I present the distribution of all income-groups which overlap with the over 100k category.

References

- [1] L. P. Argyle, E. C. Busby, N. Fulda, J. R. Gubler, C. Rytting, and D. Wingate. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351, 2023.
- [2] C. Barrie and J. C.-t. Ho. academictwitter: an r package to access the twitter academic research product track v2 api endpoint. *Journal of Open Source Software*, 6(62):3272, 2021.
- [3] J. Besag, J. York, and A. Mollié. Bayesian image restoration, with two applications in spatial statistics. *Annals of the institute of statistical mathematics*, 43(1):1–20, 1991.
- [4] B. Carpenter, A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell. Stan: A probabilistic programming language. *Journal of statistical software*, 76(1), 2017.
- [5] R. Cerina and R. Duch. Artificially intelligent opinion polling. *arXiv preprint arXiv:2309.06029*, 2023.
- [6] R. Choenni, E. Shutova, and R. van Rooij. Stepmothers are mean and academics are pretentious: What do pretrained language models learn about you? *arXiv preprint arXiv:2109.10052*, 2021.
- [7] C. Donegan. Flexible functions for icar, bym, and bym2 models in stan. *GitHub*, 2022. URL <https://github.com/ConnorDonegan/Stan-IAR>.
- [8] Y. Gao, L. Kennedy, D. Simpson, and A. Gelman. Improving multilevel regression and poststratification with structured priors. *Bayesian Analysis*, 16(3):719, 2021.
- [9] A. Gelman. Prior choice recommendations. *Stan Developer Wiki*, 2024. URL <https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations>.
- [10] A. Gelman and T. C. Little. Poststratification into many categories using hierarchical logistic regression. 1997.
- [11] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 2013.
- [12] Y. Ghitza and A. Gelman. Deep interactions with mrp: Election turnout and voting patterns among small electoral subgroups. *American Journal of Political Science*, 57(3):762–776, 2013.
- [13] M. Goplerud. Re-evaluating machine learning for mrp given the comparable performance of (deep) hierarchical models. *American Political Science Review*, 118(1):529–536, 2024.

- [14] J. Gottfried. Americans’ social media use, January 31 2024. URL <https://www.pewresearch.org/internet/2024/01/31/americans-social-media-use/>. Accessed: 2024-08-08.
- [15] M. W. Kearney. rtweet: Collecting and analyzing twitter data. *Journal of open source software*, 4(42):1829, 2019.
- [16] G. King and L. Zeng. Logistic regression in rare events data. *Political analysis*, 9(2):137–163, 2001.
- [17] B. E. Lauderdale, D. Bailey, J. Blumenau, and D. Rivers. Model-based pre-election polling for national and sub-national outcomes in the us and uk. *International Journal of Forecasting*, 36(2):399–413, 2020.
- [18] Y. LeCun. Do large language models need sensory grounding for meaning and understanding? spoiler: Yes!, 2023. URL https://drive.google.com/file/d/1BU5bV3X5w65DwSMapKcsr0ZvrMRU_Nbi/view.
- [19] L. Leemann and F. Wasserfallen. Extending the use and prediction precision of subnational public opinion estimation. *American journal of political science*, 61(4):1003–1022, 2017.
- [20] G. E. Morris. Artificial intelligence and ”big data” cannot replace public opinion polls. *Elliott’s notebook*, 2024. URL <https://gelliottmorris.substack.com/p/artificial-intelligence-and-big-data>.
- [21] M. Morris. Spatial models in stan: Intrinsic auto-regressive models for areal data. *GitHub repository*, 2018.
- [22] OpenAI. Gpt-4 technical report, 2023.
- [23] OpenAI. Introducing gpt-4o, 2024. URL <https://openai.com/index/hello-gpt-4o/>.
- [24] D. K. Park, A. Gelman, and J. Bafumi. Bayesian multilevel estimation with post-stratification: State-level estimates from national polls. *Political Analysis*, 12(4):375–385, 2004.
- [25] J. Pasek and M. J. Pasek. Package ‘anesrake’. 2018.
- [26] I. Rudnytskyi. *openai: R Wrapper for OpenAI API*, 2023. URL <https://github.com/irudnyts/openai>. R package version 0.4.1, <https://irudnyts.github.io/openai/>.
- [27] B. Schaffner, S. Ansolabehere, and M. Shih. Cooperative Election Study Common Content, 2022, 2023. URL <https://doi.org/10.7910/DVN/PR4L8P>.
- [28] J. Twyman. Getting it right: Yougov and online survey research in britain. *Journal of Elections, Public Opinion and Parties*, 18(4):343–354, 2008.

- [29] U.S. Census Bureau. American Community Survey, 2021 American Community Survey 5-Year Estimates. U.S. Census Bureau, American Community Survey (ACS), 2021. URL <https://www.census.gov/programs-surveys/acs>. Accessed: 2024-08-27.
- [30] S. Van Buuren. *Flexible imputation of missing data*. CRC press, 2018.