PoSSUM: A **Pro**tocol for **S**urveying **S**ocial-media **U**sers with **M**ultimodal LLMs

`Version 1.0.0 alpha`

Roberto Cerina

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

September 28, 2024

Contents

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 source[1] 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 online 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 15^{th} and November 5^{th}, 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

[2] Silicon 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:

- $\boldsymbol{q}$: optimal API search queries
- $\boldsymbol{w}$: weight of each query

Output:

- $\boldsymbol{\Upsilon}$: user-data object composed of profile info $\boldsymbol{\upsilon}$ and tweets $\boldsymbol{\mathcal{T}}$

Routine *get_pool*:

```
K ← length(q) ;                                # Get: number of queries
ϒ ← ∅ ;                                        # Initialize: empty users object
for k = 1 to K do
    (𝒯, υ)_kt ← 𝕏(q_k, w_k) ;                  # Call: sample of tweet-user pairs
    ϒ ← ϒ ∪ (𝒯, υ)_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 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 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 day[3]. I therefore propose to use a combination of search queries for the `tweets/search/recent` endpoint[4], 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 functions[5] reminiscent of the now-defunct `rtweet` [15] and `academictwitteR` [2], in order to specify an appropriate set of queries.

[3] I use the *basic* tier 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.

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

[5] Available 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 parties[6]. I assign a *weight* (number of tweets extracted) to this query of size w.

[6]On 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.

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

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 $\boldsymbol{q}$, which is an object of size $K = L + 1$, and a corresponding set of weights $\boldsymbol{w} = \{n, \frac{n}{L}, ..., \frac{n}{L}\}$.

We execute each of these queries in a loop, and for each we obtain a tweet-user object $(\boldsymbol{\mathcal{T}}, \boldsymbol{\upsilon})_{kt}$ 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:

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

Output:

- $\boldsymbol{Z}$: survey object with extracted features

Routine ***poll_users*****:**

```
Υ* ← Υ[τ = TRUE ∨ G ≠ ∅] ;        # Filter: recent + valid location
N ← length(Υ*) ;                    # Get: number of valid users
Z ← ∅ ;                             # Initialize: empty survey object

for i = 1 to N do
    e_i ← GPT{P^E(Υ_i)} ;           # Call: GPT entity inclusion

    if e_i ∈ E then
        g_i ← GPT{P^G(Υ_i)} ;       # Call: GPT geographic inclusion

        if g_i ∈ G then
            X_i ← GPT{P^φ(Υ_i, F^x)} ;     # Call: GPT quota inclusion

            if X_i ∈ X^Q ∪ ω'_i < ω*_i then
                T^+_i ← 𝕏(Υ_i) ;            # Call: sample last m tweets
                z_i ← GPT{P^φ(Υ_i, T^+_i, F)} ;   # Call: GPT extraction

                Z ← Z ∪ z_i;                # Store: survey object

                ω'_i ← ω'_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* 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, $\mathcal{M}(\boldsymbol{\Upsilon})$, 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, $O(\boldsymbol{\mathcal{F}})$, 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. Pesudo-code 3 presents the steps involved in the prompt-building routine.

Input:

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

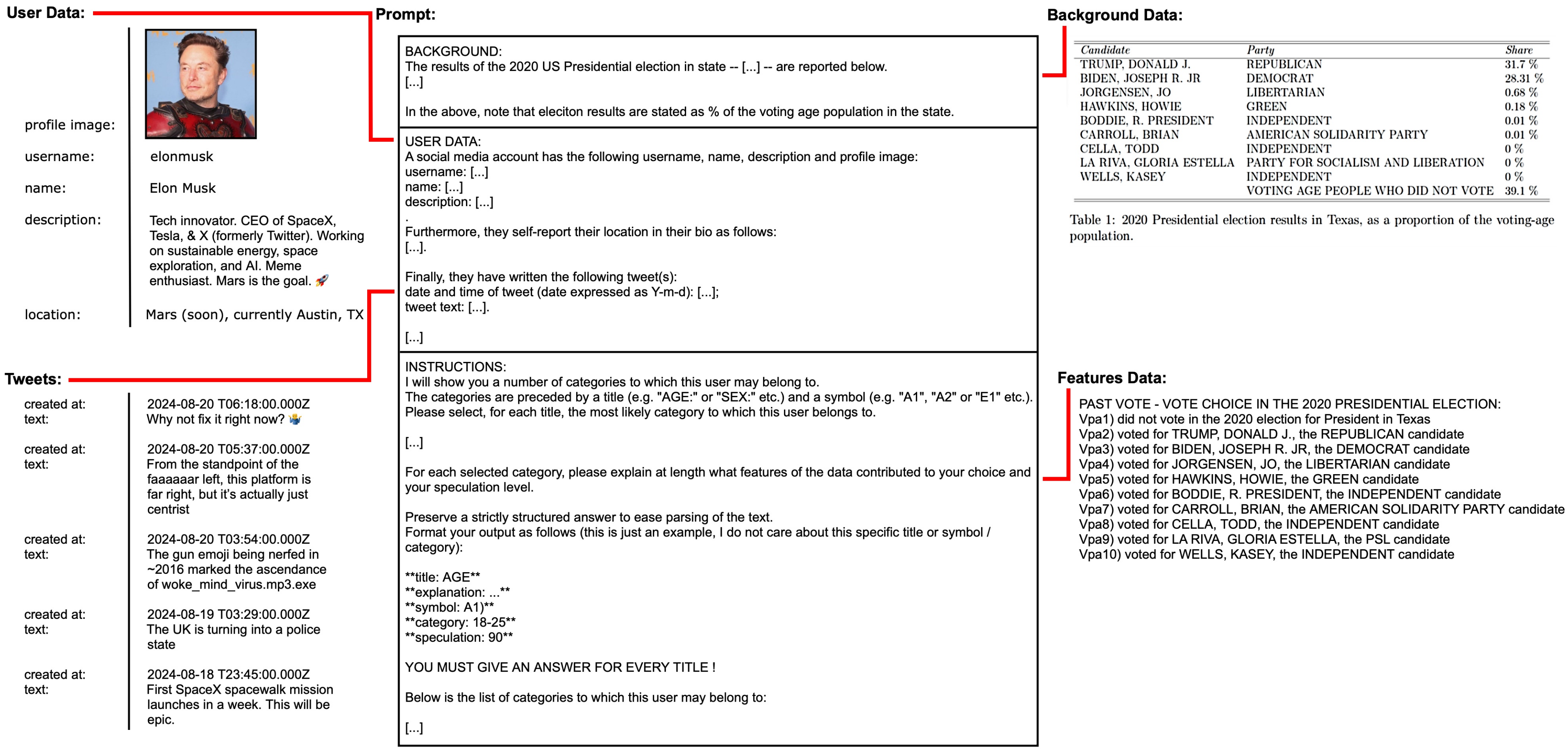
Output:

- $\mathcal{P}$: unified prompt

Routine *build_prompt***:**

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

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



Candidate	*Party*	*Share*
TRUMP, DONALD J.	REPUBLICAN	31.7 %
BIDEN, JOSEPH R. JR	DEMOCRAT	28.31 %
JORGENSEN, JO	LIBERTARIAN	0.68 %
HAWKINS, HOWIE	GREEN	0.18 %
BODDIE, R. PRESIDENT	INDEPENDENT	0.01 %
CARROLL, BRIAN	AMERICAN SOLIDARITY PARTY	0.01 %
CELLA, TODD	INDEPENDENT	0 %
LA RIVA, GLORIA ESTELLA	PARTY FOR SOCIALISM AND LIBERATION	0 %
WELLS, KASEY	INDEPENDENT	0 %
	VOTING AGE PEOPLE WHO DID NOT VOTE	39.1 %

Table 1: 2020 Presidential election results in Texas, as a proportion of the voting-age population.

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 – $\boldsymbol{\tau}$: 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:

- $\boldsymbol{\tau}$: temporal inclusion criteria
- Υ: user data

Output:

- $\Upsilon^{\star}$: filtered user data including only recent users

Routine *temporal_inclusion*:

$\Upsilon^{\star} \leftarrow \Upsilon[\boldsymbol{\tau}(\boldsymbol{t}) = \texttt{TRUE}]$; `# Filter: exclude users based on criteria`

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

Null Geography Exclusion Criteria – $\boldsymbol{\mathcal{G}} \neq \emptyset$: 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
- $\boldsymbol{\mathcal{G}}$: geographic information associated with users

Output:
- $\Upsilon^{\star}$: filtered user data with non-null geographic information

Routine ***exclude_null_geography*:**

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

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

Entity Inclusion Criteria – $\boldsymbol{E}$: 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.

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

Geographic Inclusion Criteria – $\boldsymbol{\mathcal{G}}$: 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.

```
Which state of the USA do they live in?
If they do not specify a state, but are still from the United States, write "USA".
If they are not from a state in the USA, write "Not from a state in the USA".
Write out just the full name of the state.
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 $\omega^{\star}_{c}$ from each "cell" $c \in \{1, ..., C\}$, which we would expect to capture in a random sample of target size $\Omega^{\star}$ users – we could for instance set $\Omega^{\star} = 1,500$ 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 $\omega^{\star}_{c}$, I retain the user and update the quota counter – otherwise I exclude the user from the analysis. Pseudocode 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
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
430	female	25 to 34	between 25k and 50k	asian	stayed home	1	0
431	female	65 or older	between 50k and 75k	hispanic	stayed home	1	0
432	female	18 to 24	more than 100k	asian	stayed home	1	0
433	male	18 to 24	between 50k and 75000	native	stayed home	1	0
434	female	55 to 64	between 50k and 75k	asian	stayed home	1	0
435	male	18 to 24	between 50k and 75k	asian	stayed home	1	0

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

2.2.3 General Feature Extraction Framework

Implementing the quota inclusion criteria requires a defined set of features of interest $\boldsymbol{\mathcal{F}}$. 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 $\boldsymbol{\mathcal{F}}$ 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 $O(\boldsymbol{\mathcal{F}})$ 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*. A classic critique of silicon samples is that the data generating process of the LLM is ultimately unknown. More crucially for `PoSSUM`, it is uncomfortable to be in the dark as to how much of the LLM's "own" knowledge, which it has acquired during its training phase, is responsible for a given estimate, and how much is just evident in the $\mathbb{X}$ 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.

To address this concern 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 (e.g. the text of the tweets or the pixels of the profile image) which is directly useful to the imputation task, and distinguishes this from other kinds of knowledge the LLM might leverage. The score has a categorical interpretation, which identifies "highly speculative" imputations at $S > 80$. I will present a demonstration of how to use this score to conduct a sensitivity analysis in Section 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.

```
ind.features <- c(
ETHNICITY:
E1) white - individuals with origins in any of the original peoples of europe...
E2) black or african american - individuals with origins in any of the black racial ...
E3) hispanic or latino - includes individuals of mexican, puerto rican, salvadoran, cuban...
E4) asian - individuals with origins in any of the original peoples of central or east ...
E5) american indian or alaskan native or native hawaiian or pacific islander - ...
E6) multiracial - individuals who identify explicitly as belonging to more than one of ...

',
  'AGE:
A1) under 18 years old
A2) 18 to 24 years old
A3) 25 to 34 years old
A4) 35 to 44 years old
A5) 45 to 54 years old
A6) 55 to 64 years old
A7) 65 or older

',
  'SEX:
S1) masculine sex - male
S2) feminine sex - female

',
  'INTEREST IN POLITICS:
I1) not interested at all in politics
I2) slightly interested in politics
I3) moderately interested in politics
I4) highly interested in politics

',
  'MARITAL STATUS:
M1) married - currently legally married and living with a spouse
M2) single - never married, including those who are legally separated
M3) divorced - legally divorced and not currently remarried
M4) widowed - spouse has passed away and not currently remarried

',
  "HIGHEST EDUCATIONAL QUALIFICATION:
Q1) completed education up to and including high school - high school diploma, ...
Q2) completed education at the college or university level - bachelor's degree, ...

",
  'HOUSEHOLD INCOME BRACKET:
H1) up to 25000 USD per year
H2) between 25000 and 50000 USD per year
H3) between 50000 and 75000 USD per year
H4) between 75000 and 100000 USD per year
H5) more than 100000 USD per year

',
  'GENERAL TRUST IN OTHER PEOPLE:
Tru1) always trust other people
Tru2) most of the time trust other people
Tru3) about half of the time trust other people
Tru4) some of the time trust other people
Tru5) never trust other people

',
  'PAYING ATTENTION TO THE 2024 PRESIDENTIAL ELECTION:
Att1) not paying attention at all to the 2024 Presidential election in the US
```

```
Att2) paying only a little attention to the 2024 Presidential election in the US
Att3) paying some attention to the 2024 Presidential election in the US
Att4) paying a lot of attention to the 2024 Presidential election in the US

',
  'PARTISAN LOYALTIES:
Pid1) strongly identifies with Democrats
Pid2) weakly identifies with Democrats
Pid3) independent closer to the Democrats
Pid4) independent not closer to either party
Pid5) independent closer to the Republicans
Pid6) weakly identifies with Republicans
Pid7) strongly identifies with Republicans

',
  "IDEOLOGICALLY, THIS PERSON APPEARS TO BE:
Ide2) very ideologically liberal
Ide3) somewhat ideologically liberal
Ide4) moderate in ideological orientation
Ide5) somewhat ideologically conservative
Ide6) very ideologically conservative

",
  'PAST VOTE - TURNOUT IN THE 2020 PRESIDENTIAL ELECTION:
Tpa1) no chance this individual turned out to vote - Probability: 0 ...
Tpa2) highly unlikely this individual turned out to vote - Probability: 0.15 ...
Tpa3) unlikely this individual turned out to vote - Probability: 0.3 ...
Tpa4) 50-50 likelihood that this individual voted - Probability: 0.5 ...
Tpa5) likely this individual turned out to vote - Probability: 0.7 ...
Tpa6) highly likely this individual turned out to vote - Probability: 0.85 ...
Tpa7) certain this individual turned out to vote - Probability: 1 ...

',
  'PAST VOTE - TURNOUT IN THE 2022 HOUSE OF REPRESENTATIVES ELECTION:
Thpa1) no chance this individual turned out to vote - Probability: 0 ...
Thpa2) highly unlikely this individual turned out to vote - Probability: 0.15 ...
Thpa3) unlikely this individual turned out to vote - Probability: 0.3 ...
Thpa4) 50-50 likelihood that this individual voted - Probability: 0.5 ...
Thpa5) likely this individual turned out to vote - Probability: 0.7 ...
Thpa6) highly likely this individual turned out to vote - Probability: 0.85 ...
Thpa7) certain this individual turned out to vote - Probability: 1 ...

',
  'PAST VOTE - VOTE CHOICE IN THE 2020 PRESIDENTIAL ELECTION:
Vpa1) did not vote in the 2020 election for President in their state
Vpa2) voted for Donald Trump, the Republican Party candidate, ...
Vpa3) voted for Joe Biden, the Democratic Party candidate, ...
Vpa4) voted for Jo Jorgensen, the Libertarian Party candidate, ...
Vpa5) voted for Howie Hawkins, the Green Party candidate, ...
Vpa6) voted for a candidate other than the Republican, Democrat, Libertarian, or Green ...

',
  'PAST VOTE - VOTE CHOICE IN THE 2022 HOUSE OF REPRESENTATIVES ELECTION:
Vhpa1) did not vote in the 2022 elections for the House of Representatives in their ...
Vhpa2) voted for a Republican Party candidate, ...
Vhpa3) voted for a Democratic Party candidate, ...
Vhpa4) voted for a Libertarian Party candidate, ...
Vhpa5) voted for a Green Party candidate, ...
Vhpa6) voted for a candidate other than the Republican, Democrat, Libertarian, or Green ...

')
}
```

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

```
tsk <-
      paste0(
      'I will show you a number of categories to which this user may belong to.
The categories are preceded by a title (e.g. "AGE:" or "SEX:" etc.) and a symbol (e.g. "A1", "A2" or "E1" etc.).
Please select, for each title, the most likely category to which this user belongs to.

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

For each selected symbol / category, please note the level of Speculation involved in this selection.
Present the Speculation level for each selection on a scale from 0 (not speculative at all, every single element of the user data was useful in the selection) to 100 (fully speculative, there is no information related to this title in the user data).
Speculation levels should be a direct measure of the amount of useful information available in the user data.
Speculation levels pertain only to the information available in the user data -- namely the username, name, description, location, profile picture and tweets from this user -- and should not be affected by additional information available to you from any other source.
To ensure consistency, use the following guidelines to determine speculation levels:

0-20 (Low speculation): The user data provides clear and direct information relevant to the title. (e.g., explicit mention in the profile or tweets)
21-40 (Moderate-low speculation): The user data provides indirect but strong indicators relevant to the title. (e.g., context from multiple sources within the profile or tweets)
41-60 (Moderate speculation): The user data provides some hints or partial information relevant to the title. (e.g., inferred from user interests or indirect references)
61-80 (Moderate-high speculation): The user data provides limited and weak indicators relevant to the title. (e.g., very subtle hints or minimal context)
81-100 (High speculation): The user data provides no or almost no information relevant to the title. (e.g., assumptions based on very general information)

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

Preserve a strictly structured answer to ease parsing of the text.
Format your output as follows (this is just an example, I do not care about this specific title or symbol / category):

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

YOU MUST GIVE AN ANSWER FOR EVERY TITLE !

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

',
      # randomise feature order
      paste0(
        sample(ind.features),
        collapse = '\n'
        )
      )
```

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: $m^{tredning} = \lambda \times m^{politics}$, $\forall \lambda > 1$. 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.

```
dep.features <- c(
  "UNDECIDEDNESS AROUND 2024 US PRESIDENTIAL ELECTION VOTE CHOICE - ...
Und1) no chance this individual will change their mind - Probability: 0 - ...
Und2) highly unlikely this individual will change their mind - Probability: 0.15 - ...
Und3) unlikely this individual will change their mind - Probability: 0.3 - ...
Und4) 50-50 likelihood that this individual will change their mind - Probability: 0.5 - ...
Und5) likely this individual will change their mind - Probability: 0.7 - ...
Und6) highly likely this individual will change their mind - Probability: 0.85 - ...
Und7) certain this individual will change their mind - Probability: 1 - ...

",
  "JOE BIDEN PRESIDENTIAL JOB APPROVAL:
Bap1) Strongly Approves of the way Joe Biden is handling his job as president
Bap2) Somewhat Approves of the way Joe Biden is handling his job as president
Bap3) Somewhat Disapproves of the way Joe Biden is handling his job as president
Bap4) Strongly Disapproves of the way Joe Biden is handling his job as president

",
  "FAVOURABILITY OF US POLITICIAN JOE BIDEN:
Bfa1) Very favourable view of Joe Biden
Bfa2) Somewhat favourable view of Joe Biden
Bfa3) Somewhat unfavourable view of Joe Biden
Bfa4) Very unfavourable view of Joe Biden
Bfa5) This person is unlikely to know who Joe Biden is, so they cannot have an opinion ...

",
  "FAVOURABILITY OF US POLITICIAN KAMALA HARRIS:
Kfa1) Very favourable view of Kamala Harris
Kfa2) Somewhat favourable view of Kamala Harris
Kfa3) Somewhat unfavourable view of Kamala Harris
Kfa4) Very unfavourable view of Kamala Harris
Kfa5) This person is unlikely to know who Kamala Harris is, so they cannot have an ...

",
  "FAVOURABILITY OF US POLITICIAN TIM WALZ:
Twa1) Very favourable view of Tim Walz
Twa2) Somewhat favourable view of Tim Walz
Twa3) Somewhat unfavourable view of Tim Walz
Twa4) Very unfavourable view of Tim Walz
Twa5) This person is unlikely to know who Tim Walz is, so they cannot have an opinion  ...

",
  "FAVOURABILITY OF US POLITICIAN DONALD TRUMP:
Tfa1) Very favourable view of Donald Trump
Tfa2) Somewhat favourable view of Donald Trump
Tfa3) Somewhat unfavourable view of Donald Trump
Tfa4) Very unfavourable view of Donald Trump
Tfa5) This person is unlikely to know who Donald Trump is, so they cannot have an ...

",
  "FAVOURABILITY OF US POLITICIAN JD VANCE:
Jfa1) Very favourable view of JD Vance
Jfa2) Somewhat favourable view of JD Vance
Jfa3) Somewhat unfavourable view of JD Vance
Jfa4) Very unfavourable view of JD Vance
Jfa5) This person is unlikely to know who JD Vance is, so they cannot have an opinion ...

",
  'FAVOURABILITY OF US POLITICIAN ROBERT F. KENNEDY JR.:
Rfa1) Very favourable view of Robert F. Kennedy Jr.
Rfa2) Somewhat favourable view of Robert F. Kennedy Jr.
Rfa3) Somewhat unfavourable view of Robert F. Kennedy Jr.
```

```
Rfa4) Very unfavourable view of Robert F. Kennedy Jr.
Rfa5) This person is unlikely to know who Robert F. Kennedy Jr. is, so they cannot have ...

',
  'FAVOURABILITY OF US POLITICIAN CORNEL WEST:
Wfa1) Very favourable view of Cornel West
Wfa2) Somewhat favourable view of Cornel West
Wfa3) Somewhat unfavourable view of Cornel West
Wfa4) Very unfavourable view of Cornel West
Wfa5) This person is unlikely to know who Cornel West is, so they cannot have an opinion ...

',
  'FAVOURABILITY OF US POLITICIAN JILL STEIN:
Sfa1) Very favourable view of Jill Stein
Sfa2) Somewhat favourable view of Jill Stein
Sfa3) Somewhat unfavourable view of Jill Stein
Sfa4) Very unfavourable view of Jill Stein
Sfa5) This person is unlikely to know who Jill Stein is, so they cannot have an opinion ...

',
  'FAVOURABILITY OF US POLITICIAN CHASE OLIVER:
Ofa1) Very favourable view of Chase Oliver
Ofa2) Somewhat favourable view of Chase Oliver
Ofa3) Somewhat unfavourable view of Chase Oliver
Ofa4) Very unfavourable view of Chase Oliver
Ofa5) This person is unlikely to know who Chase Oliver is, so they cannot have an ...

',
  'BELIEF IN MOST IMPORTANT ISSUE TODAY:
Moi01) jobs and the economy are the most important issue for this user today
Moi02) immigration is the most important issue for this user today
Moi03) climate change and the environment are the most important issue for this user today
Moi04) foreign policy is the most important issue for this user today
Moi05) national security is the most important issue for this user today
Moi06) education is the most important issue for this user today
Moi07) healthcare is the most important issue for this user today
Moi08) taxes and government spending are the most important issue for this user today
Moi09) abortion is the most important issue for this user today
Moi10) civil rights and racism are the most important issue for this user today
Moi11) guns are the most important issue for this user today
Moi12) crime is the most important issue for this user today
Moi13) criminal justice reform and over-incarceration are the most important issue for ...
Moi14) inflation and the cost of living is the most important issue for this user today

',
  'HAPPINESS LEVEL:
Hap1) very happy - consistently feels a high level of joy and satisfaction with ...
Hap2) rather happy - generally feels happy and satisfied with their life - ...
Hap3) not very happy - experiences some level of dissatisfaction or lack of joy in their ...
Hap4) not happy at all - consistently feels unhappy and dissatisfied with their life ...

',

  'BIG 5 PERSONALITY TRAIT -- OPENNESS TO EXPERIENCE:
Ope1) very open to experience - highly curious, imaginative, and open to exploring ...
Ope2) somewhat open to experience - open to new experiences and ideas, but they may ...
Ope3) neither open nor closed to experience - shows a balanced approach, being neither ...
Ope4) somewhat closed to experience - tends to prefer familiar routines and may be ...
Ope5) very closed to experience - prefers a highly structured and predictable ...

',
  'BIG 5 PERSONALITY TRAIT -- CONSCIENTIOUSNESS:
Con1) very conscientious - highly organized, responsible, and dependable - they ...
Con2) somewhat conscientious - generally reliable and organized but may occasionally ...
Con3) neither conscientious nor unconscientious - demonstrates an average level of ...
```

```
Con4) somewhat unconscientious - tends to be less organized and may struggle with ...
Con5) very unconscientious - often lacks organization and discipline - they may be ...

',
  'BIG 5 PERSONALITY TRAIT -- EXTRAVERSION:
Ext1) very extraverted - highly outgoing, energetic, and enjoys social interactions - ...
Ext2) somewhat extraverted - enjoys social interactions and can be outgoing, but ...
Ext3) neither extraverted nor introverted - displays a balanced mix of extraverted ...
Ext4) somewhat introverted - tends to be more reserved and prefers quieter, less ...
Ext5) very introverted - highly reserved, enjoys solitude, and prefers minimal social ...

',
  "BIG 5 PERSONALITY TRAIT -- AGREEABLENESS:
Agr1) very agreeable - highly cooperative, compassionate, and trusting - they ...
Agr2) somewhat agreeable - generally kind and cooperative but may occasionally ...
Agr3) neither agreeable nor disagreeable - generally kind and cooperative but ...
Agr4) somewhat disagreeable - tends to be more competitive and less concerned with ...
Agr5) very disagreeable - often uncooperative and less empathetic - they prioritize ...

",
  'BIG 5 PERSONALITY TRAIT -- NEUROTICISM:
Neu1) very high in neuroticism - frequently experiences intense emotions such ...
Neu2) somewhat high in neuroticism - tends to experience negative emotions more ...
Neu3) neither high nor low in neuroticism - has an average level of emotional stability ...
Neu4) somewhat low in neuroticism - generally calm and emotionally stable - they ...
Neu5) very low in neuroticism - exceptionally calm, resilient, and emotionally stable - ...

',
  'CURRENT VOTING PREFERENCES - TURNOUT IN THE 2024 PRESIDENTIAL ELECTION IF THE ELECTION
      WERE HELD ON THE DATE OF THEIR MOST RECENT TWEET:
Tcu1) no chance this individual would turn-out to vote - Probability: 0 - in the ...
Tcu2) highly unlikely this individual would turn-out to vote - Probability: 0.15 - ...
Tcu3) unlikely this individual would turn-out to vote - Probability: 0.3 - in the ...
Tcu4) 50-50 likelihood that this individual would vote - Probability: 0.5 - in the ...
Tcu5) likely this individual would turn-out to vote - Probability: 0.7 - in the ...
Tcu6) highly likely this individual would turn-out to vote - Probability: 0.85 - in the ...
Tcu7) certain this individual would turn-out to vote - Probability: 1 - in the ...

',
  'CURRENT VOTING PREFERENCES - VOTE CHOICE IN THE 2024 PRESIDENTIAL ELECTION IF THE
      ELECTION WERE HELD ON THE DATE OF THEIR MOST RECENT TWEET:
Vcu1) would not vote in the 2024 elections for President in their state
Vcu2) would vote for Donald Trump, the Republican Party candidate ...
Vcu3) would vote for Kamala Harris, the Democratic Party candidate ...
Vcu4) would vote for Robert F. Kennedy Jr., who is not affiliated with any ...
Vcu5) would vote for Jill Stein, the Green Party candidate ...
Vcu6) would vote for Chase Oliver, the Libertarian Party candidate ...
Vcu7) would vote for Dr. Cornel West, who is not affiliated with any political party, ...

)
```

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

```
**title: HIGHEST EDUCATIONAL QUALIFICATION**
**explanation: The user mentions being a former instructor at University, which strongly implies a high level of education.**
**symbol: Q2**
**category: completed education at the college or university level - bachelor's degree, master's degree, doctorate**
**speculation: 15**

**title: ETHNICITY**
**explanation: The user includes the Italian flag in their bio, indicating a likely Italian heritage, which falls under the white category.**
**symbol: E1**
**category: white**
**speculation: 5**

**title: HOUSEHOLD INCOME BRACKET**
**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.**
**symbol: H5**
**category: more than 100000 USD per year**
**speculation: 50**

**title: INTEREST IN POLITICS**
**explanation: The user identifies as a Democrat and engages in political commentary on Twitter, indicating a high interest in politics.**
**symbol: I4**
**category: highly interested in politics**
**speculation: 10**

**title: SEX**
**explanation: The user's name "Karen" and profile picture indicate a feminine sex.**
**symbol: S2**
**category: feminine sex - female**
**speculation: 5**

**title: IDEOLOGICALLY, THIS PERSON APPEARS TO BE**
**explanation: The user identifies as a Democrat and expresses strong opinions about political figures, suggesting a strong ideological leaning.**
**symbol: Ide2**
**category: very ideologically liberal**
**speculation: 20**

**title: MARITAL STATUS**
**explanation: There is no direct information about marital status in the user data, making this highly speculative.**
**symbol: M2**
**category: single**
**speculation: 85**

**title: PAST VOTE - TURNOUT IN THE 2020 PRESIDENTIAL ELECTION**
**explanation: The user's strong political engagement and identification as a Democrat make it highly likely they turned out to vote.**
**symbol: Tpa7**
**category: certain this individual turned out to vote**
**speculation: 20**

**title: PAST VOTE - VOTE CHOICE IN THE 2020 PRESIDENTIAL ELECTION**
**explanation: The user identifies as a Democrat, making it likely they voted for Joe Biden.**
**symbol: Vpa3**
**category: voted for Joe Biden, the Democratic Party candidate, in the 2020 election for
```

```
President in their state**
**speculation: 10**

**title: GENERAL TRUST IN OTHER PEOPLE**
**explanation: There is no information directly indicating this user's level of trust in other people, making this highly speculative.**
**symbol: Tru3**
**category: about half of the time trust other people**
**speculation: 80**

**title: PAYING ATTENTION TO THE 2024 PRESIDENTIAL ELECTION**
**explanation: The user's active political engagement and commentary indicate they are paying a lot of attention to the upcoming election.**
**symbol: Att4**
**category: paying a lot of attention to the 2024 Presidential election in the US**
**speculation: 20**

**title: PAST VOTE - VOTE CHOICE IN THE 2022 HOUSE OF REPRESENTATIVES ELECTION**
**explanation: As a self-identified Democrat, it is likely they voted for a Democratic Party candidate in the 2022 elections.**
**symbol: Vhpa3**
**category: voted for a Democratic Party candidate in the 2022 elections for the House of Representatives in their congressional district**
**speculation: 30**

**title: AGE**
**explanation: Based on the professional experience detailed in the bio, including 20 years writing about jazz and previous hard news experience, the user is likely in an older age bracket.**
**symbol: A6**
**category: 55 to 64 years old**
**speculation: 30**

**title: PARTISAN LOYALTIES**
**explanation: The user identifies as a Democrat, suggesting strong partisan loyalty.**
**symbol: Pid1**
**category: strongly identifies with Democrats**
**speculation: 10**

**title: PAST VOTE - TURNOUT IN THE 2022 HOUSE OF REPRESENTATIVES ELECTION**
**explanation: Given the user's high political engagement, it is highly likely they turned out to vote in the 2022 elections.**
**symbol: Thpa7**
**category: certain this individual turned out to vote**
**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
- $\boldsymbol{\mathcal{F}}_1$: empty features object
- $\mathcal{O}_2$: operation to extract features
- $\mathcal{M}$: mould structure

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

Routine *background_informed_feature_extraction*:

$\mathcal{P}_1 = \mathcal{B} \parallel \mathcal{O}_1(\boldsymbol{\mathcal{F}}_1)$; `# Concatenate: to form` 1^{st} `prompt`

$\boldsymbol{\mathcal{F}}_2 \leftarrow \mathrm{GPT}\,(\mathcal{P}_1)$; `# Call: GPT to generate bg-informed features`

$\mathcal{P}_2 = \mathcal{B} \parallel \mathcal{M}(\Upsilon) \parallel \mathcal{O}_2(\boldsymbol{\mathcal{F}}_2)$; `# Concatenate: to form` 2^{nd} `prompt`

$\boldsymbol{X}_i \leftarrow \mathrm{GPT}\,(\mathcal{P}_2)$; `# 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 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 Lab[7] 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.

```
BACKGROUND INFORMATION:
The candidates running in the 2024 US Presidential election in state -- Massachusetts -- are reported below.

Kamala Harris, the Democratic Party candidate
Robert F. Kennedy Jr., who is not affiliated with any political party
Donald Trump, the Republican Party candidate'

INSTRUCTIONS:
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.
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.
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.
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.
Do not produce any other text beyond the completed set of questions.

CURRENT VOTING PREFERENCES - TURNOUT IN THE 2024 PRESIDENTIAL ELECTION IF THE ELECTION WERE HELD ON THE DATE OF THEIR MOST RECENT TWEET:
Tcu1) no chance this individual will turn-out to vote - Probability: 0 - in the 2024 election for President in <INSERT_STATE_NAME_HERE>
Tcu2) highly unlikely this individual will turn-out to vote - Probability: 0.15 - in the 2024 election for President in <INSERT_STATE_NAME_HERE>
Tcu3) unlikely this individual will turn-out to vote - Probability: 0.3 - in the 2024 election for President in <INSERT_STATE_NAME_HERE>
Tcu4) 50-50 likelihood that this individual will vote - Probability: 0.5 - in the 2024 election for President in <INSERT_STATE_NAME_HERE>
Tcu5) likely this individual will turn-out to vote - Probability: 0.7 - in the 2024 election for President in <INSERT_STATE_NAME_HERE>
Tcu6) highly likely this individual will turn-out to vote - Probability: 0.85 - in the 2024 election for President in <INSERT_STATE_NAME_HERE>
Tcu7) certain this individual will turn-out to vote - Probability: 1 - in the 2024 election for President in <INSERT_STATE_NAME_HERE>

CURRENT VOTING PREFERENCES - VOTE CHOICE IN THE 2024 PRESIDENTIAL ELECTION IF THE ELECTION WERE HELD ON THE DATE OF THEIR MOST RECENT TWEET:
Vcu1) would not vote in the 2024 election for President in <INSERT_STATE_NAME_HERE>
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>
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>
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>
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>
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

This Section documents `PoSSUM`'s polling effort for the 2024 election. A number of `PoSSUM` polls are fielded starting August 15^{th}, each reported in their respective Sections below (i. Section 4; ii. Section 5). For each poll, the anonymised individual-level data is freely available on https://github.com/robertocerinaprojects/PoSSUM.

For each poll I describe the subject pool (i. Section 4.1; ii. Section 5.1), and analyse sampling efficiency according to user-acceptance and rejection statistics over the sampling period. Word-clouds are used to visualise the distribution of $\mathbb{X}$ API query search-terms across users in the subject pool and amongst accepted users.

I compare the characteristic distribution of the accepted-users, and the known sample of unfilled cell quotas, against the target distribution for each poll (i. Section 4.2; ii. Section 5.2). The reference stratification frame, which I use to inform the target distribution for each poll as described in Section 2.2.2, is defined by the following variables: `sex` $\in$ {`male`,`female`}; `age` $\in$ {`25-34`, `35-44`, `45-54`, `55-64 18-24`, `65+`}; `hh_income` $\in \{[0, 25k), [25k, 50k), [50k, 75k), [75k, 100k), [100k, \infty]\}$; `race_ethnicity` $\in$ {`black`, `white`, `hispanic`, `asian`, `multiracial`, `native` }; `vote2020` $\in$ {`D`, `G`, `L`,`R`, `stayed home`}. Note that the marginal characteristic distributions of the accepted sample may differ slightly from the marginals of the final survey-object – 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).

Poll-wise `PoSSUM` estimates for topline vote share and socio-demographic crosstabs, derived upon applying the weighting procedure described in Section 3.1, are presented in Section 4.4 for the first poll; Section **??** for the second 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.

These estimates are compared with distributions published by other pollsters who fielded surveys around the same time in Sections 4.5 and **??** respectively. The statistical model used to obtain posterior distributions for each crosstab presented by a comparison-pollster is described in Section 3.2. In comparing these distributions, note that the primary `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 records[8]. 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 speculation distribution of each poll is presented respectively in Sections 4.3; 5.3.

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

3.1 Representative Inference

Given the sampling protocol, `PoSSUM` samples will tend to be 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 the joint distribution of 2020 Vote Choice as derived from an auxiliary survey. It differs from traditional MrsP in that it doesn't use the auxiliary 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 in the auxiliary survey is relatively small. I use the 2022 Cooperative Election Study (CES) [27] as the auxiliary survey to get estimates of 2020 recall vote[9]. 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.

[9] I 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.

Structured Priors: The final Hierarchical Model used to generate smoothed estimates 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; state[10] 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 2 presents the covariates and parameters used in the model for 2024 vote choice.

predictor	*level*	*description*	*index*	*domain*	*parameter*	*prior correlation structure*
1	global	/	/	/	α_j	iid
/	state	state_id	l	{1,...,54}	λ_{sj}	spatial (BYM2)
/	individual	date_id	t	{1,...,T}	η^D_{tj}	random-walk
/		age_id	a	{1,...,6}	η^A_{aj}	random-walk
/		income_id	h	{1,...,5}	η^H_{hj}	random-walk
/		sex_id	g	{1,2}	γ^G_{gj}	unstructured + shared variance
/		race_id	r	{1,...,6}	γ^R_{rj}	unstructured + shared variance
/		vote20_id	v	{1,...,5}	γ^V_{vj}	unstructured + shared variance
z_1	state	2020 *R* share	/	$\mathbb{R}$	$\beta_{1j=\text{R}}$	iid
z_2		On ballot: R.F.K. Jr.			$\beta_{1j=\text{K}}$	
z_3		On ballot: Jill Stein			$\beta_{1j=\text{G}}$	
z_4		2020 *G* share			$\beta_{2j=\text{G}}$	
z_5		On ballot: Chase Oliver			$\beta_{1j=\text{K}}$	
z_6		2020 *L* share			$\beta_{2j=\text{R}}$	
z_7		On ballot: Cornel West			$\beta_{1j=\text{W}}$	
z_8		2020 "stay home" share			$\beta_{1j=\text{stay\_home}}$	

Table 2: Model Predictors and Parameters for the 2024 vote-choice model. 'iid' refers to fully independent parameters, or 'fixed' effects [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.

[10]Because we have an interest in being able to estimate the number of electoral votes won by each candidate, we treat the congressional districts of Nebraska and Maine as separate states.

The covariates used for other models involving different dependent variables are very similar, and available in the GitHub repository. I present the full[11] Hierarchical Bayesian model below – see [5] for a more attentive explanation of each model component. I describe the generation of given choice $j \in \{1, ..., J\}$, made by a sampled user $i \in \{1, ..., n\}$, as follows:

$$
\begin{aligned}
y_{ij} \sim& \text{Categorical}(\pi_{i1}, ..., \pi_{iJ}) && \texttt{likelihood} \\
\pi_{ij} =& \frac{\exp(\mu_{ij})}{\sum_j \exp(\mu_{i,j})}; && \texttt{softmax link} \\
\mu_{ij} =& \alpha_j + \\
& \lambda_{\text{state\_id}[i],j} + \\
& \eta^P_{\text{poll\_id}[i],j} + \eta^A_{\text{age\_id}[i],j} + \eta^H_{\text{income\_id}[i],j} + \\
& \gamma^G_{\text{sex\_id}[i],j} + \gamma^R_{\text{race\_id}[i],j} + \gamma^V_{\text{vote20\_id}[i],j} + \\
& \boldsymbol{\beta}_j \boldsymbol{Z}^j; && \texttt{linear predictor} \\
\alpha_j \sim& \text{N}(0,1); && \texttt{intercept} \\
\lambda_{sj} =& \ \sigma^\lambda_j \left(\phi_{sj} \sqrt{(1-\xi_j)} + \psi_{sj} \sqrt{(\xi_j/\epsilon)} \right); && \texttt{BYM2 effects} \\
\phi_{sj} \sim& N(0,1); && \texttt{unstructured effect} \\
\psi_{sj} \mid \psi_{s'j} \sim& N\left(\frac{\sum_{l' \neq l} \psi_{s'j}}{\nu_l}, \frac{1}{\sqrt{\nu_l}} \right); && \texttt{conditional auto-regressive effects} \\
\xi_j \sim& \text{Beta}\left(\frac{1}{2}, \frac{1}{2} \right); && \texttt{mixing weights} \\
\sigma^\lambda_j \sim& N^+(0,1); && \texttt{state-level scale} \\
\gamma^U_{uj} \mid \gamma^U_{u-1\ j} \dots \gamma^U_{1j} \sim& N(\gamma^U_{u-1\ j}, \sigma^U_j), \ \forall \ U \in \{A, H, P\}; && \texttt{random walk effects} \\
\gamma^U_{uj} \sim& N(0, \sigma^U_j), \ \forall \ U \in \{G, R, V\}; && \texttt{unstructured effects} \\
\sigma^U_j \sim& N^+(0,1); && \texttt{random effect scales} \\
\boldsymbol{\beta}_j \sim& N^+(0,1). && \texttt{fixed effects}
\end{aligned}
$$

[11]Note: this model is the 'saturated' model used to aggregate learning from multiple polls conducted during a given period – hence the poll-level random effects η^P. These poll effects are ordered in time, according to the fieldwork dates of the poll, and as such are assigned random-walk priors to allow for temporal smoothing. The model fit to weight individual polls is identical, with the only exception being the absence of poll-level effects.

3.1.1 Learning from Stateless Users

`PoSSUM`'s geographic inclusion criteria ensures users who are selected for analysis are most likely based in the US. The protocol does however allow for the inclusion in the sample of users whose state of residency within the US – their 2^{nd} level geography – is unknown. Learning from these users can bring to bear evidence pertaining the relationship between individual-level attributes, such as age, gender, education, past-vote, etc., and 2024 voting preferences. How to best allow for this in the hierarchical model described above ?

I consider an approach that uses two separate linear predictors, one for the observations missing a state, and one for those observations which are complete. For the latter, the linear predictor is exactly as described above; for the former, the following linear predictor is used:

$$
\begin{aligned}
\mu_{ij}^{s'} =& \alpha_j + \\
& \xi_j + \\
& \eta^{P}_{\text{poll\_id}[i],j} + \eta^{A}_{\text{age\_id}[i],j} + \eta^{H}_{\text{income\_id}[i],j} + \\
& \gamma^{G}_{\text{sex\_id}[i],j} + \gamma^{R}_{\text{race\_id}[i],j} + \gamma^{V}_{\text{vote20\_id}[i],j} + \\
& \boldsymbol{\beta}_j \boldsymbol{Z}^j; && \texttt{no-state linear predictor}; \\
\xi_j \sim& \mathrm{N}(0,1); && \texttt{no-state fixed effect};
\end{aligned}
$$

where ξ_j is a 'no-state' fixed effect, which captures the average difference between the baseline level of support relative to the average state, and the 'no-state' pool's support. Effectively, I'm treating the 'no-state' label as a separate, independent state, which is not pooled towards the state-level effect average. The remaining individual-level coefficients are still informed by these users' preferences. Making out-of-sample predictions I then use the linear predictor for the observations with known states, effectively discarding the no-state effect.

There are alternatives to this formulation – for instance we could retain the two separate linear predictors, but not include a separate 'no-state' effect. This would effectively let the 'no-state' pool inform the intercept in full, without netting-out the difference with the rest of the sample. I consider this approach unwise: composition effects in this particular pool of users could lead to distortions of the baseline rate of support. Imagine for instance that all of these users would belong to one particularly polarised state – say California or Wyoming. If their state of residency was known, its effect would largely be captured by the respective state-level random effect, and the intercept would be largely untouched. We recover similar behaviour by introducing a 'no-state' fixed effect. However, were we to let this pool of 'masked' Californians or Wyomingites inform the intercept, our out-of-sample predictions would become skewed according to the unknown states' respective bias. According to a principle of caution, we opt to net-out the unknown-state effects as described.

3.2 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 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, 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.

$$
\begin{aligned}
y \sim & \text{Multinomial}(\boldsymbol{\pi}, n); \\
\boldsymbol{\pi} \sim & \text{Dirichlet}\left(\frac{\mathbf{1}}{\mathbf{2}}\right).
\end{aligned}
$$

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 available for comparison is typically 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 drop responses associated with 'would not vote' and 'unsure' choice-options, 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.

4 1^{st} Poll: August 15^{th} - August 23^{rd}

I field a `PoSSUM` poll starting on August 15^{th} and ending on August 23^{rd}. The quotas are 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 3). During this time in the field I obtain valid digital traces for 945 unique users [12]. The costs of the exercise can be broken down as follows: \$145 were spent on 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 \$0.94.

4.1 Subject Pool

The `get_pool` routine is executed on four occasions.Table 3 presents the total number of candidate and accepted users by pool-date, broken down by the trending and political search-terms. Table 4 presents the distribution of reasons-for-exclusion over time. Figure 2 presents the distribution of trending topics amongst candidate and accepted users.

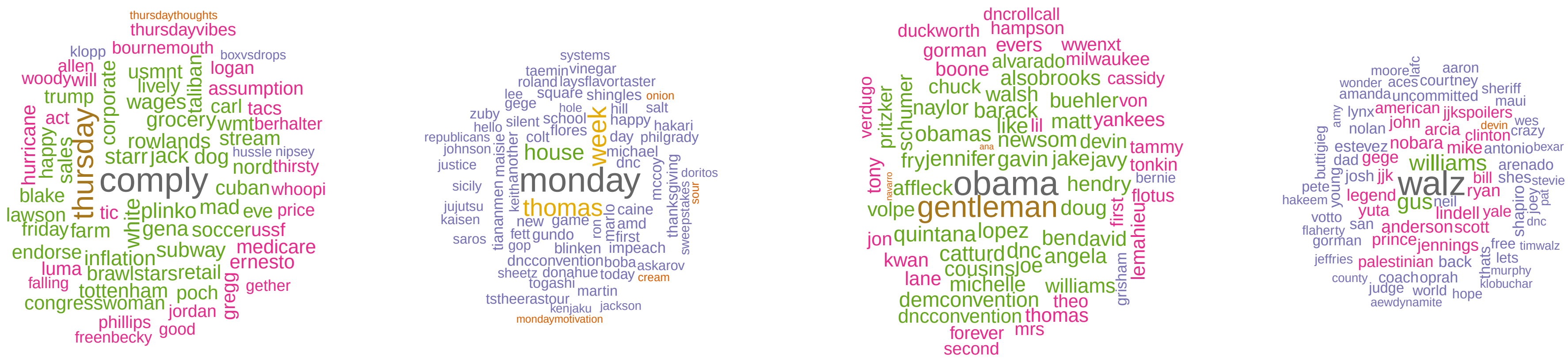
	2024-08-15	2024-08-19	2024-08-21	2024-08-22
"Fresh" Users in the Subject Pool				
Total	3,670	3,721	3,615	3,520
Politics	1,960	1,944	1,955	1,867
Trends	1,710	1,777	1,660	1,653
Accepted Users				
Total	489	258	129	80
Politics	190	95	56	33
Trends	299	163	73	47

Table 3: Breakdown of users by subject-pool type and inclusion in the survey-object, for each instance of the `get_pool` routine.

[12]There can be a minor discrepancy between this number and the number of accepted users presented in the tables below. This is due to post-sampling cleaning, which can highlight invalid survey-entries or accidental duplicates.

	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 4: Breakdown of reasons for rejecting users from being included in the sample. Note that each 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.



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

(e) Sample - 15/08 (f) Sample - 19/08 (g) Sample - 21/08 (h) Sample - 22/08

Figure 2: Word-cloud presenting words from the ‘trending’ queries to the 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. Fieldwork dates 08/15 to 08/23.

4.2 Unfilled Quotas

The sample frame is composed of 435 mutually exclusive cells. Figure 3 presents the marginal distributions associated with the filled quotas, against those of the unfilled quotas. 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%).

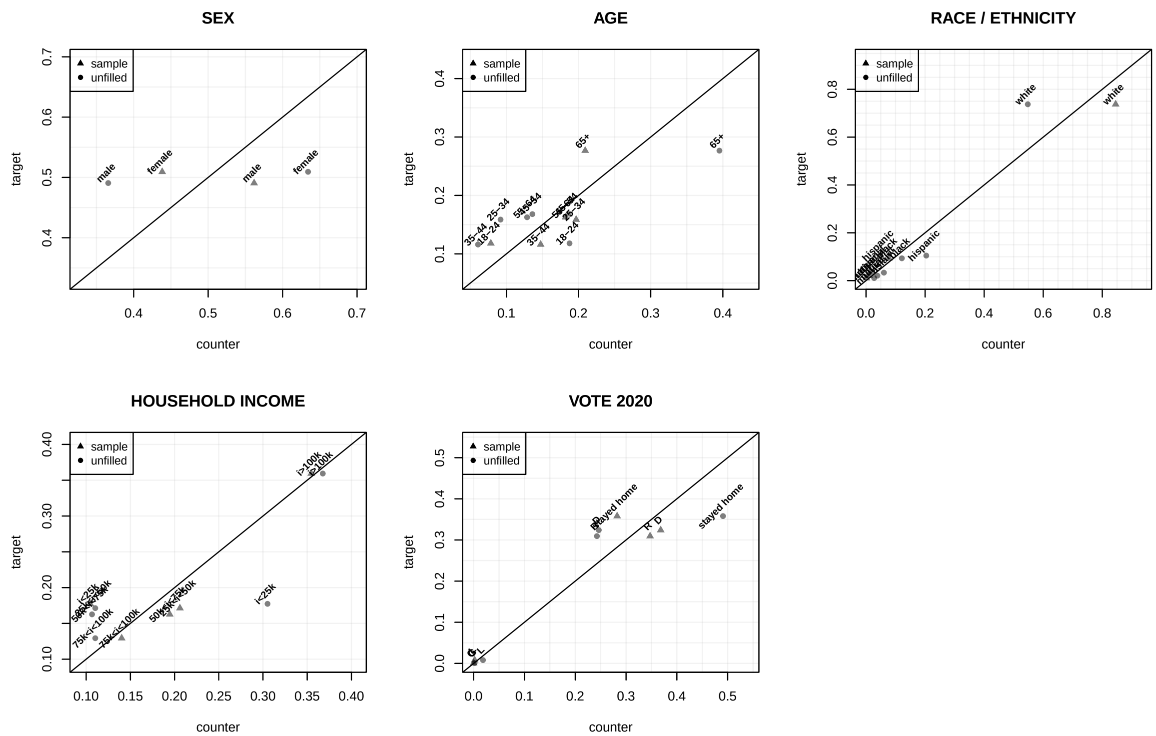


Figure 3: Comparing the filled and unfilled quotas (x-axis) against the sample-level stratification frame (y-axis). Fieldwork dates 08/15 to 08/23.

4.3 Speculation

Figure 4 presents the general distribution of speculation levels across covariates and dependent variables.

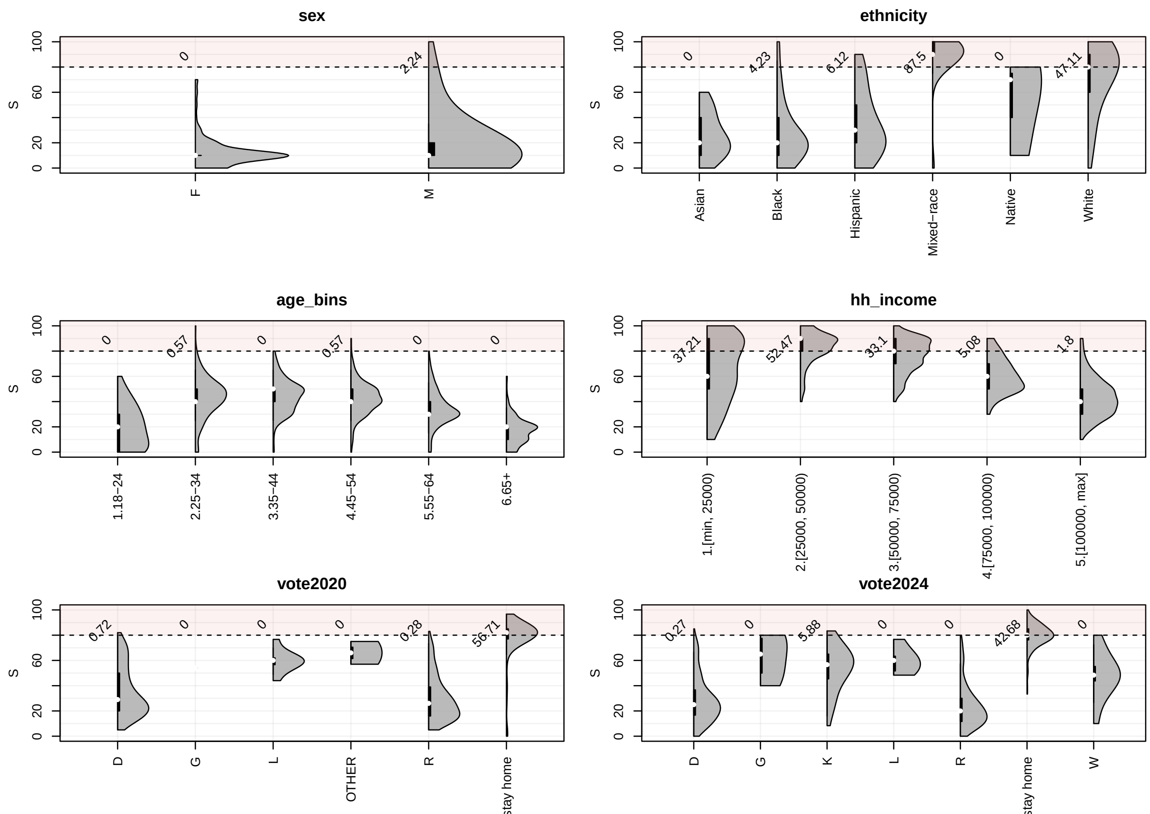


Figure 4: Distribution of the speculation score within variables used for modeling. Each distribution is accompanied by a number on the top-left, representing the % of the speculation scores which are greater than the “highly speculative” threshold. Fieldwork dates 08/15 to 08/23.

4.4 Topline & Crosstabs - 2024 Vote Choice

Table 5 presents the topline, national vote share estimates. Crosstabs by Region, Sex, Race/Ethnicity, Education, Household Income, Age and 2020 Vote Choice are available below. Note: upon news of Robert F. Kennedy Jr.'s endorsement of Donald Trump, and subsequent effects to the ballot-access map in each state, we chose to re-weight the data via an MrP model that accounts for these changes. We did this to ensure this first can be directly compared to subsequent polls, holding ballot access constant. This improves the ability to make statements about movements in support over the course of the campaign.

Table 5: Topline estimates of 2024 Vote Choice, 08/15 to 08/23.

Pop.	Vote2024	Topline
LV	Harris (D)	46.4 (44.2 , 48.3)
LV	Trump (R)	47.2 (45.1 , 49.3)
LV	RFK Jr (Ind)	3.7 (2.4 , 5.3)
LV	Stein (G)	1.1 (0.4 , 2.5)
LV	West (Ind)	0.2 (0 , 0.7)
LV	Oliver (L)	1 (0.5 , 2)
A	Abstention	30 (27.6 , 32.2)
A	Turnout	70 (67.8 , 72.4)

Table 6: Estimates of 2024 Vote Choice by Region, 08/15 to 08/23

Pop.	Vote2024	Midwest	Northeast	South	West
LV	Harris (D)	43.4 (41.4 , 45.3)	52.3 (49.5 , 54.5)	43.1 (40.7 , 45)	50.1 (46.6 , 52.4)
LV	Trump (R)	49 (46 , 51.5)	44.3 (42 , 46.8)	51.5 (48.8 , 53.9)	41.5 (38.6 , 44.4)
LV	RFK Jr. (Ind)	5.6 (3.5 , 8.4)	1.5 (0.8 , 2.4)	2.7 (1.6 , 4.2)	5.1 (3.2 , 7.5)
LV	Stein (G)	0.6 (0.2 , 1.4)	0.9 (0.3 , 2.4)	1.1 (0.3 , 2.5)	1.6 (0.5 , 4.3)
LV	West (Ind)	0.2 (0 , 0.9)	0 (0 , 0.1)	0.2 (0 , 0.9)	0.2 (0 , 0.8)
LV	Oliver (L)	0.9 (0.4 , 2)	0.8 (0.3 , 1.5)	1 (0.4 , 2.1)	1.1 (0.4 , 2.4)
A	Abstention	29.1 (26.3 , 31.8)	30.8 (27.6 , 34.3)	31.9 (28.8 , 34.4)	27.1 (23.3 , 30.7)
A	Turnout	70.9 (68.2 , 73.7)	69.2 (65.7 , 72.4)	68.1 (65.6 , 71.2)	72.9 (69.3 , 76.7)

Table 7: Estimates of 2024 Vote Choice by Sex, 08/15 to 08/23

Pop.	Vote2024	F	M
LV	Harris (D)	51.3 (48.4 , 53.7)	41 (38.4 , 43.1)
LV	Trump (R)	43.4 (40.6 , 45.9)	51.6 (49 , 54.3)
LV	RFK Jr. (Ind)	3.3 (1.9 , 5.1)	4.3 (2.6 , 6.3)
LV	Stein (G)	1.1 (0.4 , 3)	1 (0.3 , 2.5)
LV	West (Ind)	0.1 (0 , 0.6)	0.2 (0 , 0.9)
LV	Oliver (L)	0.5 (0 , 1.6)	1.5 (0.7 , 3)
A	Abstention	27.3 (24.1 , 30.5)	32.8 (30.1 , 35.4)
A	Turnout	72.7 (69.5 , 75.9)	67.2 (64.6 , 69.9)

Table 8: Estimates of 2024 Vote Choice by Race/Ethnicity, 08/15 to 08/23

Pop.	Vote2024	Asian	Black	Hispanic	Mixed-Race	Native	White
LV	Harris (D)	61.9 (49.4 , 68.9)	78.1 (72 , 83.4)	59.2 (52.7 , 64.5)	42.5 (35.5 , 50.3)	41.4 (29.4 , 51.4)	40.5 (38.4 , 42.4)
LV	Trump (R)	30.8 (24.8 , 41.5)	16.7 (11.6 , 21.7)	35.4 (30.2 , 41.3)	48.1 (35.3 , 56.3)	46.9 (32.2 , 60)	53.2 (50.9 , 55.4)
LV	RFK Jr. (Ind)	1.8 (0.2 , 6)	1.2 (0.1 , 4)	1.7 (0.2 , 5.5)	5.2 (1.1 , 14.5)	5.3 (1.3 , 18.5)	4.2 (2.6 , 6)
LV	Stein (G)	2.5 (0.5 , 13.6)	1.5 (0.3 , 4.8)	1.4 (0.2 , 5.2)	1.2 (0.2 , 4.6)	1.5 (0.2 , 9.2)	0.7 (0.3 , 1.9)
LV	West (Ind)	0.1 (0 , 0.6)	0.5 (0.1 , 2)	0.2 (0 , 0.7)	0.4 (0 , 1.8)	0.3 (0 , 1.5)	0.1 (0 , 0.6)
LV	Oliver (L)	0.8 (0.1 , 2.6)	1 (0.2 , 2.7)	1 (0.2 , 3.4)	1.4 (0.2 , 4.5)	1.3 (0.3 , 5.6)	0.9 (0.4 , 1.9)
A	Abstention	25.7 (16.9 , 32.8)	37.7 (33.2 , 42.1)	38 (32.3 , 43.1)	30.4 (24.7 , 38)	37.8 (25.9 , 46.2)	28 (25.5 , 30.3)
A	Turnout	74.3 (67.2 , 83.1)	62.3 (57.9 , 66.8)	62 (56.9 , 67.7)	69.6 (62 , 75.3)	62.2 (53.8 , 74.1)	72 (69.7 , 74.5)

Table 9: Estimates of 2024 Vote Choice by College Grad., 08/15 to 08/23

Pop.	Vote2024	Does not Have a College Degree	Has a College Degree
LV	Harris (D)	42.1 (39.6 , 44.4)	53.6 (51.3 , 55.4)
LV	Trump (R)	51.3 (48.6 , 53.9)	40.4 (38.5 , 42.4)
LV	RFK Jr. (Ind)	3.7 (2.3 , 5.4)	3.7 (2.4 , 5.4)
LV	Stein (G)	1.1 (0.4 , 2.9)	1 (0.4 , 2)
LV	West (Ind)	0.2 (0 , 0.8)	0.1 (0 , 0.6)
LV	Oliver (L)	1 (0.5 , 2.2)	0.9 (0.4 , 1.7)
A	Abstention	35.6 (32.8 , 38.3)	18.1 (15.9 , 20.1)
A	Turnout	64.4 (61.7 , 67.2)	81.9 (79.9 , 84.1)

Table 10: Estimates of 2024 Vote Choice by Houshold Income, 08/15 to 08/23

Pop.	Vote2024	[0, 25k)	[25k, 50k)	[50k, 75k)	[75k, 100k)	[100k, +)
LV	Harris (D)	47.5 (39.9 , 54.1)	43.9 (40.8 , 47.1)	44.2 (41.6 , 46.9)	44.8 (41.6 , 47.4)	48.4 (45.5 , 50.6)
LV	Trump (R)	44.6 (36.5 , 50.2)	50.8 (47.4 , 53.9)	50.9 (48 , 53.8)	49.8 (46.2 , 53.3)	44.6 (42.1 , 47.4)
LV	RFK Jr. (Ind)	3.2 (1.1 , 6.4)	2.8 (1.5 , 4.9)	2.8 (1.4 , 4.5)	3.2 (1.7 , 5.5)	4.7 (2.8 , 7.3)
LV	Stein (G)	1.8 (0.5 , 9.7)	1 (0.3 , 2.8)	0.7 (0.2 , 1.7)	0.6 (0.1 , 1.8)	0.9 (0.3 , 2.2)
LV	West (Ind)	0.4 (0.1 , 1.8)	0.2 (0 , 0.9)	0.2 (0 , 0.7)	0.1 (0 , 0.6)	0.1 (0 , 0.6)
LV	Oliver (L)	1 (0.2 , 3.2)	0.8 (0.2 , 2)	1 (0.4 , 2.2)	1 (0.4 , 2.2)	0.9 (0.4 , 2.2)
A	Abstention	48.7 (42.8 , 54.7)	37.1 (33.4 , 40.8)	28.8 (25.3 , 31.8)	22.7 (19.6 , 26.8)	20.7 (17.1 , 23.4)
A	Turnout	51.3 (45.3 , 57.2)	62.9 (59.2 , 66.6)	71.2 (68.2 , 74.7)	77.3 (73.2 , 80.4)	79.3 (76.6 , 82.9)

Table 11: Estimates of 2024 Vote Choice by Age, 08/15 to 08/23

Pop.	Vote2024	18-24	25-34	35-44	45-54	55-64	65+
LV	Harris (D)	57.9 (50.7 , 64)	50.4 (45.7 , 54.8)	50.6 (47.5 , 54.5)	44.8 (41.5 , 48)	43.8 (41.4 , 46.3)	42.8 (39.5 , 45.6)
LV	Trump (R)	30.4 (22.5 , 36.5)	37 (31.9 , 41.2)	40.4 (36.1 , 44.1)	49 (45.6 , 52.8)	52.1 (49.3 , 54.9)	54.3 (51.6 , 57)
LV	RFK Jr. (Ind)	6.6 (2.3 , 14.5)	8.8 (5 , 14.6)	5.9 (3.4 , 9.7)	3.3 (1.7 , 5.3)	2.1 (1 , 3.7)	1.1 (0.3 , 2.5)
LV	Stein (G)	1.5 (0.4 , 5.1)	1.2 (0.4 , 3)	1.2 (0.4 , 2.8)	1 (0.3 , 2.4)	0.9 (0.3 , 2.4)	0.8 (0.2 , 3)
LV	West (Ind)	0.8 (0.1 , 3.8)	0.4 (0.1 , 1.7)	0.2 (0 , 0.9)	0.1 (0 , 0.4)	0.1 (0 , 0.4)	0 (0 , 0.3)
LV	Oliver (L)	1.3 (0.2 , 5.3)	1.5 (0.5 , 3.8)	1 (0.3 , 2.3)	1.4 (0.6 , 3.4)	0.6 (0.1 , 1.6)	0.5 (0 , 1.5)
A	Abstention	53.8 (46.8 , 61.4)	42.5 (38.2 , 46.9)	31.4 (27.2 , 35.6)	24.5 (18.8 , 28.8)	22.3 (18.5 , 25.9)	21.5 (16.8 , 25.5)
A	Turnout	46.2 (38.6 , 53.2)	57.5 (53.1 , 61.8)	68.6 (64.4 , 72.8)	75.5 (71.2 , 81.2)	77.7 (74.1 , 81.5)	78.5 (74.5 , 83.2)

Table 12: Estimates of 2024 Vote Choice by 2020 Vote Choice, 08/15 to 08/23

Pop.	Vote2024	D	R	L	G	OTHER	Stay Home
LV	Harris (D)	94.8 (92 , 96.5)	0.7 (0.2 , 1.7)	14.1 (3.7 , 34.1)	21 (3.3 , 56.4)	10.6 (1.3 , 33.1)	33 (20.8 , 47.8)
LV	Trump (R)	1.1 (0.5 , 2.4)	95.6 (93.2 , 97.4)	18 (3.4 , 44.4)	22.3 (0.4 , 80.5)	44.4 (13.2 , 83.8)	45.9 (30.1 , 60.8)
LV	RFK Jr. (Ind)	1.7 (0.9 , 3.1)	2.9 (1.6 , 5.1)	35.8 (20.9 , 50.9)	17.4 (0.8 , 54.8)	26.4 (4.4 , 52.8)	12 (5.7 , 20.9)
LV	Stein (G)	1.4 (0.5 , 3.5)	0 (0 , 0.3)	1.1 (0.1 , 9.3)	2.4 (0.1 , 26.8)	1 (0 , 16.7)	3.3 (0.4 , 12)
LV	West (Ind)	0.2 (0 , 0.9)	0.1 (0 , 0.4)	1.8 (0.1 , 8.9)	3.9 (0.3 , 22.4)	0.6 (0 , 11.9)	0.2 (0 , 1)
LV	Oliver (L)	0.4 (0.1 , 1.2)	0.5 (0.1 , 1.3)	21.2 (6.1 , 44.7)	4.7 (0.1 , 50.5)	2.7 (0 , 27.5)	2.5 (0.4 , 9.1)
A	Abstention	8.3 (6.1 , 11.4)	2.3 (1.1 , 4.4)	5.7 (0.4 , 29.8)	12.4 (0.6 , 61.6)	8.4 (0.2 , 44)	80.3 (73.5 , 86.1)
A	Turnout	91.7 (88.6 , 93.9)	97.7 (95.6 , 98.9)	94.3 (70.2 , 99.6)	87.6 (38.4 , 99.4)	91.6 (56 , 99.8)	19.7 (13.9 , 26.5)

4.5 Cross-pollsters comparison - 2024 Vote Choice

`PoSSUM MrP` displays the following house-effects: i. lower-than-average (3 to 4 % points on the point estimate) share of Harris voters in the topline, largely driven by a lower share amongst whites and males around (2 to 3% % points on the point estimate) and seemingly lower support with voters earning $50k$$ or less; ii. lower-than-average % of 'switchers' between the two major parties, relative to the 2020 vote.

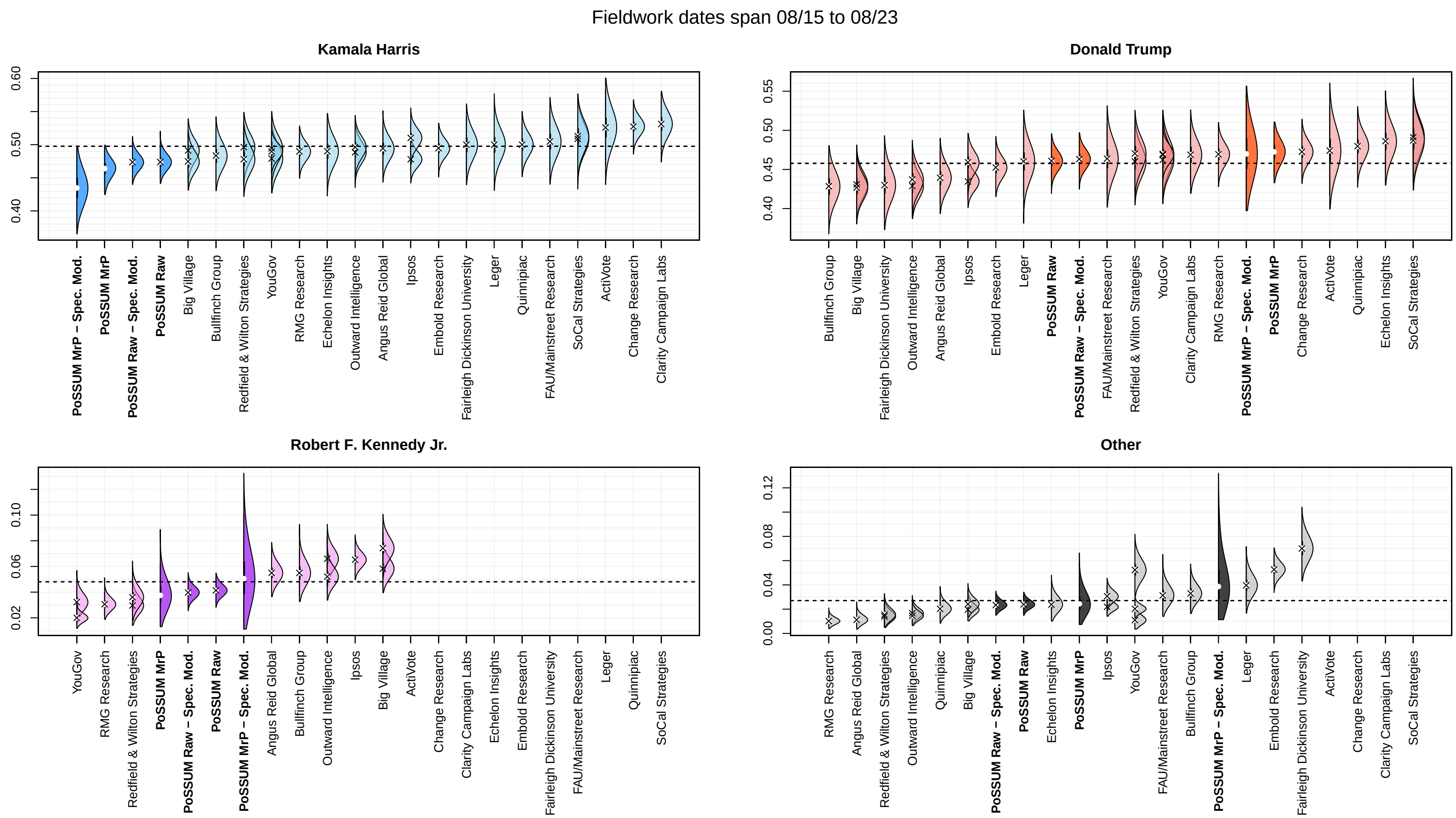


Figure 5: A comparison of the topline distribution of 2024 vote choice across pollsters. The dotted line represents the simple arithmetic average of all polls. `PoSSUM MrP` is the primary estimate of interest.

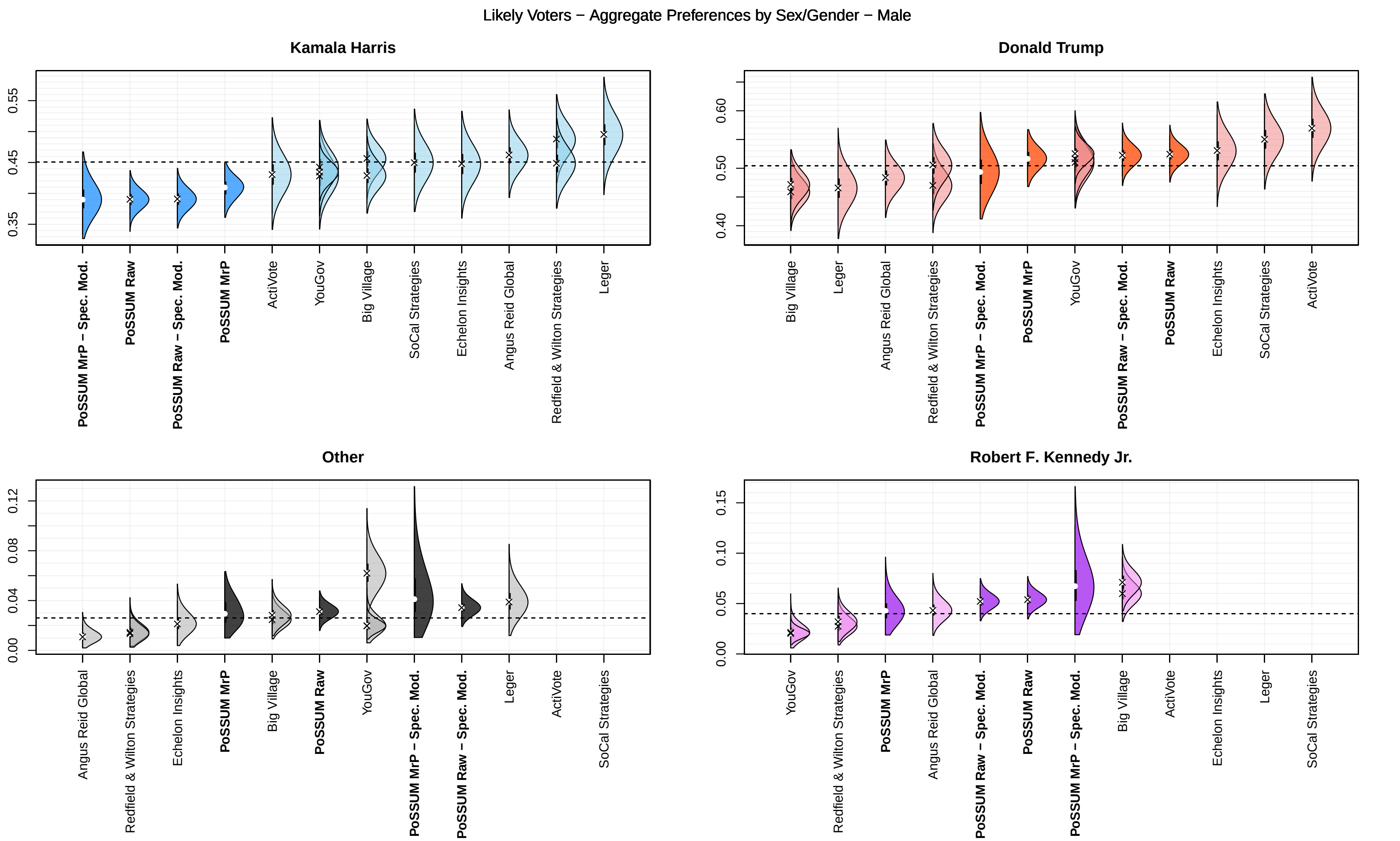


Figure 6: A comparison of the Male distribution of 2024 vote choice across pollsters. The dotted line represents the simple arithmetic average of all polls. `PoSSUM MrP` is the primary estimate of interest. Large intervals for a given pollster correspond to small n for the given crosstab.

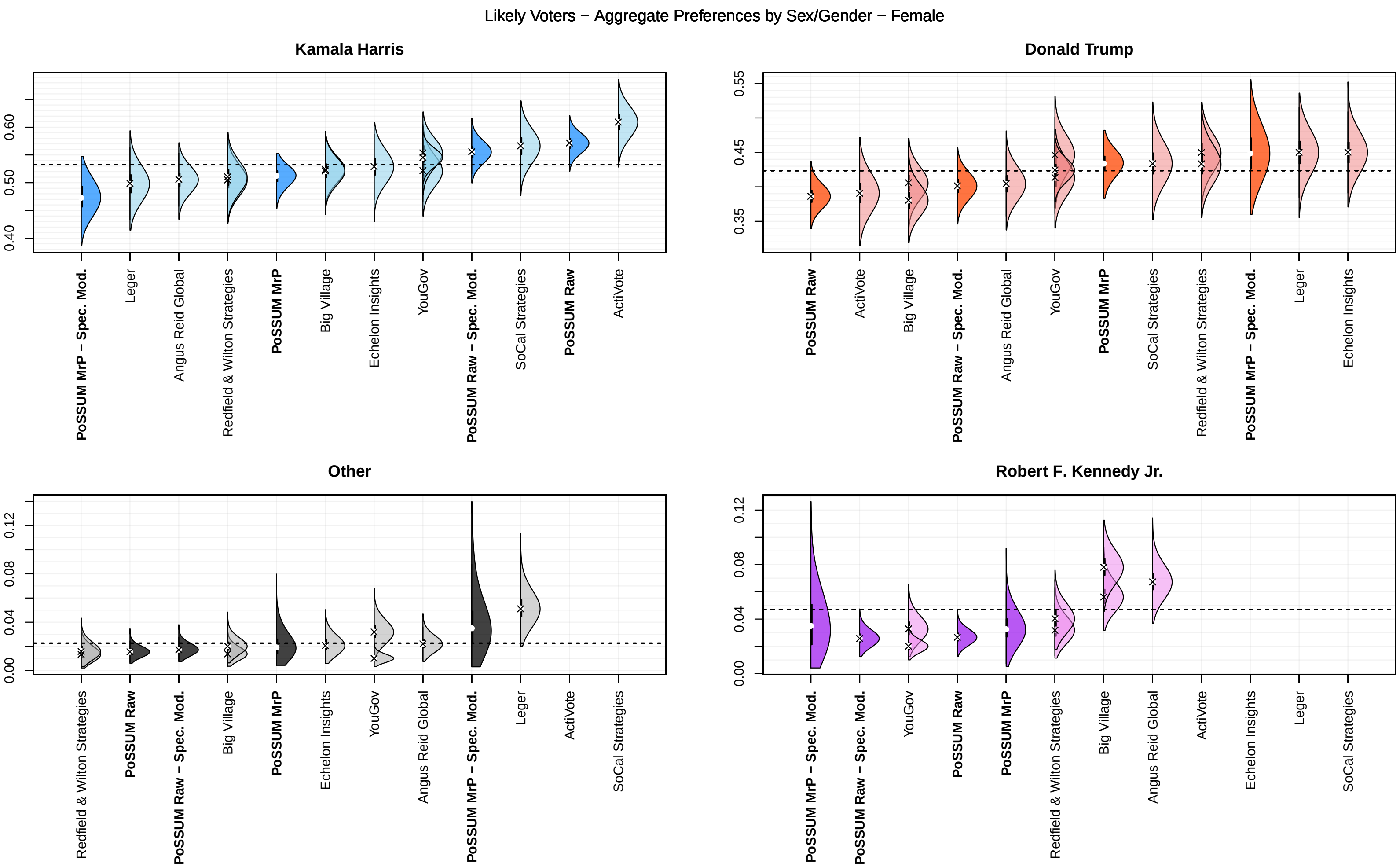


Figure 7: A comparison of the Female distribution of 2024 vote choice across pollsters. The dotted line represents the simple arithmetic average of all polls. `PoSSUM MrP` is the primary estimate of interest. `PoSSUM MrP` is the primary estimate of interest. Large intervals for a given pollster correspond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

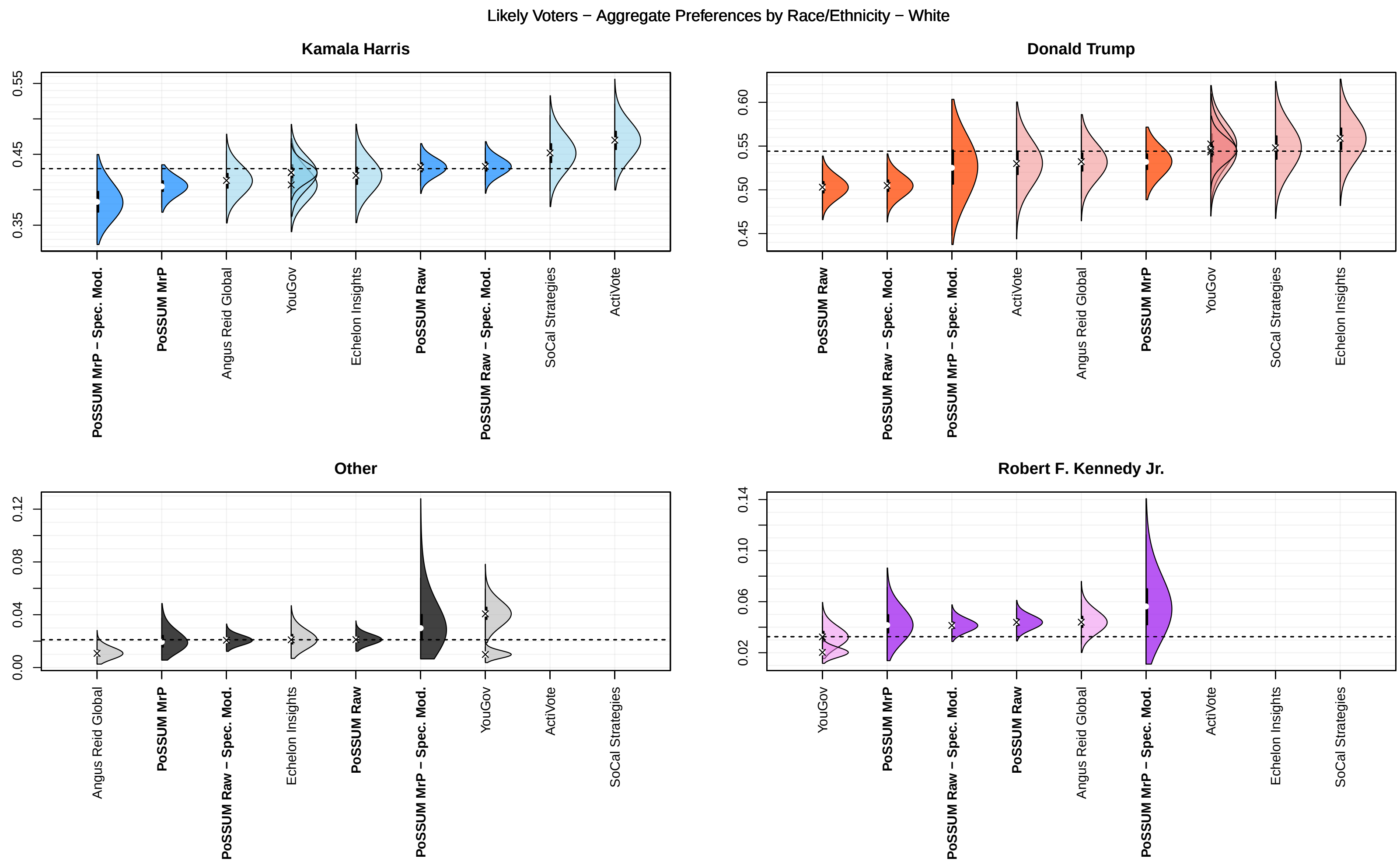


Figure 8: A comparison of the White distribution of 2024 vote choice across pollsters. The dotted line represents the simple arithmetic average of all polls. `PoSSUM MrP` is the primary estimate of interest. Large intervals for a given pollster correspond to small n for the given crosstab

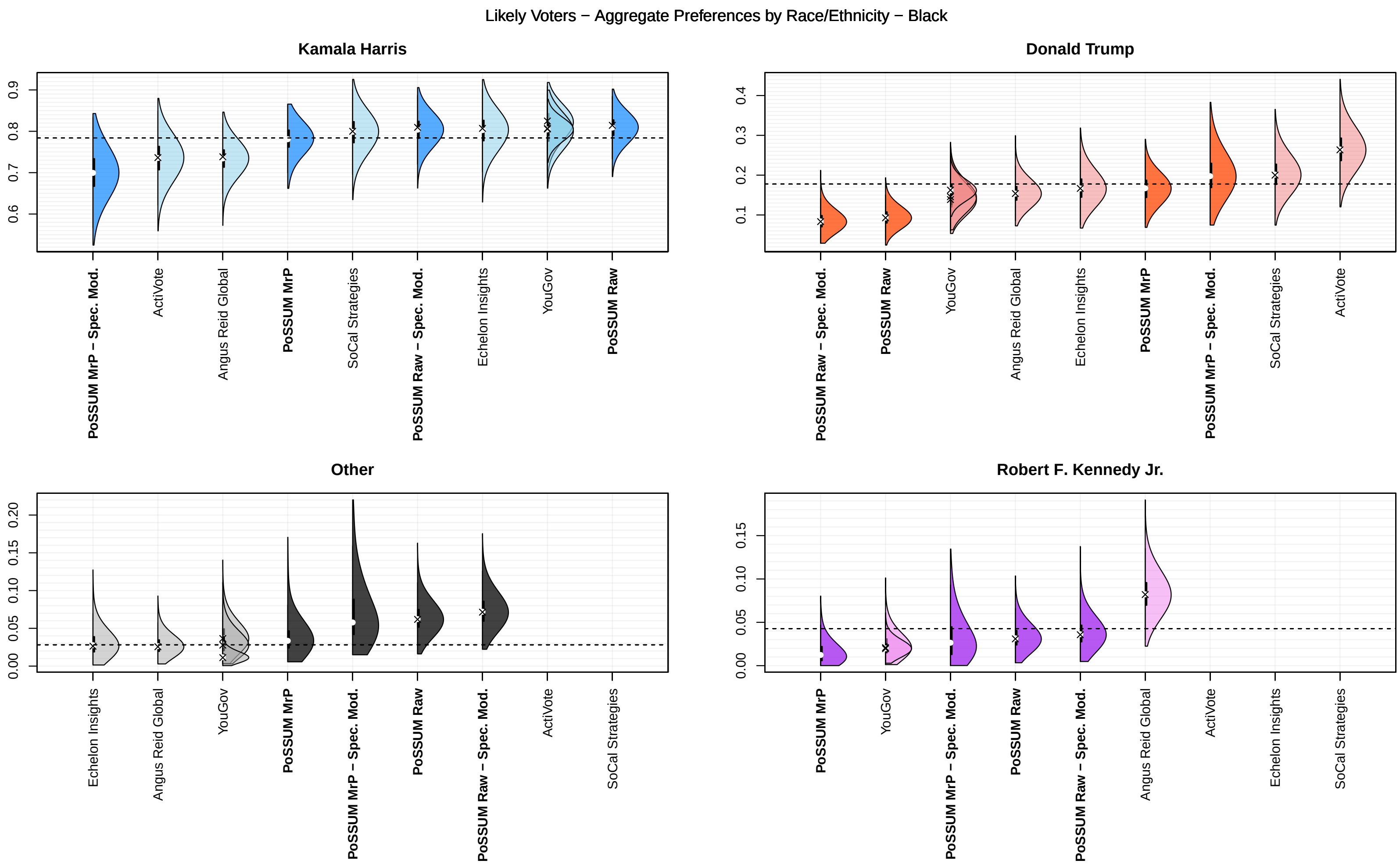


Figure 9: A comparison of the Black distribution of 2024 vote choice across pollsters. The dotted line represents the simple arithmetic average of all polls. `PoSSUM MrP` is the primary estimate of interest. Large intervals for a given pollster correspond to small n for the given crosstab

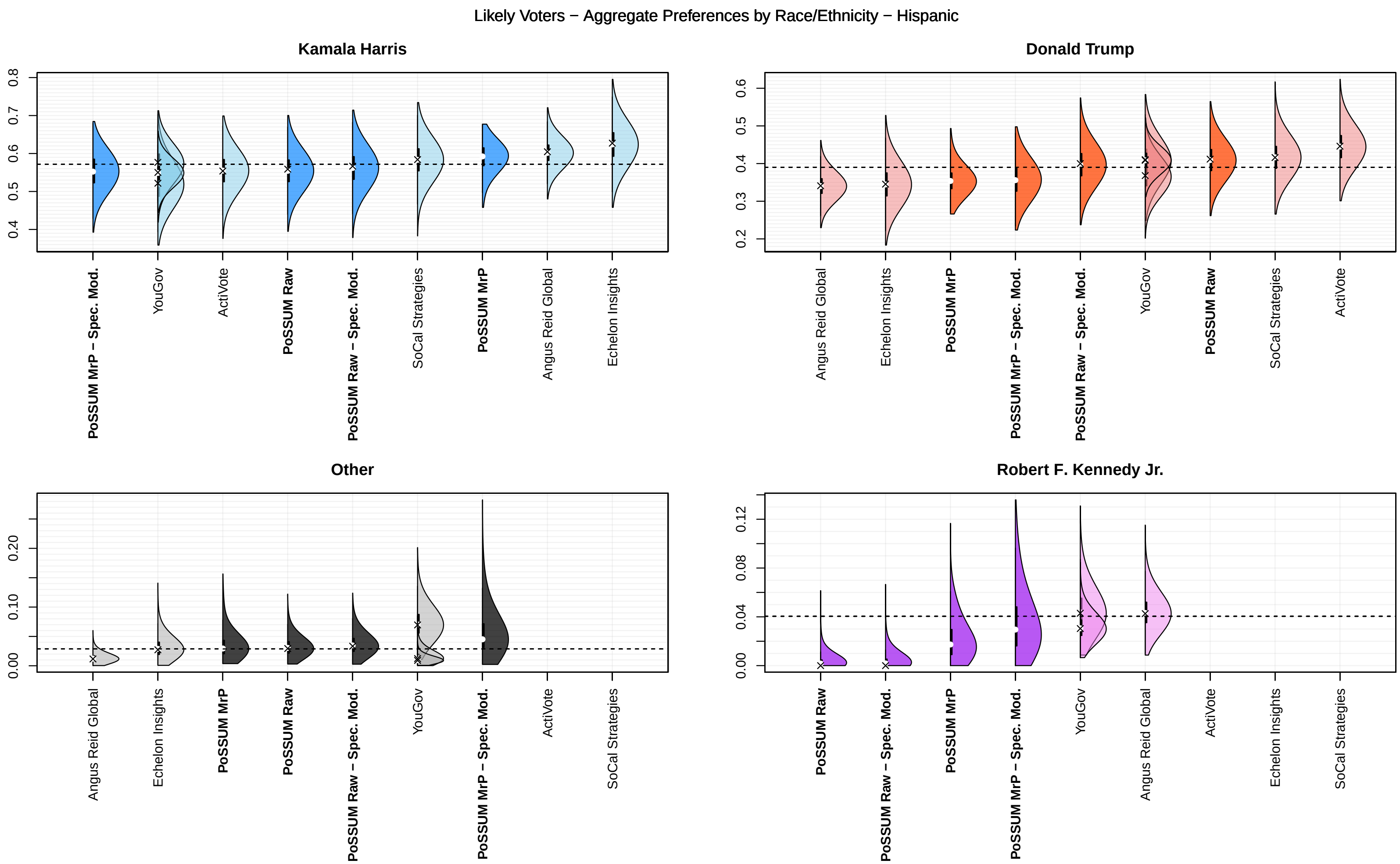


Figure 10: A comparison of the Hispanic distribution of 2024 vote choice across pollsters. The dotted line represents the simple arithmetic average of all polls. `PoSSUM MrP` is the primary estimate of interest. Large intervals for a given pollster correspond to small n for the given crosstab.

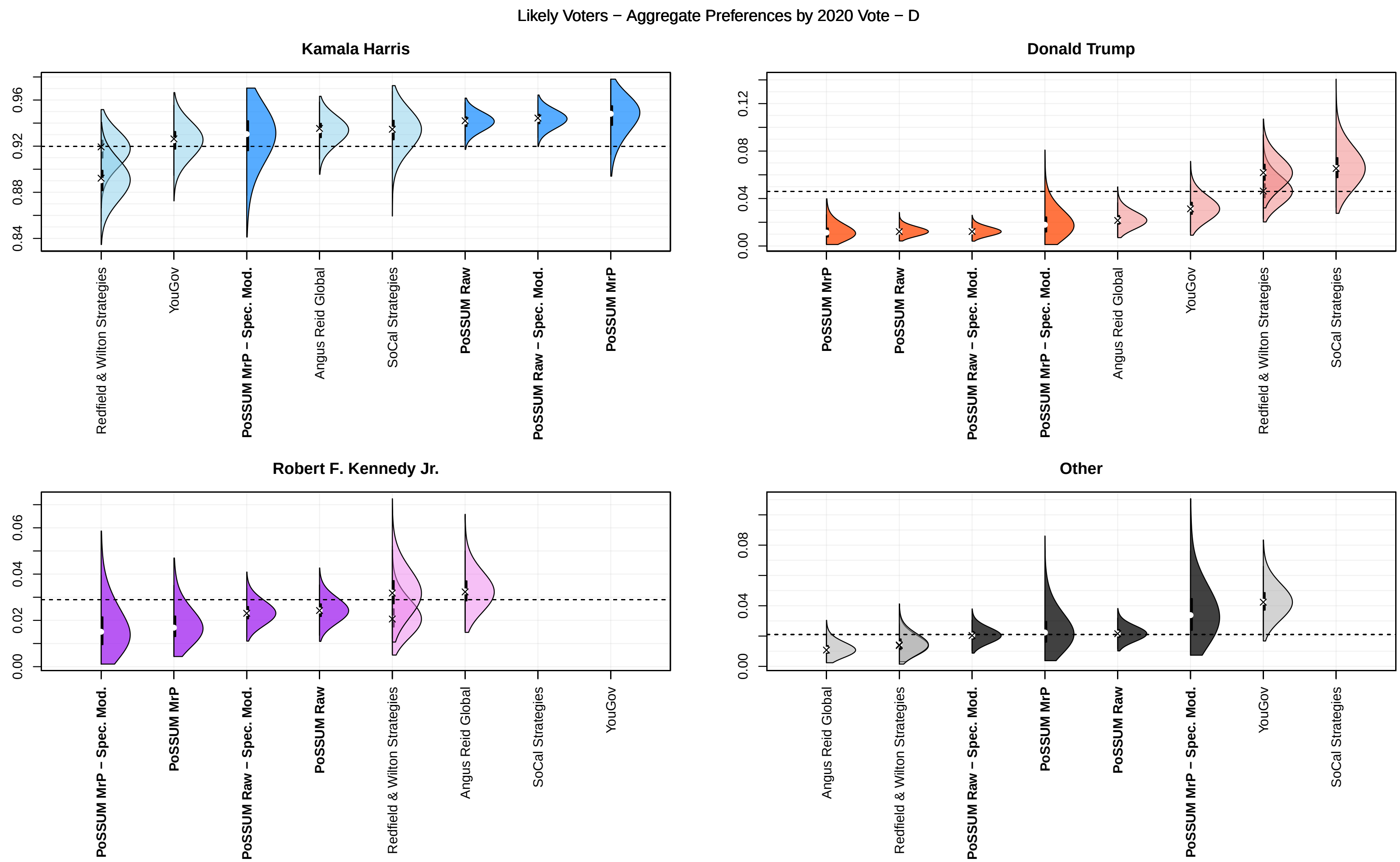


Figure 11: A comparison of the distribution of 2024 vote choice across pollsters, for those who voted for Joe Biden (D) in 2020. The dotted line represents the simple arithmetic average of all polls. `PoSSUM MrP` is the primary estimate of interest. Large intervals for a given pollster correspond to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

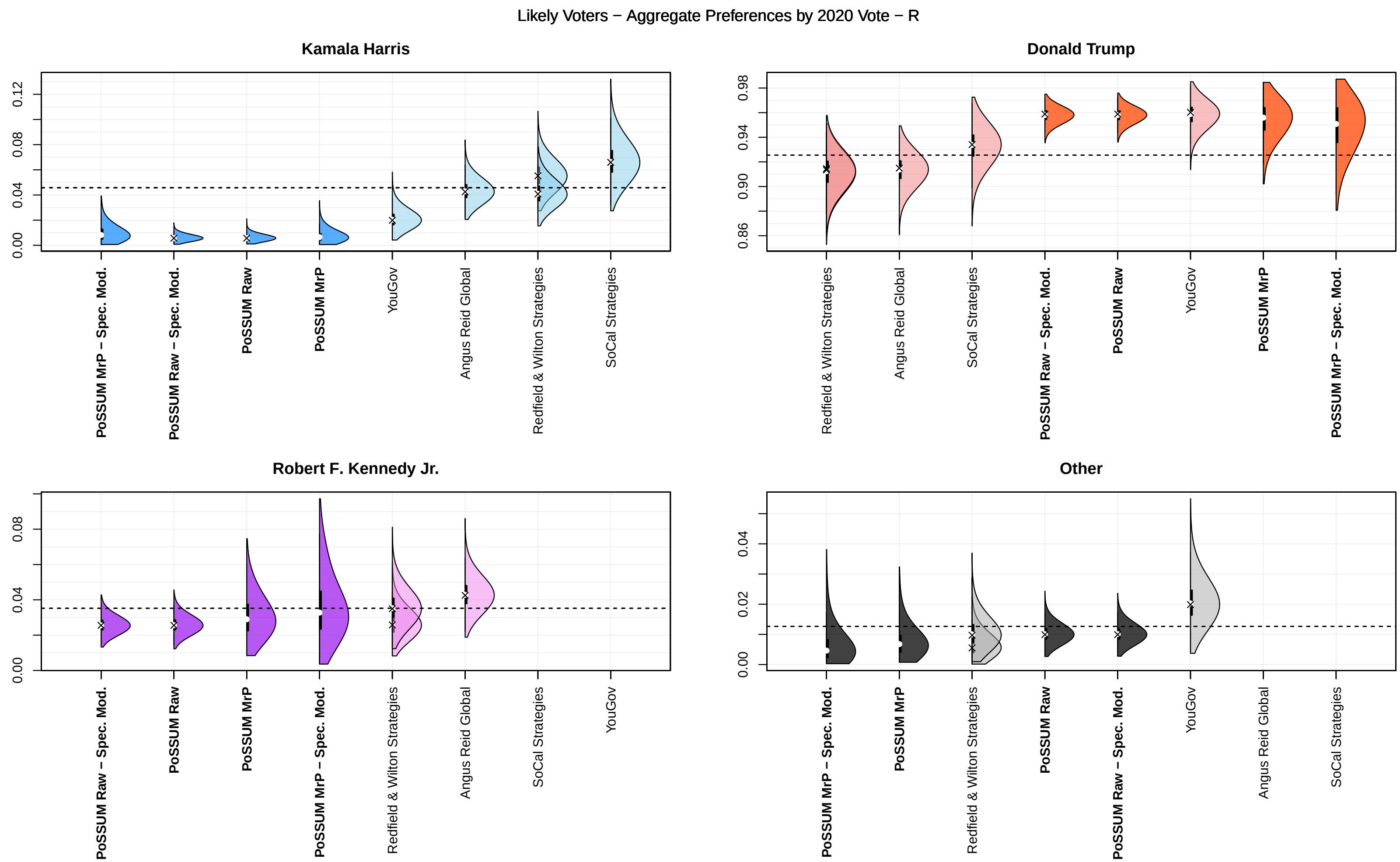


Figure 12: A comparison of the distribution of 2024 vote choice across pollsters, for those who voted for Donlad Trump (R) in 2020. The dotted line represents the simple arithmetic average of all polls. `PoSSUM MrP` is the primary estimate of interest. Large intervals for a given pollster correspond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

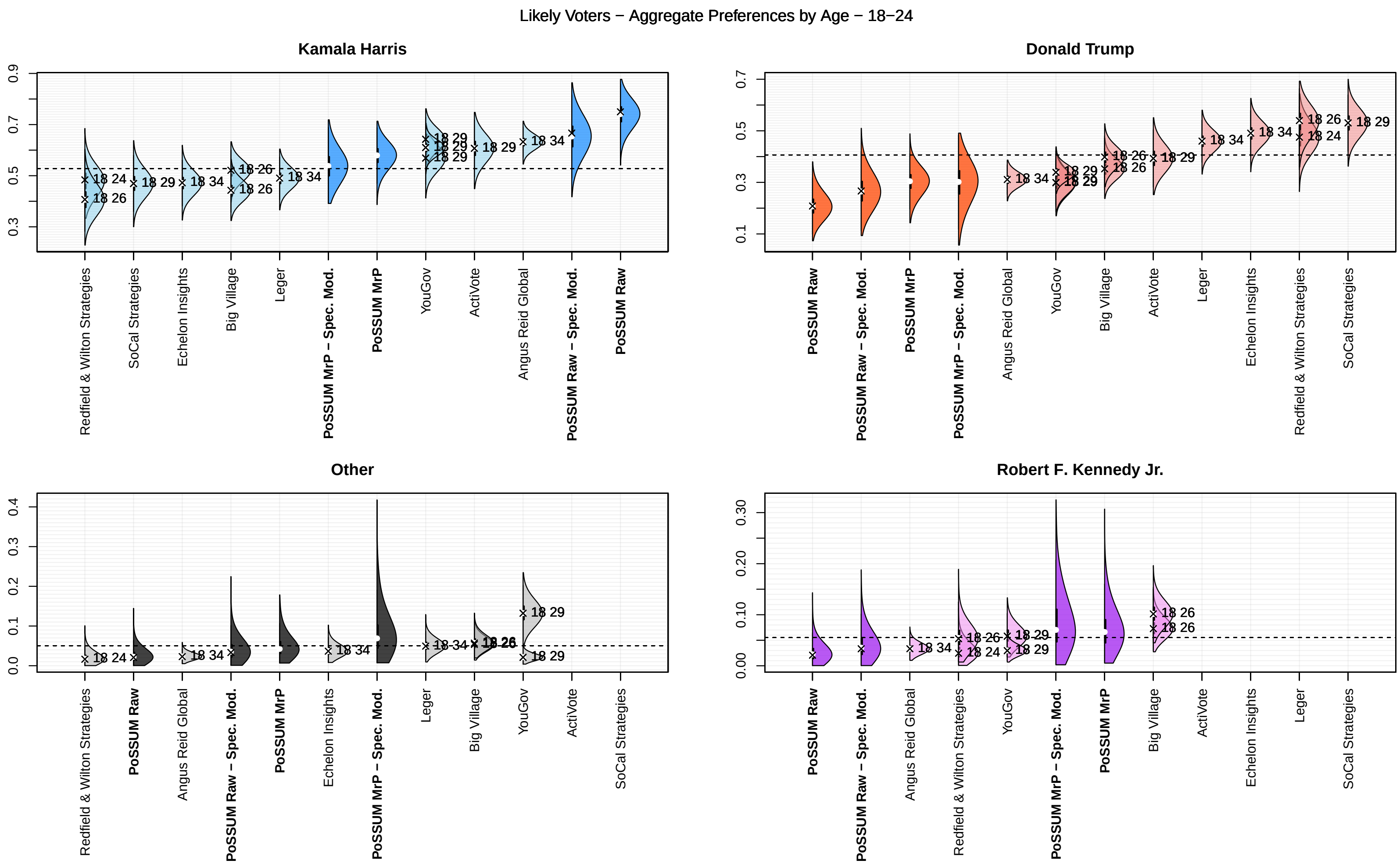


Figure 13: A comparison of the 18-24 years old distribution of 2024 vote choice across pollsters. The dotted line represents the simple arithmetic average of all polls. `PoSSUM MrP` is the primary estimate of interest. Large intervals for a given pollster correspond to small n for the given crosstab. 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.

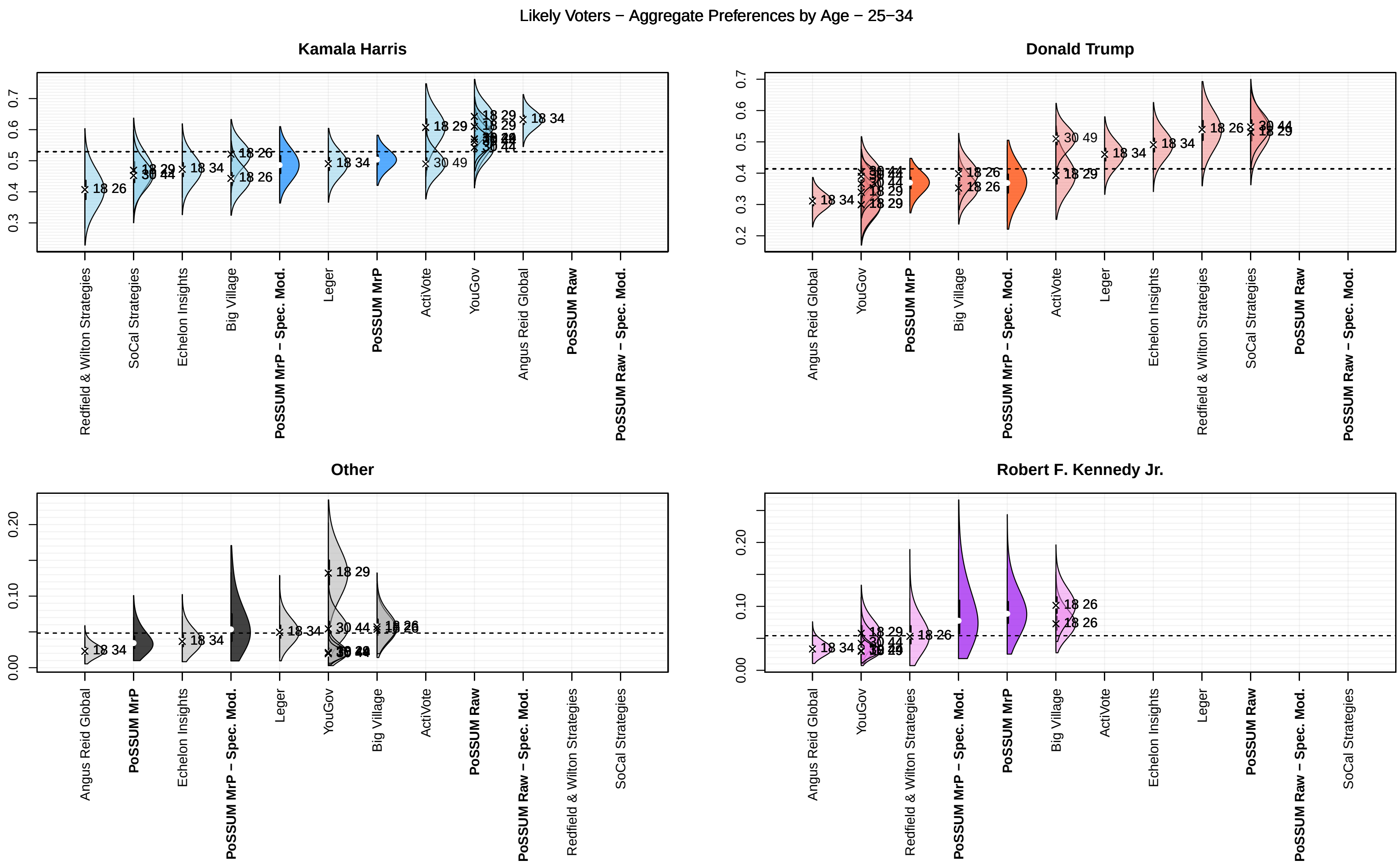


Figure 14: A comparison of the 25-34 years old distribution of 2024 vote choice across pollsters. The dotted line represents the simple arithmetic average of all polls. `PoSSUM MrP` is the primary estimate of interest. Large intervals for a given pollster correspond to small n for the given crosstab. 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.

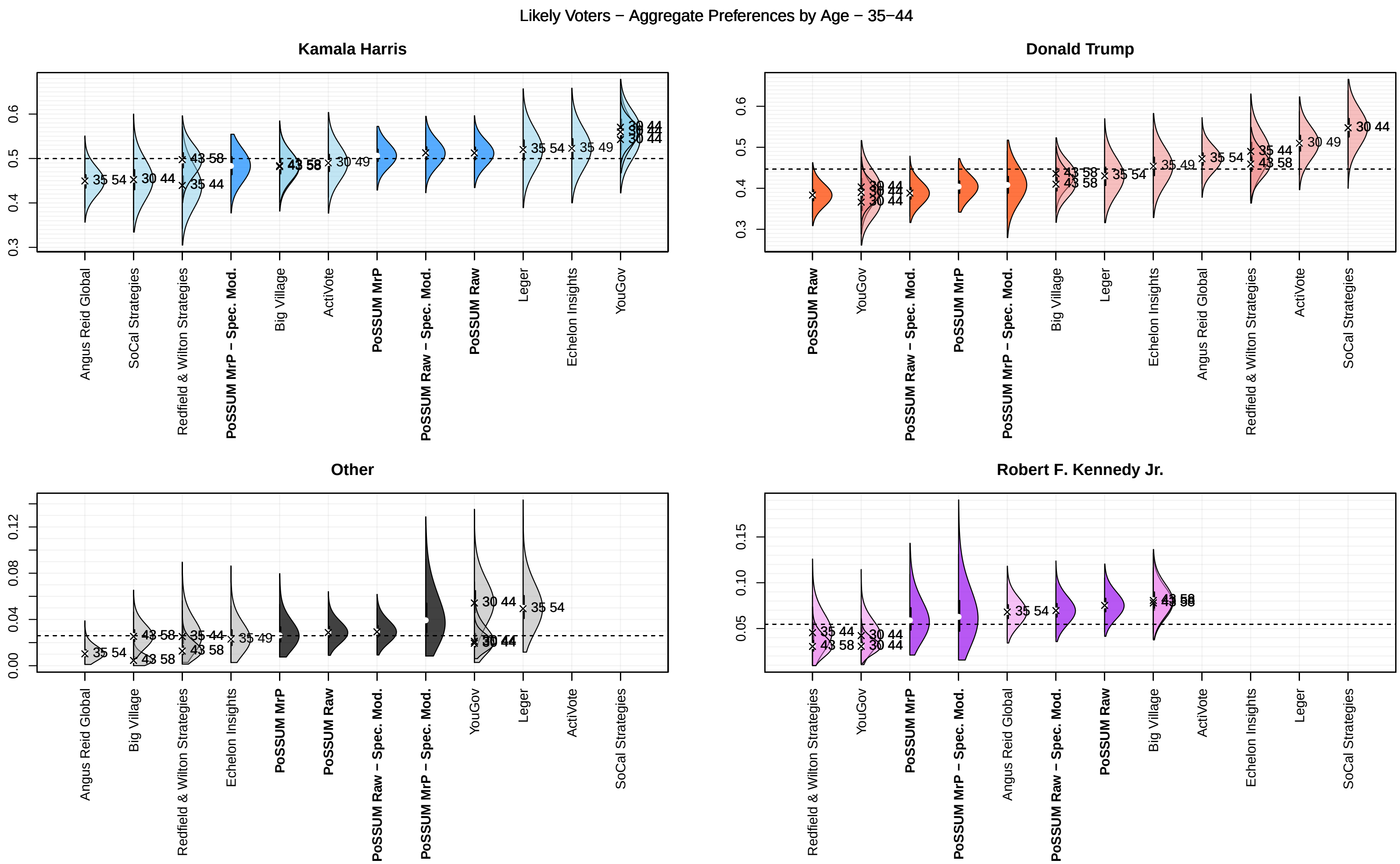


Figure 15: A comparison of the 35-44 years old distribution of 2024 vote choice across pollsters. The dotted line represents the simple arithmetic average of all polls. `PoSSUM MrP` is the primary estimate of interest. Large intervals for a given pollster correspond to small n for the given crosstab. 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.

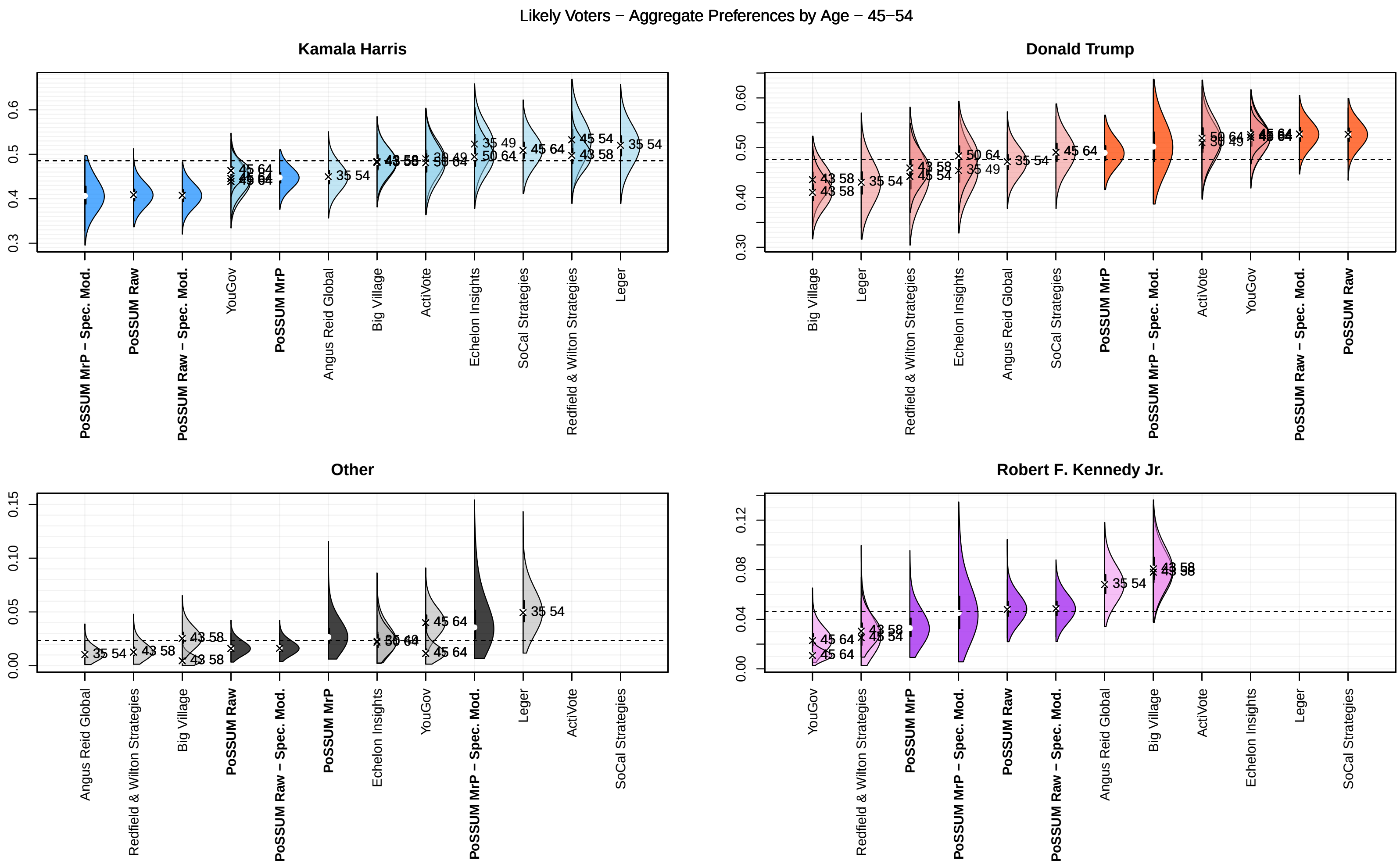


Figure 16: A comparison of the 45-54 years old distribution of 2024 vote choice across pollsters. The dotted line represents the simple arithmetic average of all polls. PoSSUM MrP is the primary estimate of interest. Large intervals for a given pollster correspond to small n for the given crosstab. To 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.

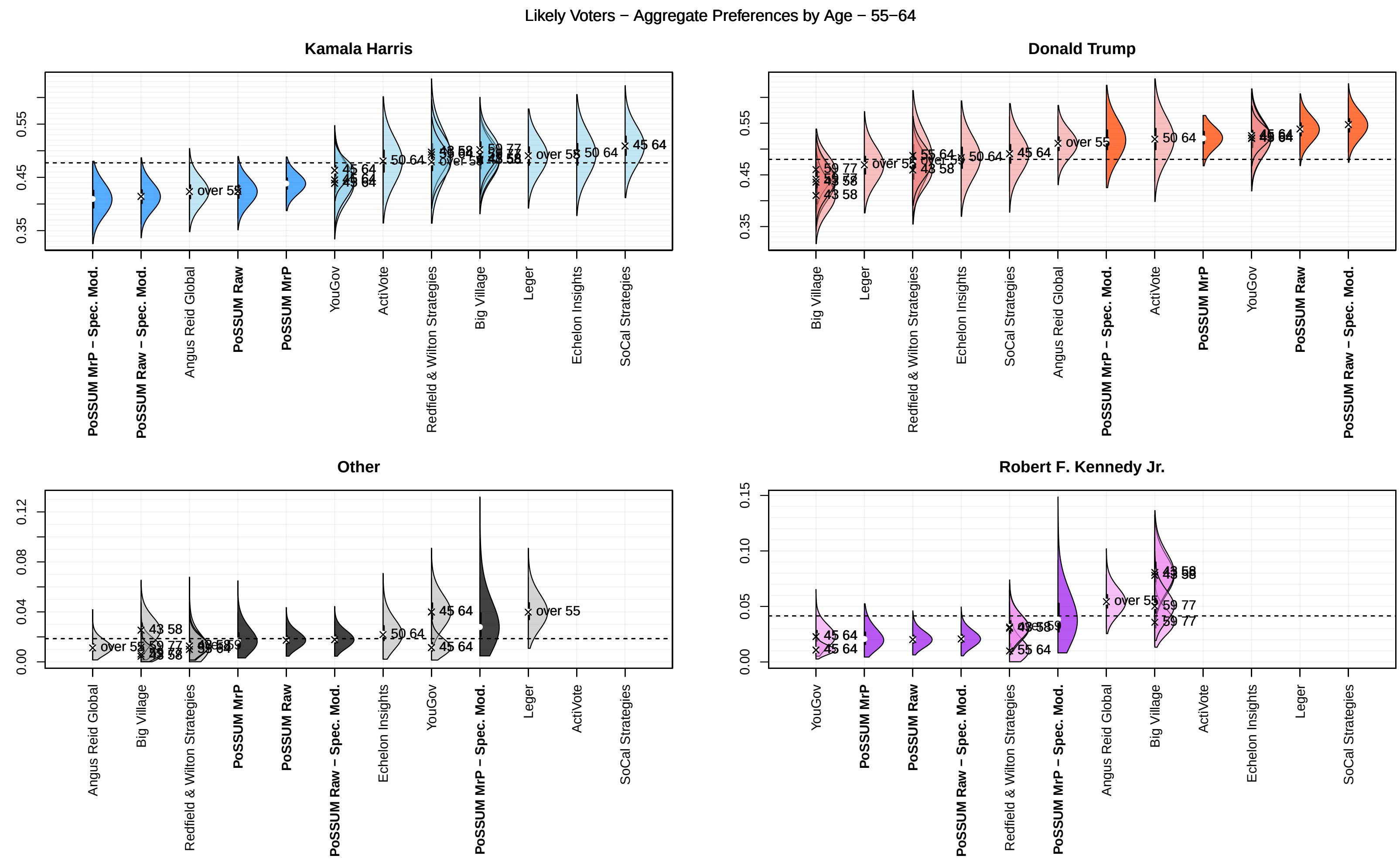


Figure 17: A comparison of the 55-64 years old distribution of 2024 vote choice across pollsters. The dotted line represents the simple arithmetic average of all polls. `PoSSUM MrP` is the primary estimate of interest. Large intervals for a given pollster correspond to small n for the given crosstab. 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.

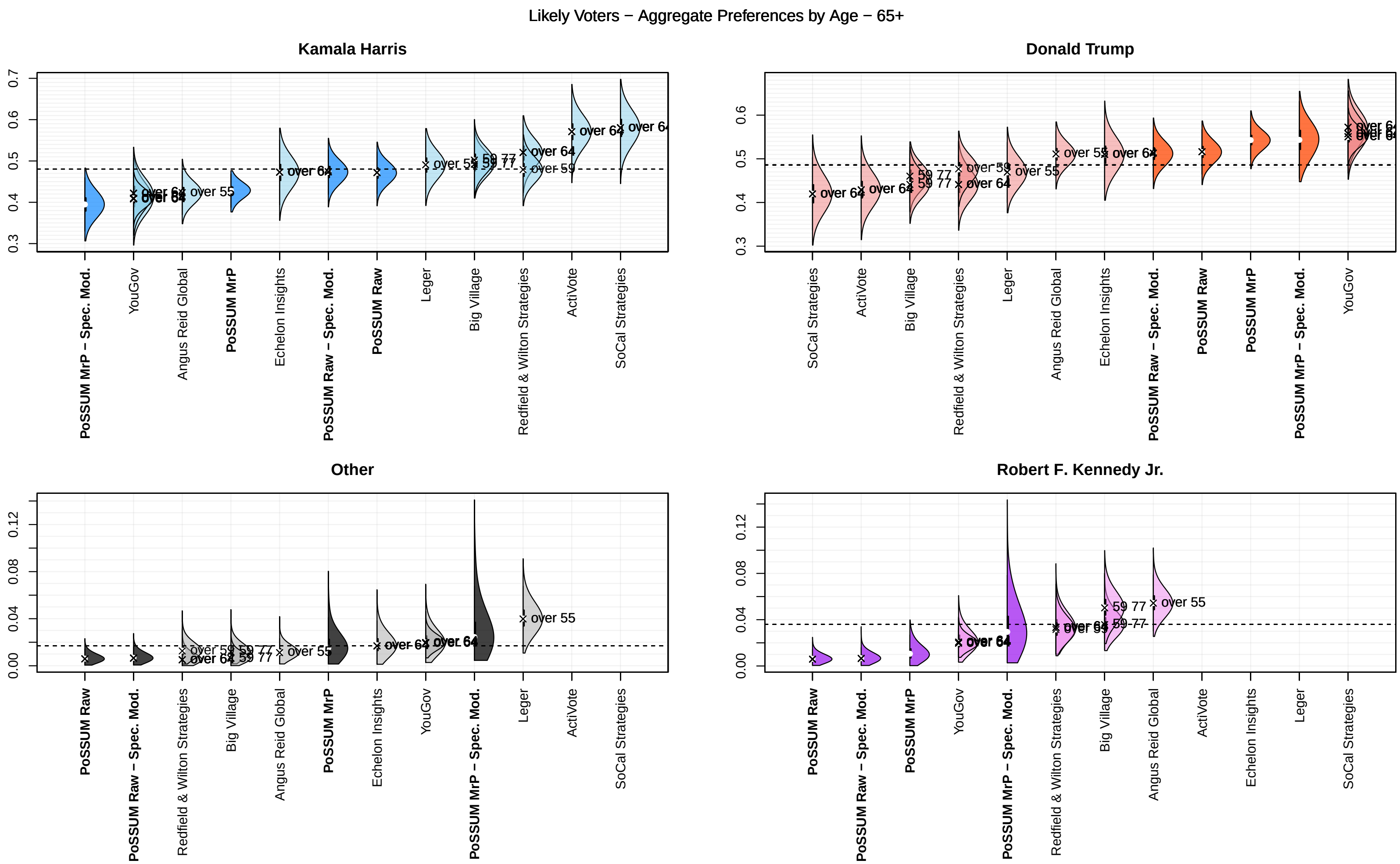


Figure 18: A comparison of the 65 years old and older distribution of 2024 vote choice across pollsters. The dotted line represents the simple arithmetic average of all polls. `PoSSUM MrP` is the primary estimate of interest. Large intervals for a given pollster correspond to small n for the given crosstab. 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.

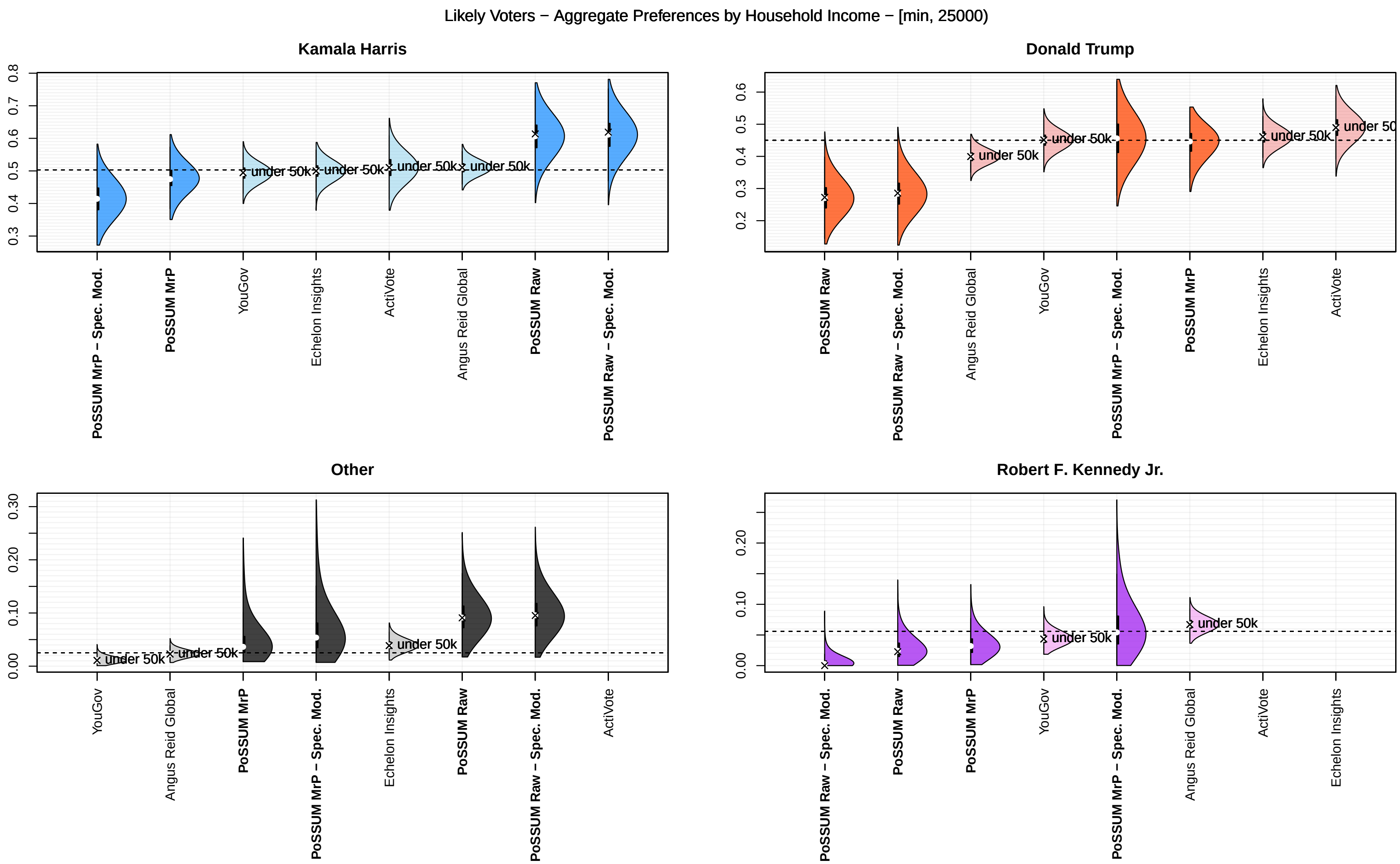


Figure 19: A comparison of the 2024 vote choice across pollsters, for voters with household income below $25k$. The dotted line represents the simple arithmetic average of all polls. `PoSSUM MrP` is the primary estimate of interest. Large intervals for a given pollster correspond to small n for the given 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.

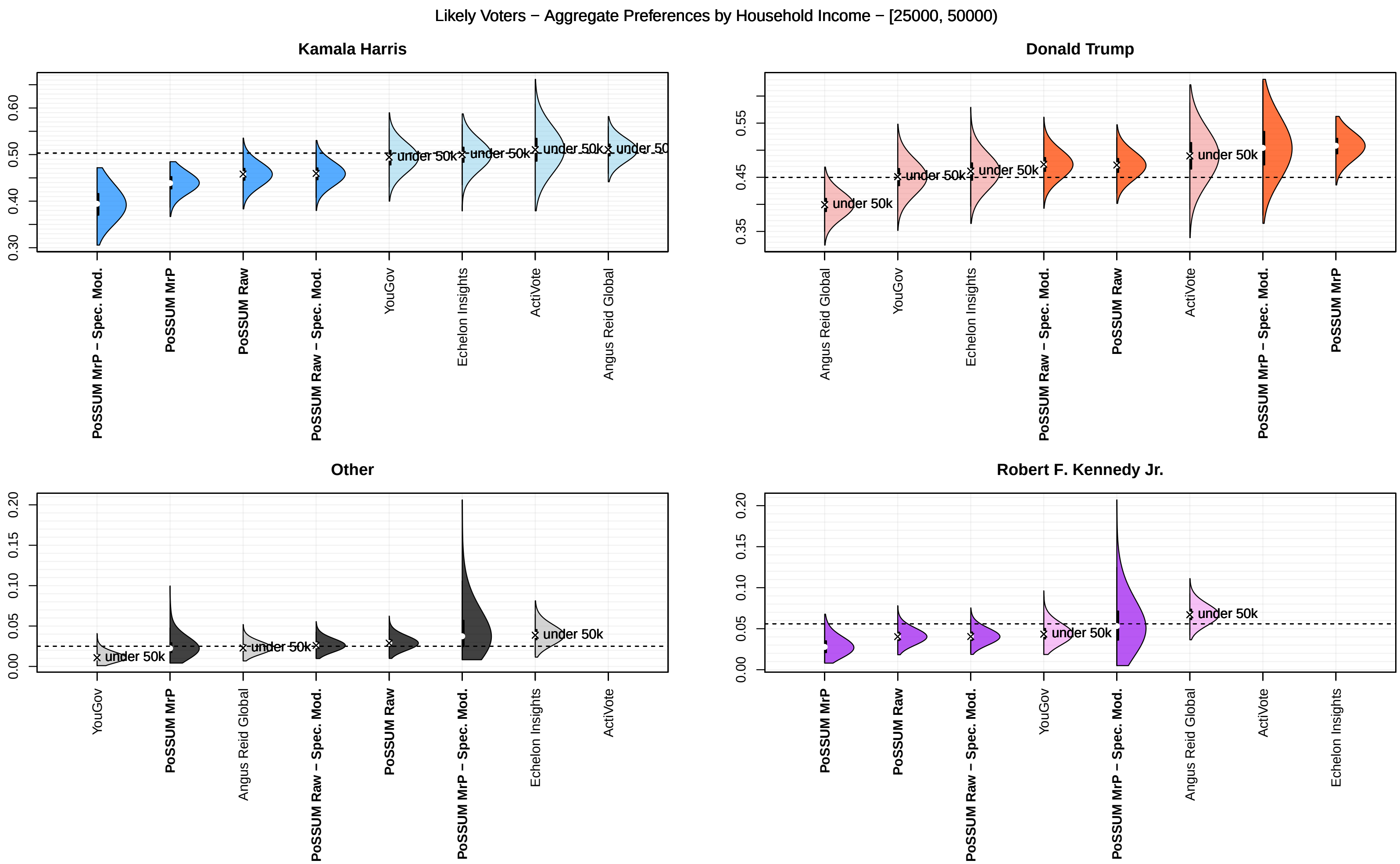


Figure 20: A comparison of the 2024 vote choice across pollsters, for voters with household income between $25k$ and $50k$. The dotted line represents the simple arithmetic average of all polls. `PoSSUM MrP` is the primary estimate of interest. Large intervals for a given pollster correspond to small 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.

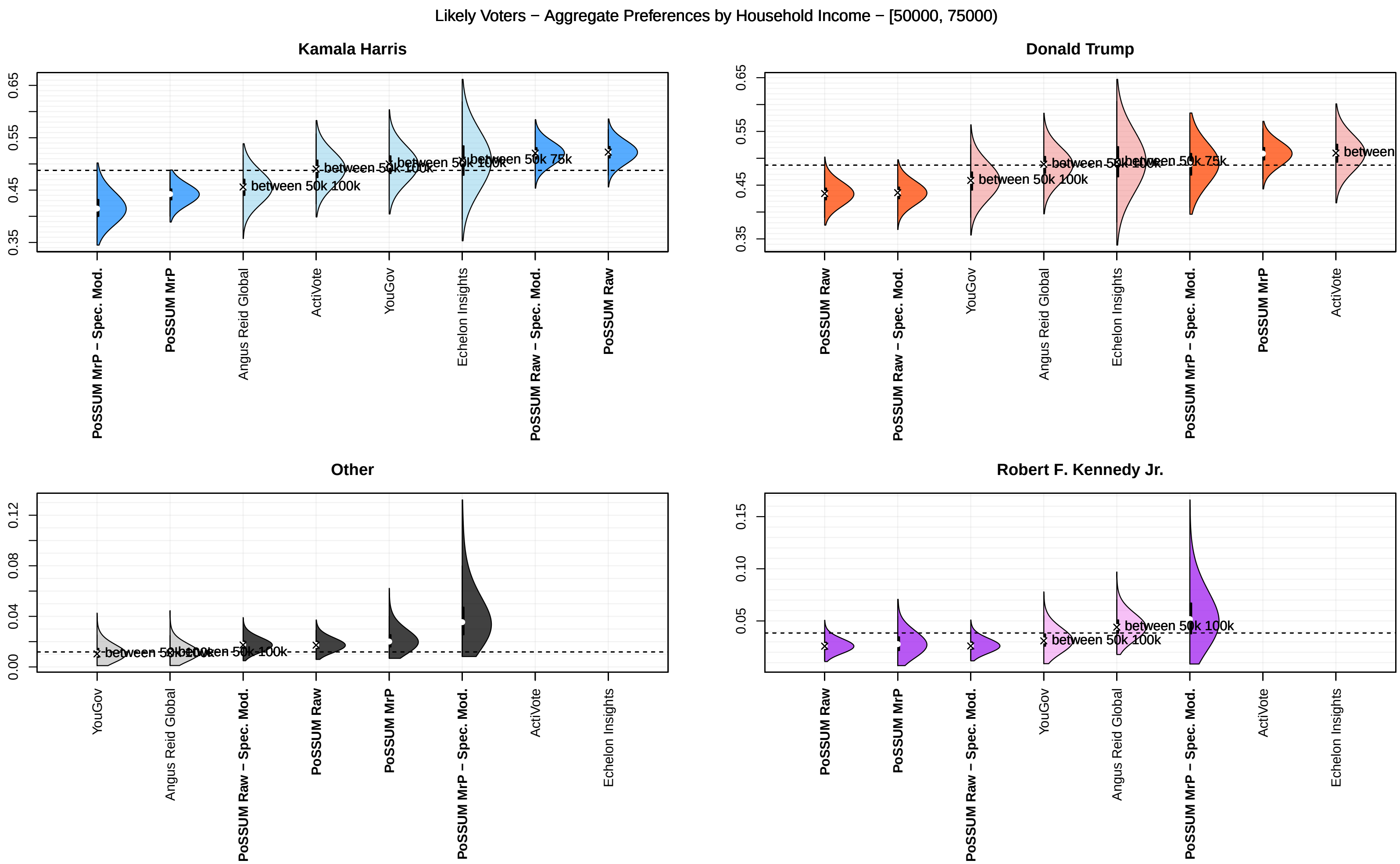


Figure 21: A comparison of the 2024 vote choice across pollsters, for voters with household income between $50k$ and $75k$. The dotted line represents the simple arithmetic average of all polls. `PoSSUM MrP` is the primary estimate of interest. Large intervals for a given pollster correspond to small 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.

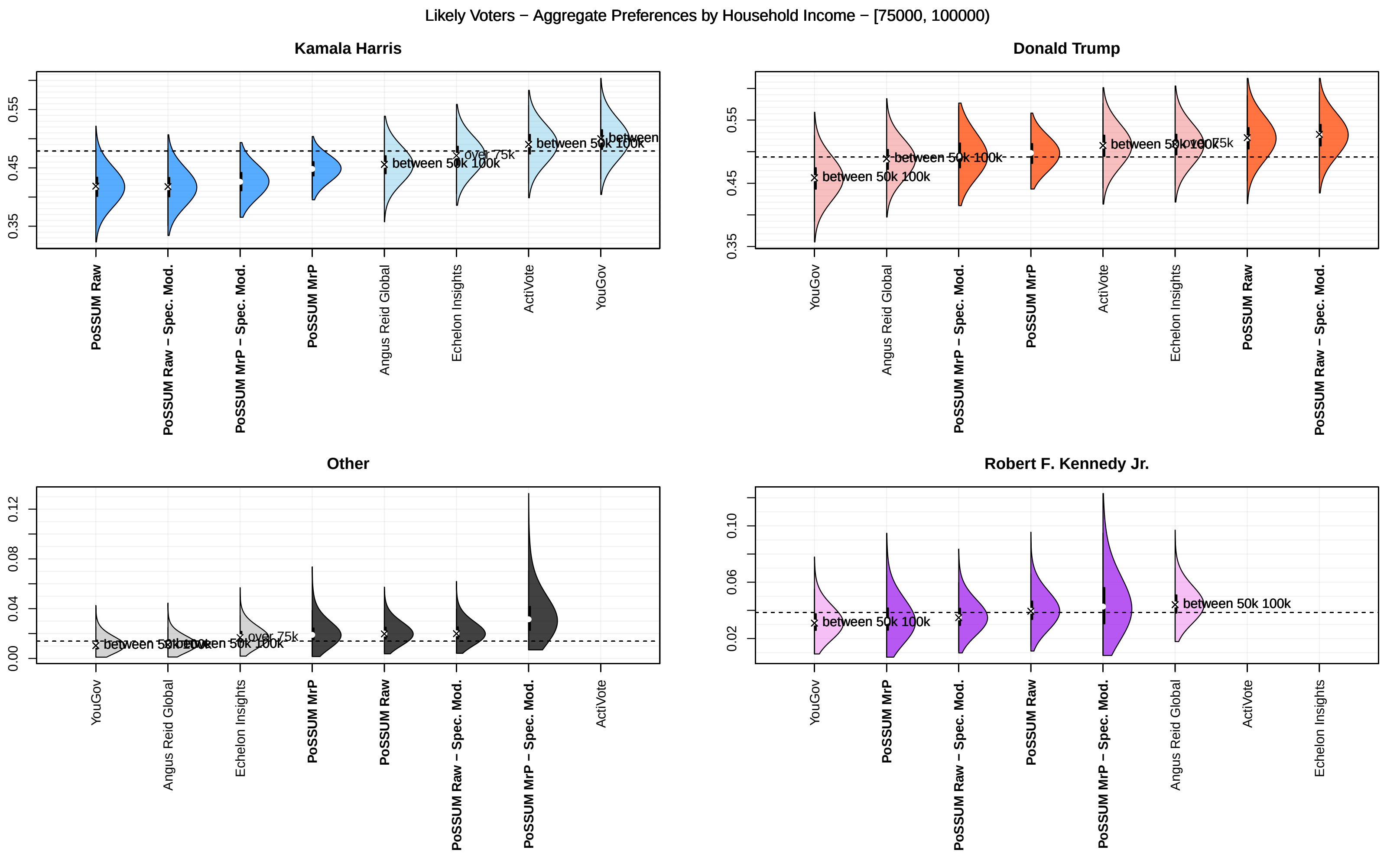


Figure 22: A comparison of the 2024 vote choice across pollsters, for voters with household income between $75k$ and $100k$. The dotted line represents the simple arithmetic average of all polls. `PoSSUM MrP` is the primary estimate of interest. Large intervals for a given pollster correspond to small 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.

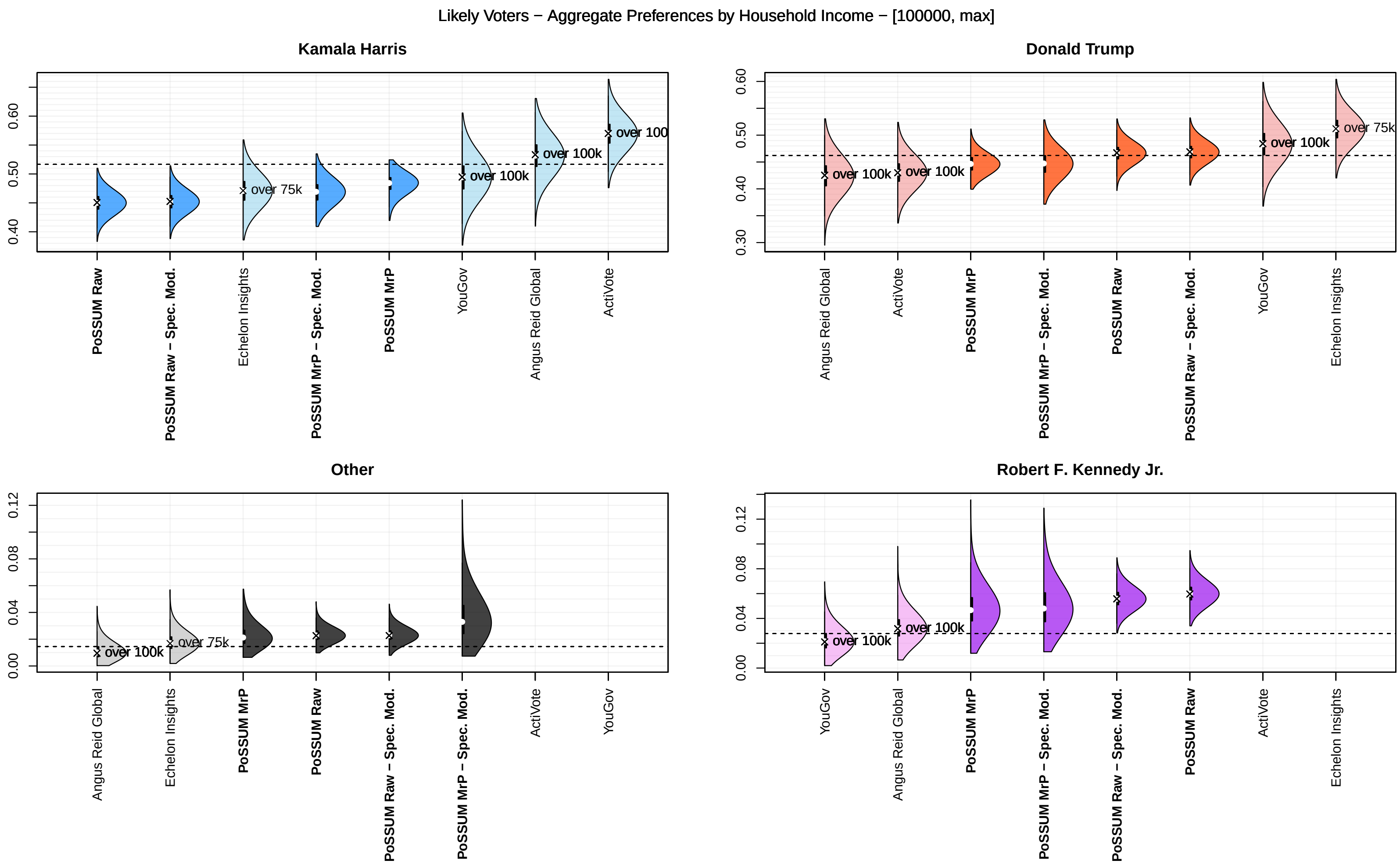


Figure 23: A comparison of the 2024 vote choice across pollsters, for voters with household income above $100k$. The dotted line represents the simple arithmetic average of all polls. `PoSSUM MrP` is the primary estimate of interest. Large intervals for a given pollster correspond to small n for the 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.

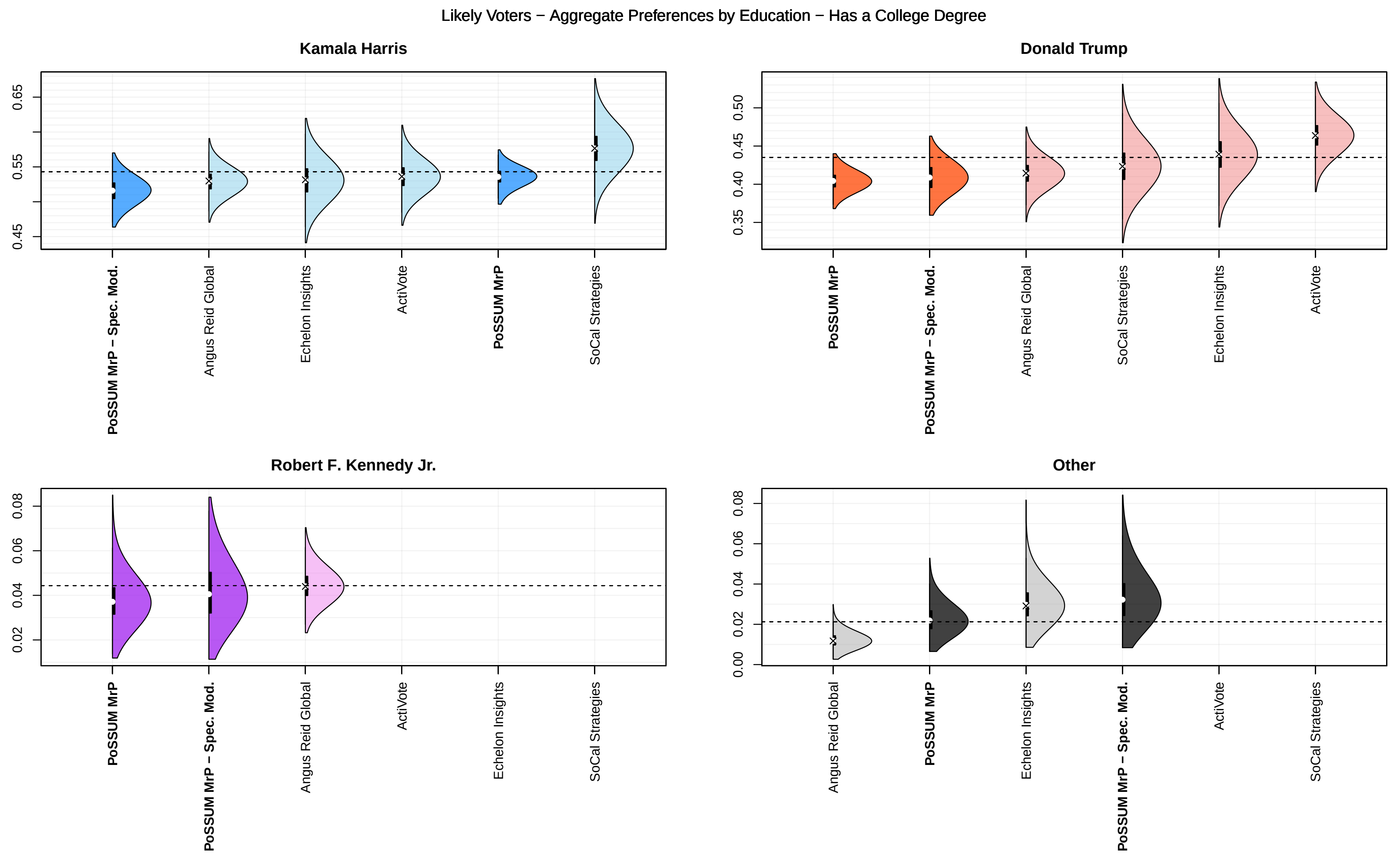


Figure 24: A comparison of the 2024 vote choice across pollsters, for voters with a college degree. The dotted line represents the simple arithmetic average of all polls. `PoSSUM MrP` is the primary estimate of interest. Large intervals for a given pollster correspond to small n for the given crosstab.

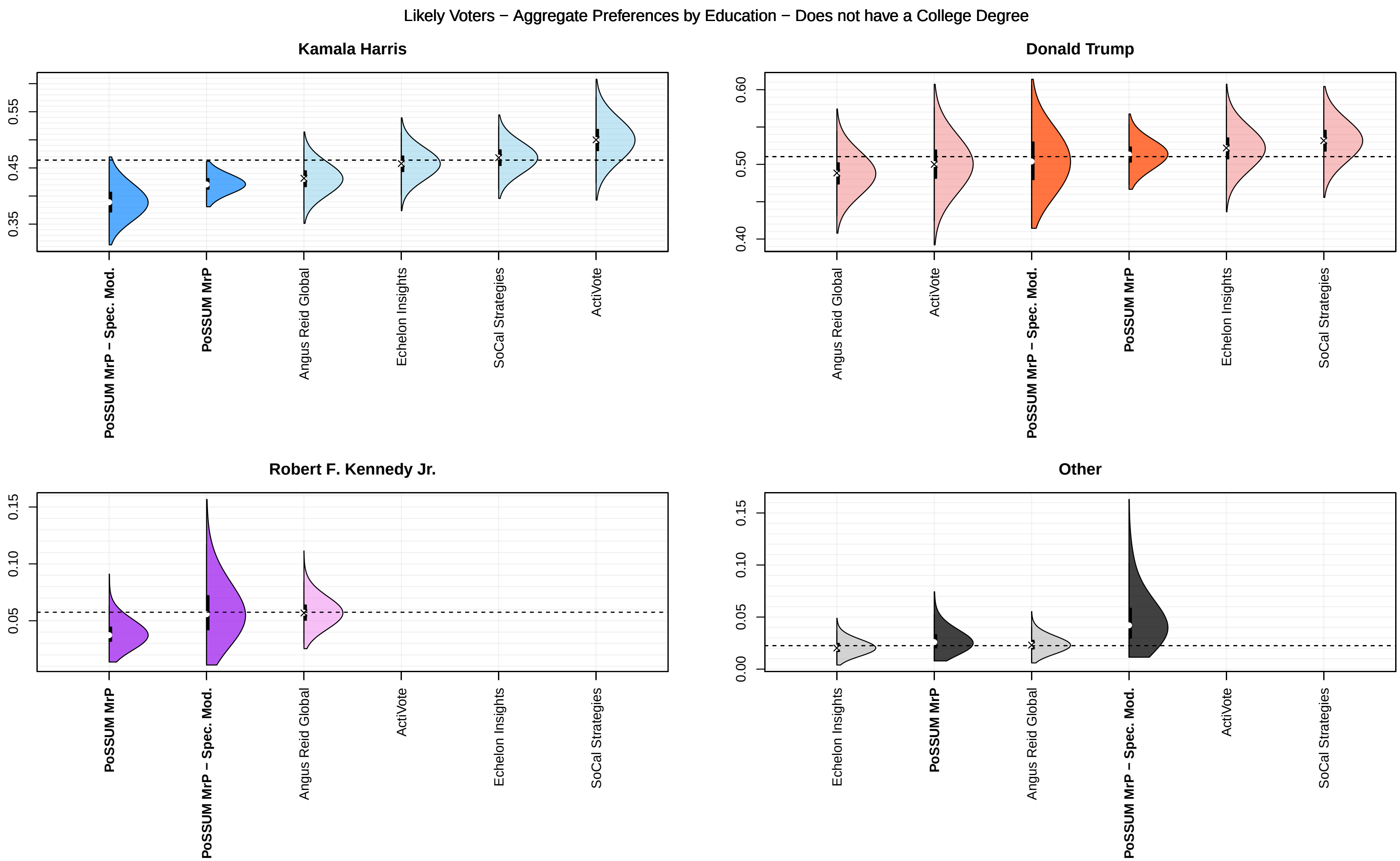


Figure 25: A comparison of the 2024 vote choice across pollsters, for voters without a college degree. The dotted line represents the simple arithmetic average of all polls. `PoSSUM MrP` is the primary estimate of interest. Large intervals for a given pollster correspond to small n for the given crosstab.

5 2^{nd} Poll: September 8^{th} - September 12^{th}

I field a PoSSUM poll starting on September 8^{th} and ending on September 23^{rd}. The quotas are tuned to target a representative sample of $1,500$ individuals, though the protocol was again stopped early, when the expected number of new acceptable users in a "fresh" pool dropped to cost-ineffective levels (see Table 13). During this time in the field I obtain valid digital traces for 940 unique users. The costs of the exercise can be broken down as follows: 180$ were spent on 𝕏 API calls to generate a subject pool via the `get_pool` routine; 280$ (upper-bound) were spent on querying user-specific timelines to get their latest tweets; 514$ (upper bound) were allocated to LLM (OpenAI API) processing costs. The total cost of the poll was 974$; the cost-per-accepted-user was 1.04$.

5.1 Subject Pool

The `get_pool` routine is executed on five occasions – typically once per day, with the exception of 11/09, where it is executed twice. This was necessary due to lower-than-usual number of valid users extracted from the subject-pool on 10/09. Table 13 presents the total number of candidate and accepted users by pool-date, broken down by the trending and political search-terms. Table 14 presents the distribution of reasons-for-exclusion over time. Figure 26 presents the distribution of trending topics amongst candidate and accepted users.

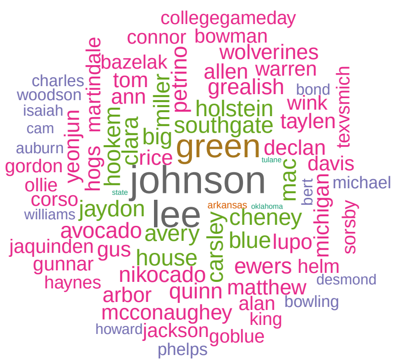
Note: the televised debate between Donald Trump and Kamala Harris took place on the 10^{th} of September – before the debate started, 88% of the sample had been collected, as can be seen in 13.

	2024-09-07	2024-09-09	2024-09-10	2024-09-11
"Fresh" Users in the Subject Pool				
Total	3,531	3,741	3,716	6,990
Politics	1,977	1,970	1,929	3,815
Trends	1,554	1,771	1,787	3,175
Accepted Users				
Total	518	235	97	116
Politics	285	135	60	53
Trends	233	100	37	63

Table 13: Breakdown of users by subject-pool type and inclusion in the survey object for each instance of the `get_pool` routine.

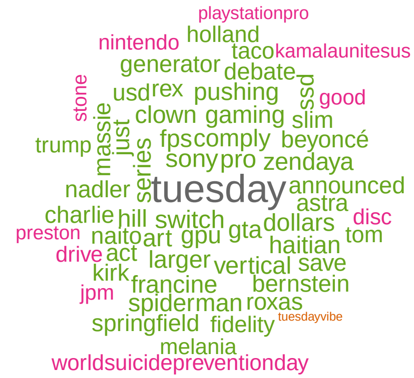
	2024-09-07	2024-09-09	2024-09-10	2024-09-11
Total				
accept	518 (14.7%)	235 (6.3%)	97 (2.6%)	116 (1.7%)
no_quota_match	260 (7.4%)	177 (4.7%)	215 (5.8%)	404 (5.8%)
not_a_person	354 (10.0%)	489 (13.1%)	496 (13.3%)	777 (11.1%)
not_from_USA	327 (9.3%)	455 (12.2%)	495 (13.3%)	1,185 (17.0%)
null_location	1,450 (41.0%)	1,531 (40.9%)	1,560 (42.0%)	2,937 (42.0%)
quota_is_full	622 (17.6%)	854 (22.8%)	853 (23.0%)	1,571 (22.5%)
Politics				
accept	285 (14.4%)	135 (6.9%)	60 (3.1%)	53 (1.4%)
no_quota_match	121 (6.1%)	90 (4.6%)	114 (5.9%)	203 (5.3%)
not_a_person	185 (9.4%)	197 (10.0%)	194 (10.1%)	397 (10.4%)
not_from_USA	159 (8.0%)	149 (7.6%)	162 (8.4%)	651 (17.1%)
null_location	910 (46.0%)	918 (46.6%)	880 (45.6%)	1,717 (45.0%)
quota_is_full	317 (16.0%)	481 (24.4%)	519 (26.9%)	794 (20.8%)
Trends				
accept	233 (15.0%)	100 (5.6%)	37 (2.1%)	63 (2.0%)
no_quota_match	139 (8.9%)	87 (4.9%)	101 (5.7%)	201 (6.3%)
not_a_person	169 (10.9%)	292 (16.5%)	302 (16.9%)	380 (12.0%)
not_from_USA	168 (10.8%)	306 (17.3%)	333 (18.6%)	534 (16.8%)
null_location	540 (34.7%)	613 (34.6%)	680 (38.0%)	1,220 (38.4%)
quota_is_full	305 (19.6%)	373 (21.1%)	334 (18.7%)	777 (24.5%)

Table 14: Breakdown of reasons for rejecting users from being included in the sample. Note that each 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.

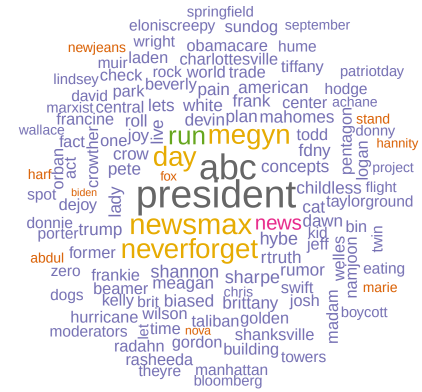


(a) Subject Pool - 07/09

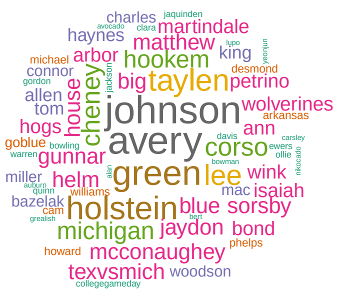
(b) Subject Pool - 09/09



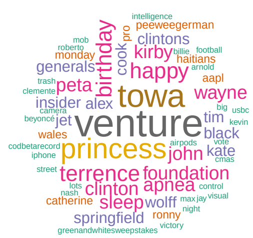
(c) Subject Pool - 10/09



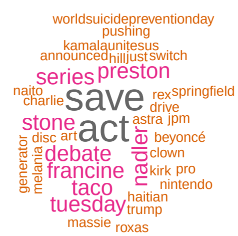
(d) Subject Pool - 11/09



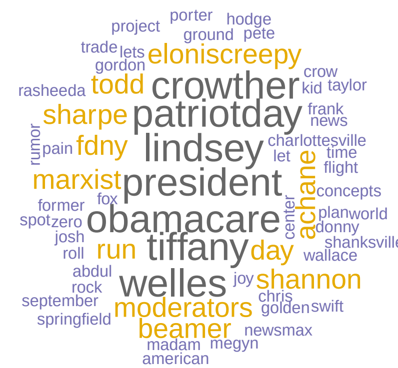
(e) Sample - 07/09



(f) Sample - 09/09



(g) Sample - 10/09



(h) Sample - 11/09

Figure 26: Word-cloud presenting words from the 'trending' queries to the 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. Fieldwork dates 09/07 to 09/12.

5.2 Unfilled Quotas

The sample frame is composed of 477 mutually exclusive cells. Figure 3 presents the marginal distributions associated with the filled quotas, against those of the unfilled quotas. Comparing the unfilled subjects with the quota targets, the missing sample is more female (Female + 14.1% relative to target); closer to the upper and lower tails of the age distribution ([18-24) + 4.3%, [65,∞) +10.3%); substantially more non-white (White -21.8%); significantly poorer (i<25k +14.9%); and finally substantially more likely to have not voted in the 2020 election (+ 14.8%).

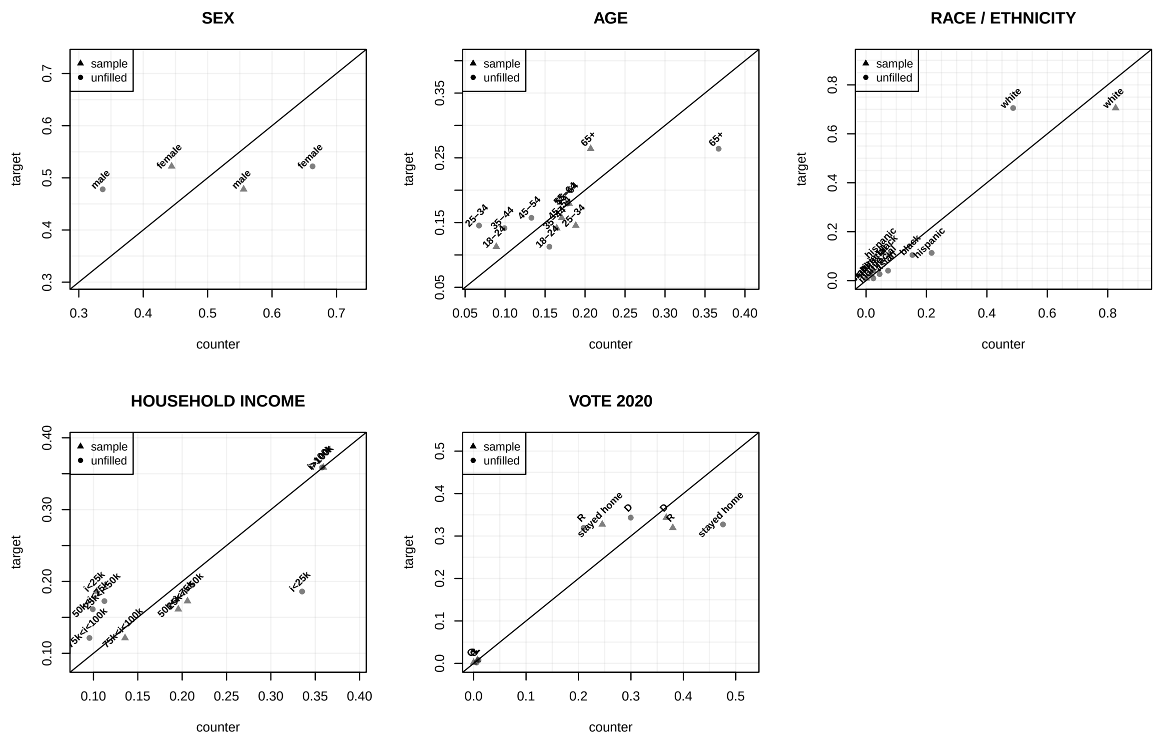


Figure 27: Comparing the filled and unfilled quotas (x-axis) against the sample-level stratification frame (y-axis). Fieldwork dates 09/07 to 09/12.

5.3 Speculation

Figure 28 presents the general distribution of speculation levels across covariates and dependent variables.

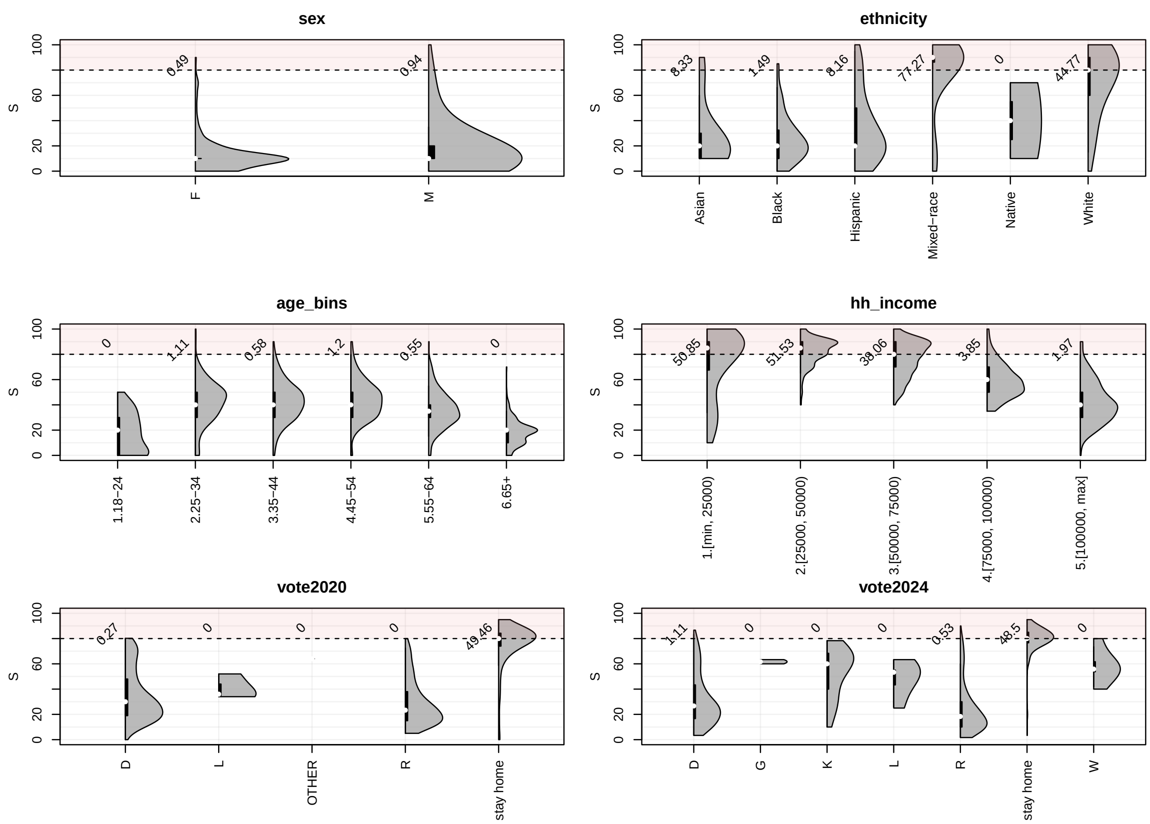


Figure 28: Distribution of the speculation score within variables used for modeling. Each distribution is accompanied by a number on the top-left, representing the % of the speculation scores which are greater than the "highly speculative" threshold. Fieldwork dates 09/07 to 09/12.

5.4 Topline & Crosstabs - 2024 Vote Choice

Table 15 presents the topline, national vote share estimates. Crosstabs by Region, Sex, Race/Ethnicity, Education, Household Income, Age, and 2020 Vote Choice are available below.

Table 15: Topline estimates of 2024 Vote Choice, 09/07 to 09/12.

Pop.	Vote2024	Topline
LV	Harris (D)	47.6 (45.4 , 50)
LV	Trump (R)	46.8 (44.4 , 49.6)
LV	RFK Jr (Ind)	3 (1.7 , 4.8)
LV	Stein (G)	0.4 (0.1 , 1)
LV	West (Ind)	0.8 (0.2 , 2.1)
LV	Oliver (L)	0.9 (0.4 , 1.7)
A	Abstention	24.6 (21.4 , 27.6)
A	Turnout	75.4 (72.4 , 78.6)

Table 16: Estimates of 2024 Vote Choice by Region, 09/07 to 09/12

Pop.	Vote2024	Midwest	Northeast	South	West
LV	Harris (D)	44.1 (41.7 , 46.5)	53.9 (50.9 , 56.9)	43.9 (41.5 , 46.6)	52.3 (49.2 , 55.3)
LV	Trump (R)	49.5 (46.8 , 52.3)	43.3 (39.8 , 46.6)	51.3 (48.4 , 54.5)	40.3 (36.8 , 43.7)
LV	RFK Jr. (Ind)	4 (2.2 , 6.7)	1.5 (0.8 , 2.9)	1.9 (1 , 3.3)	4.6 (2.4 , 7.7)
LV	Stein (G)	0.2 (0 , 0.7)	0.3 (0.1 , 1)	0.4 (0.1 , 1.4)	0.3 (0.1 , 1.1)
LV	West (Ind)	1 (0.3 , 2.5)	0.2 (0 , 0.4)	1 (0.3 , 2.5)	1 (0.3 , 2.3)
LV	Oliver (L)	0.7 (0.3 , 1.5)	0.7 (0.3 , 1.5)	0.8 (0.3 , 1.9)	1.1 (0.4 , 2.4)
A	Abstention	23.2 (19.9 , 26.4)	25.3 (20.2 , 29)	26.1 (22.8 , 29.4)	23.3 (19.4 , 26.9)
A	Turnout	76.8 (73.6 , 80.1)	74.7 (71 , 79.8)	73.9 (70.6 , 77.2)	76.7 (73.1 , 80.6)

Table 17: Estimates of 2024 Vote Choice by Sex, 09/07 to 09/12

Pop.	Vote2024	F	M
LV	Harris (D)	52.1 (49.2 , 55.1)	42.6 (40 , 45.3)
LV	Trump (R)	43.1 (40.3 , 46.4)	51.1 (48.1 , 54.3)
LV	RFK Jr. (Ind)	2.4 (1 , 4.6)	3.5 (2 , 5.7)
LV	Stein (G)	0.5 (0.1 , 1.6)	0.2 (0 , 0.8)
LV	West (Ind)	0.9 (0.2 , 2.3)	0.7 (0.2 , 2)
LV	Oliver (L)	0.4 (0 , 1.2)	1.3 (0.6 , 2.7)
A	Abstention	22.1 (17.8 , 25.9)	27.4 (24 , 30.2)
A	Turnout	77.9 (74.1 , 82.2)	72.6 (69.8 , 76)

Table 18: Estimates of 2024 Vote Choice by Race/Ethnicity, 09/07 to 09/12

Pop.	Vote2024	Asian	Black	Hispanic	Mixed-Race	Native	White
LV	Harris (D)	67.4 (59.4 , 75.3)	80 (73.9 , 85)	61 (53.5 , 67.1)	47.7 (39.4 , 57.2)	44.3 (30.9 , 56)	41.1 (38.9 , 43.5)
LV	Trump (R)	24.6 (14.3 , 33.5)	11.6 (6.6 , 17.2)	33.9 (27.6 , 42)	40.1 (23.1 , 52.5)	41.8 (22.2 , 57.8)	54.2 (51.7 , 57.1)
LV	RFK Jr. (Ind)	4.6 (0.9 , 11.7)	4.2 (1.8 , 8.4)	2.7 (0.5 , 5.7)	6 (2 , 16.8)	6.9 (2 , 25.5)	2.5 (1.3 , 4.3)
LV	Stein (G)	0.4 (0.1 , 2.4)	0.6 (0.1 , 2.2)	0.4 (0 , 2.2)	0.5 (0.1 , 2.7)	0.5 (0.1 , 4.5)	0.2 (0.1 , 0.8)
LV	West (Ind)	0.6 (0.1 , 1.9)	1.5 (0.4 , 4.4)	0.5 (0.1 , 1.6)	1.3 (0.2 , 4.1)	1.6 (0.3 , 5.9)	0.8 (0.2 , 1.9)
LV	Oliver (L)	1.2 (0.3 , 3.9)	1 (0.2 , 3.2)	0.9 (0.2 , 2.4)	2.1 (0.5 , 6.2)	1.3 (0.2 , 5.8)	0.7 (0.3 , 1.5)
A	Abstention	23 (13.6 , 30.3)	31 (24 , 37)	32.5 (24.9 , 39.1)	28.5 (20.7 , 38.6)	30.7 (17.2 , 42.4)	22.6 (19.4 , 25.7)
A	Turnout	77 (69.7 , 86.4)	69 (63 , 76)	67.5 (60.9 , 75.1)	71.5 (61.4 , 79.3)	69.3 (57.6 , 82.8)	77.4 (74.3 , 80.6)

Table 19: Estimates of 2024 Vote Choice by College Grad., 09/07 to 09/12

Pop.	Vote2024	Does not Have a College Degree	Has a College Degree
LV	Harris (D)	44 (41.2 , 47.1)	54.3 (51.9 , 56.2)
LV	Trump (R)	49.8 (46.9 , 53.2)	41.4 (39.5 , 43.9)
LV	RFK Jr. (Ind)	3.3 (1.9 , 5.4)	2.3 (1.3 , 3.9)
LV	Stein (G)	0.4 (0.1 , 1.2)	0.3 (0.1 , 0.8)
LV	West (Ind)	1 (0.3 , 2.6)	0.5 (0.1 , 1.3)
LV	Oliver (L)	0.9 (0.4 , 1.8)	0.8 (0.4 , 1.7)
A	Abstention	29.1 (25.4 , 32.4)	15.1 (12.5 , 17.6)
A	Turnout	70.9 (67.6 , 74.6)	84.9 (82.4 , 87.5)

Table 20: Estimates of 2024 Vote Choice by Houshold Income, 09/07 to 09/12

Pop.	Vote2024	[0, 25k)	[25k, 50k)	[50k, 75k)	[75k, 100k)	[100k, +)
LV	Harris (D)	53.2 (46.2 , 61.9)	46.1 (41.8 , 51.3)	44.6 (41.2 , 48.1)	44.5 (40.4 , 48)	48.4 (45.8 , 51.5)
LV	Trump (R)	39 (28.6 , 47)	47.4 (42 , 52.5)	51.2 (47.2 , 54.8)	51.7 (47.9 , 56)	46.5 (43.1 , 49.7)
LV	RFK Jr. (Ind)	3.1 (0.7 , 7.4)	3.3 (1.5 , 5.9)	2.3 (1.2 , 4.5)	2 (0.8 , 4)	3.1 (1.5 , 5.7)
LV	Stein (G)	0.6 (0.1 , 3.6)	0.4 (0.1 , 1.2)	0.3 (0.1 , 0.9)	0.2 (0 , 0.8)	0.2 (0 , 0.7)
LV	West (Ind)	1.7 (0.4 , 5)	1.2 (0.3 , 3.3)	0.6 (0.1 , 1.5)	0.4 (0.1 , 1.2)	0.5 (0.1 , 1.6)
LV	Oliver (L)	1.1 (0.2 , 3.2)	0.8 (0.3 , 1.8)	0.7 (0.2 , 1.6)	0.7 (0.2 , 1.6)	0.8 (0.3 , 2)
A	Abstention	39.9 (31.1 , 48.3)	27.1 (21.4 , 32.6)	24.7 (20.6 , 29)	24.5 (19.2 , 29.5)	15.8 (12.1 , 19.5)
A	Turnout	60.1 (51.7 , 68.9)	72.9 (67.4 , 78.6)	75.3 (71 , 79.4)	75.5 (70.5 , 80.8)	84.2 (80.5 , 87.9)

Table 21: Estimates of 2024 Vote Choice by Age, 09/07 to 09/12

Pop.	Vote2024	18-24	25-34	35-44	45-54	55-64	65+
LV	Harris (D)	61.7 (56.3 , 67.5)	53.5 (49.7 , 57.7)	51.6 (47.6 , 54.9)	45.7 (41.9 , 49)	44.4 (41.7 , 47.3)	43.5 (40.3 , 46.9)
LV	Trump (R)	32.3 (25.4 , 37.9)	39.9 (35.7 , 44.2)	42.3 (39.1 , 47)	48.5 (44.6 , 52.8)	50.5 (47.1 , 53.6)	52 (48.1 , 56)
LV	RFK Jr. (Ind)	2.7 (0.8 , 6.1)	3.2 (1.6 , 5.5)	3.6 (1.9 , 6.4)	3.4 (1.7 , 5.9)	2.9 (1.4 , 5.2)	2.3 (0.9 , 4.5)
LV	Stein (G)	0.5 (0.1 , 1.9)	0.4 (0.1 , 1.5)	0.3 (0.1 , 1.1)	0.3 (0.1 , 1.2)	0.3 (0.1 , 0.9)	0.3 (0 , 1.1)
LV	West (Ind)	0.9 (0.2 , 2.7)	0.9 (0.2 , 2.4)	0.7 (0.2 , 1.9)	0.8 (0.2 , 2.1)	0.8 (0.2 , 2.3)	0.7 (0.2 , 2.3)
LV	Oliver (L)	1.3 (0.3 , 3.9)	1.3 (0.4 , 3.4)	0.8 (0.3 , 1.9)	0.8 (0.3 , 2)	0.7 (0.2 , 1.7)	0.5 (0.1 , 1.5)
A	Abstention	47.4 (41.1 , 54.9)	40.1 (35.7 , 44.7)	26.1 (21.7 , 31)	20.7 (15.7 , 25.3)	15.4 (10.2 , 20.5)	15 (8.7 , 20.3)
A	Turnout	52.6 (45.1 , 58.9)	59.9 (55.3 , 64.3)	73.9 (69 , 78.3)	79.3 (74.7 , 84.3)	84.6 (79.5 , 89.8)	85 (79.7 , 91.3)

Table 22: Estimates of 2024 Vote Choice by 2020 Vote Choice, 09/07 to 09/12

Pop.	Vote2024	D	R	L	G	OTHER	Stay Home
LV	Harris (D)	96 (93.8 , 97.5)	1.1 (0.4 , 2.4)	23 (6.1 , 56)	20.3 (1.6 , 62.4)	22.3 (5 , 54.8)	40 (27.5 , 53.8)
LV	Trump (R)	1.7 (0.8 , 3.1)	95.9 (93.5 , 97.7)	32.7 (7.3 , 71.1)	30.6 (1.3 , 93.2)	22.1 (0.8 , 73.3)	41.6 (29.2 , 57.6)
LV	RFK Jr. (Ind)	1 (0.4 , 2.1)	1.6 (0.7 , 3.2)	4 (0.2 , 26.5)	6.3 (0.3 , 48.9)	5 (0.1 , 39.3)	11.3 (5.6 , 19.5)
LV	Stein (G)	0.4 (0.1 , 1.2)	0.1 (0 , 0.6)	0.9 (0 , 14.5)	0.5 (0 , 21.5)	0.6 (0 , 10.9)	0.5 (0 , 2.8)
LV	West (Ind)	0.3 (0 , 1)	0.2 (0 , 0.8)	1.9 (0.1 , 17.4)	2.1 (0 , 23.2)	21 (1.7 , 47.7)	3.1 (0.7 , 8.8)
LV	Oliver (L)	0.4 (0.1 , 1.3)	0.7 (0.2 , 1.8)	19.7 (2.4 , 55.2)	4.4 (0 , 63.8)	4.3 (0.1 , 38.6)	0.5 (0 , 2.6)
A	Abstention	7.4 (5.2 , 10)	2.3 (1.1 , 4)	10.9 (0.8 , 42.3)	13.7 (0.3 , 74.1)	12.8 (0.9 , 56.9)	64.7 (55.5 , 72.6)
A	Turnout	92.6 (90 , 94.8)	97.7 (96 , 98.9)	89.1 (57.7 , 99.2)	86.3 (25.9 , 99.7)	87.2 (43.1 , 99.1)	35.3 (27.4 , 44.5)

5.5 Cross-pollsters comparison - 2024 Vote Choice

`PoSSUM MrP`'s estimates for this poll seem exchangeable with any other poll in the campaign, with the exception again of lower-than-average % of ‘switchers’ between the two major parties, relative to the 2020 vote.

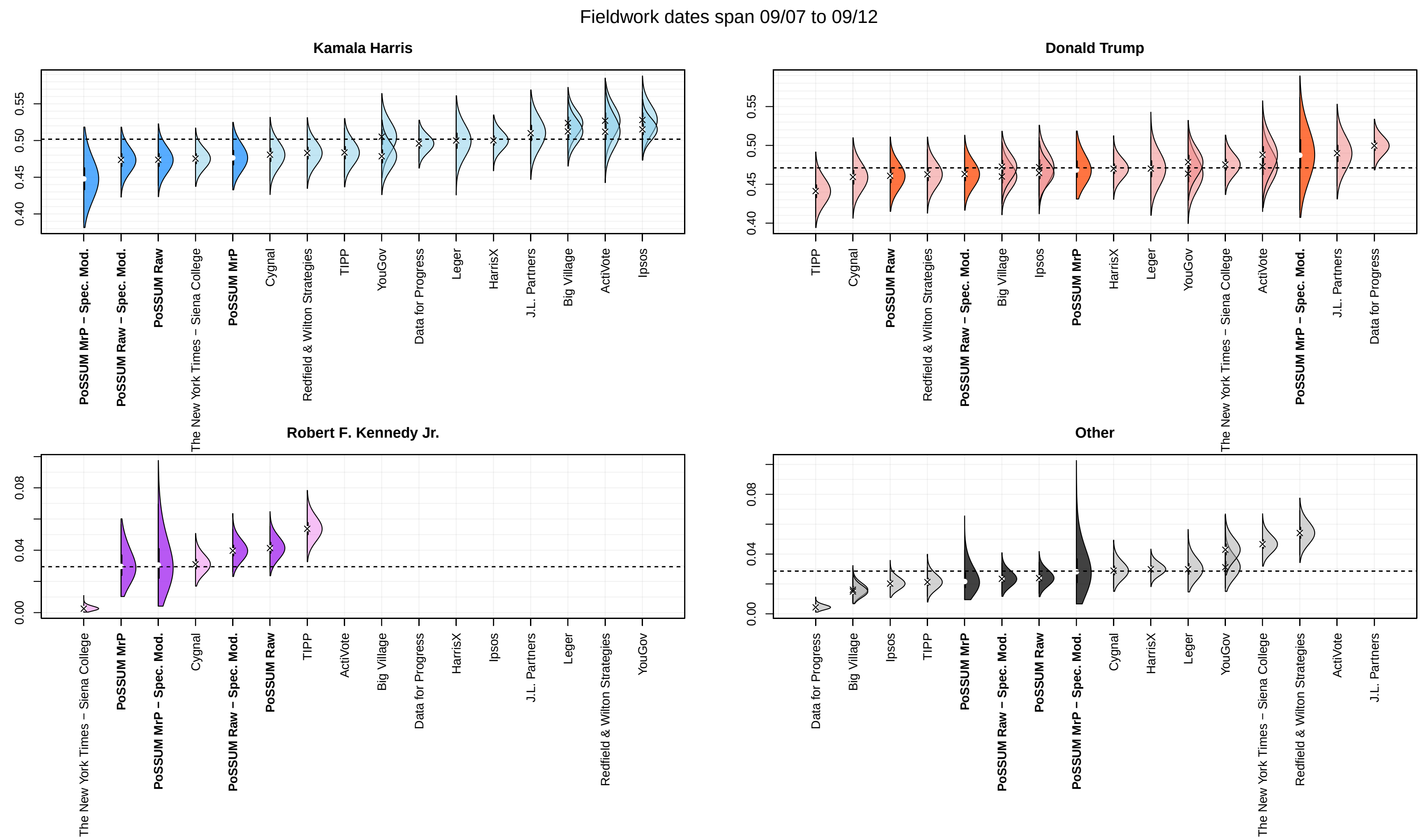


Figure 29: A comparison of the topline distribution of 2024 vote choice across pollsters. The dotted line represents the simple arithmetic average of all polls. `PoSSUM MrP` is the primary estimate of interest.

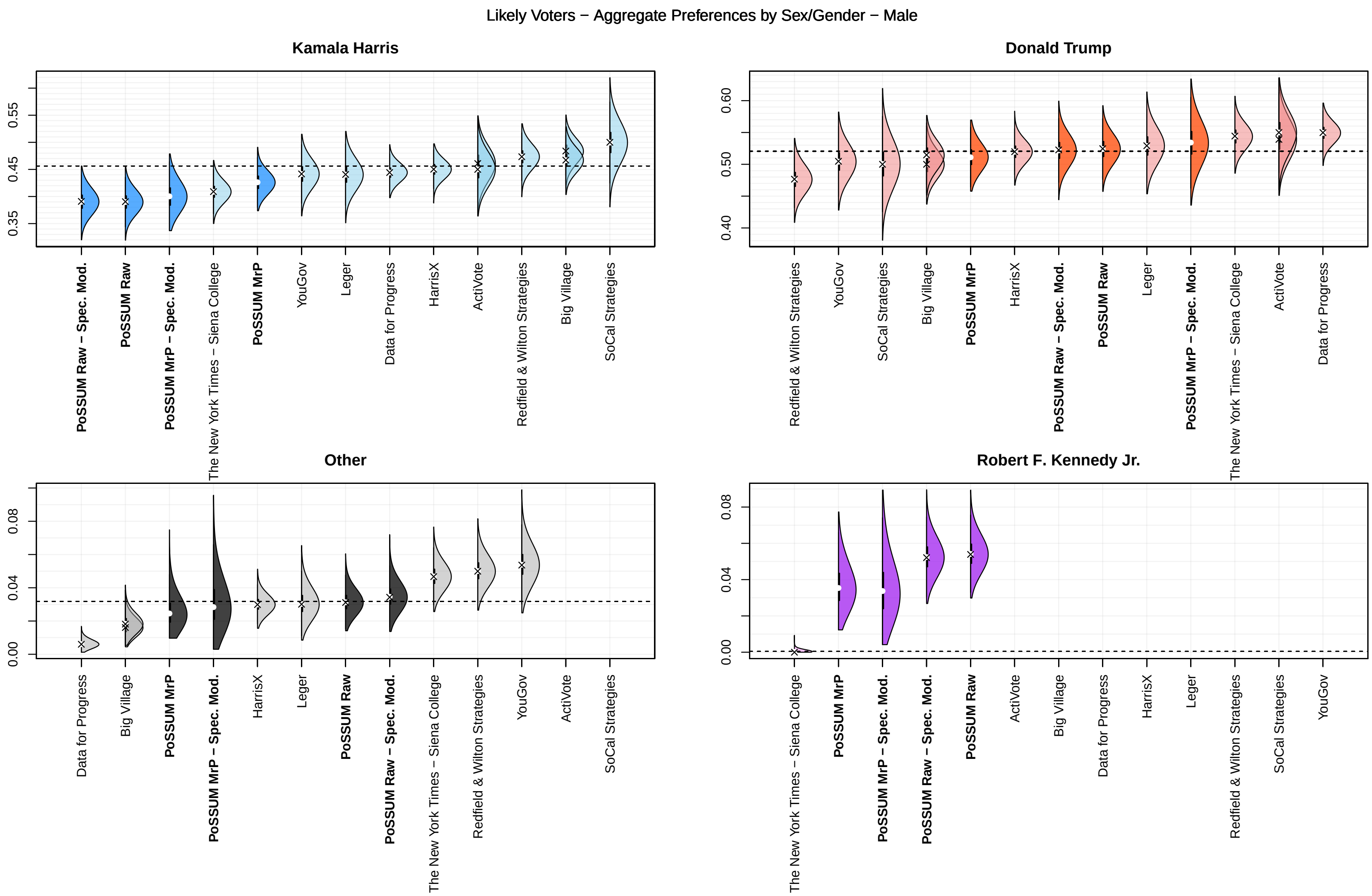


Figure 30: A comparison of the Male distribution of 2024 vote choice across pollsters. The dotted line represents the simple arithmetic average of all polls. `PoSSUM MrP` is the primary estimate of interest. Large intervals for a given pollster correspond to small n for the given crosstab.

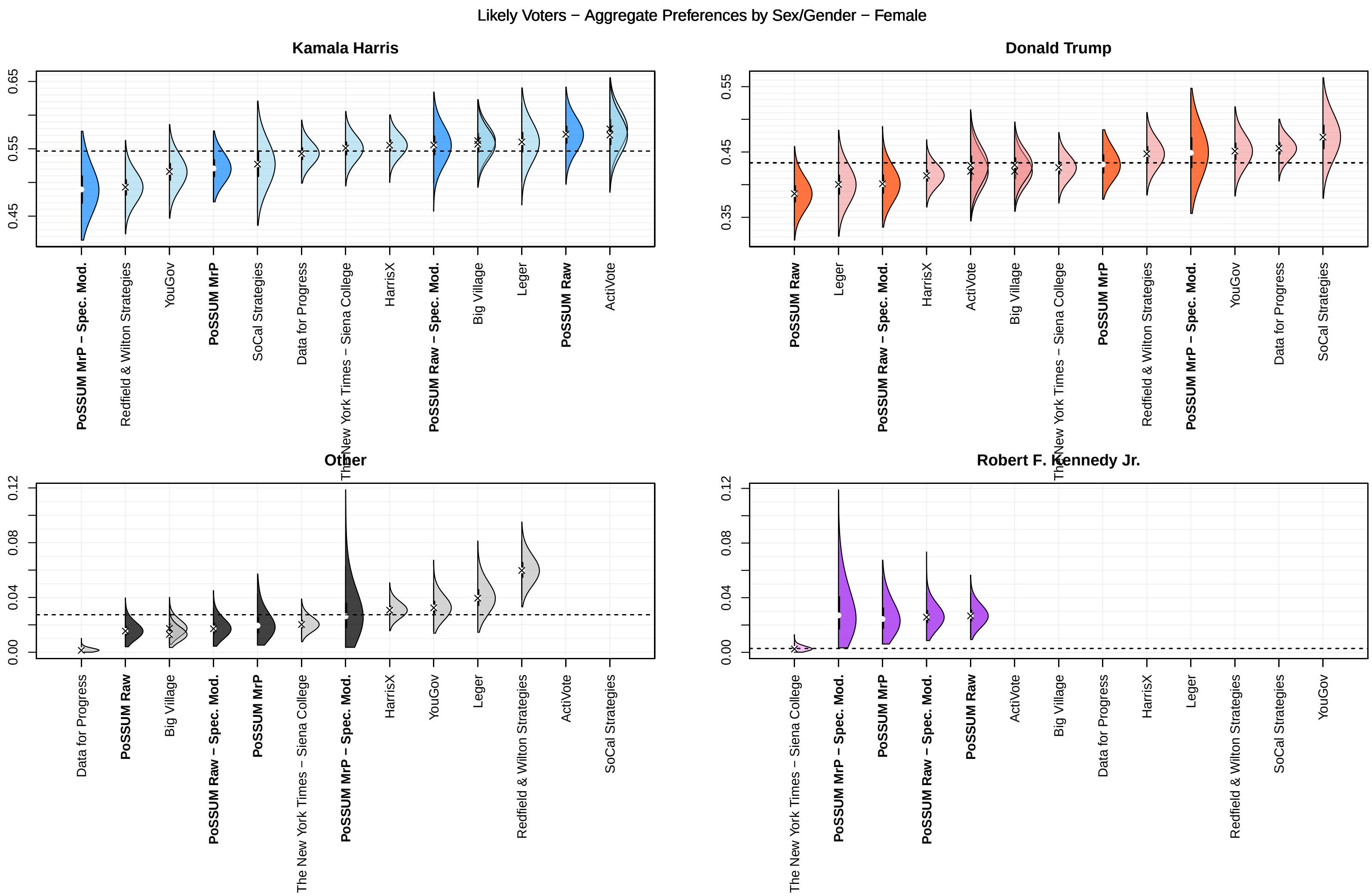


Figure 31: A comparison of the Female distribution of 2024 vote choice across pollsters. The dotted line represents the simple arithmetic average of all polls. `PoSSUM MrP` is the primary estimate of interest. `PoSSUM MrP` is the primary estimate of interest. Large intervals for a given pollster correspond to small n for the given crosstab

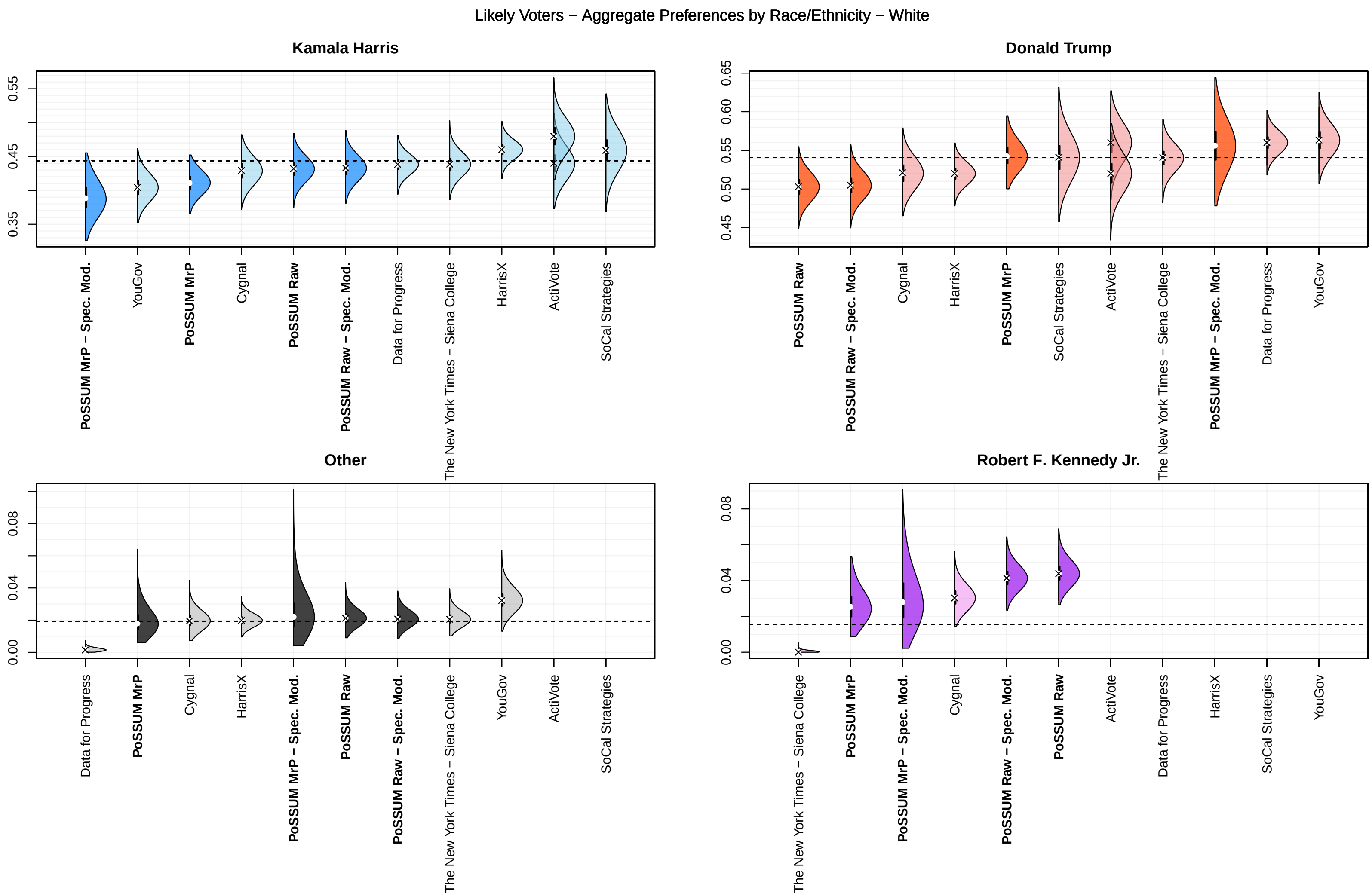


Figure 32: A comparison of the White distribution of 2024 vote choice across pollsters. The dotted line represents the simple arithmetic average of all polls. `PoSSUM MrP` is the primary estimate of interest. Large intervals for a given pollster correspond to small n for the given crosstab

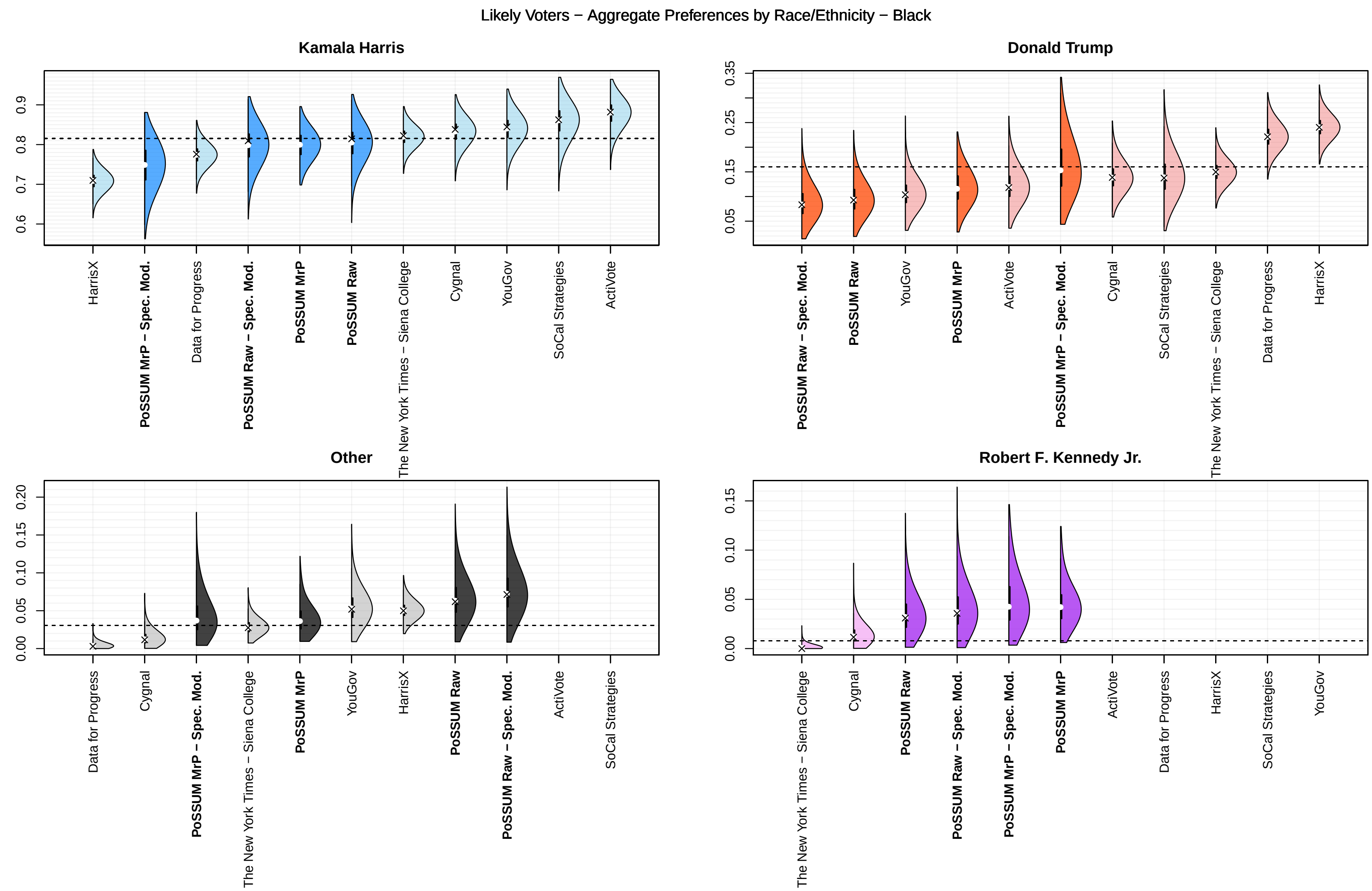


Figure 33: A comparison of the Black distribution of 2024 vote choice across pollsters. The dotted line represents the simple arithmetic average of all polls. `PoSSUM MrP` is the primary estimate of interest. Large intervals for a given pollster correspond to small n for the given crosstab

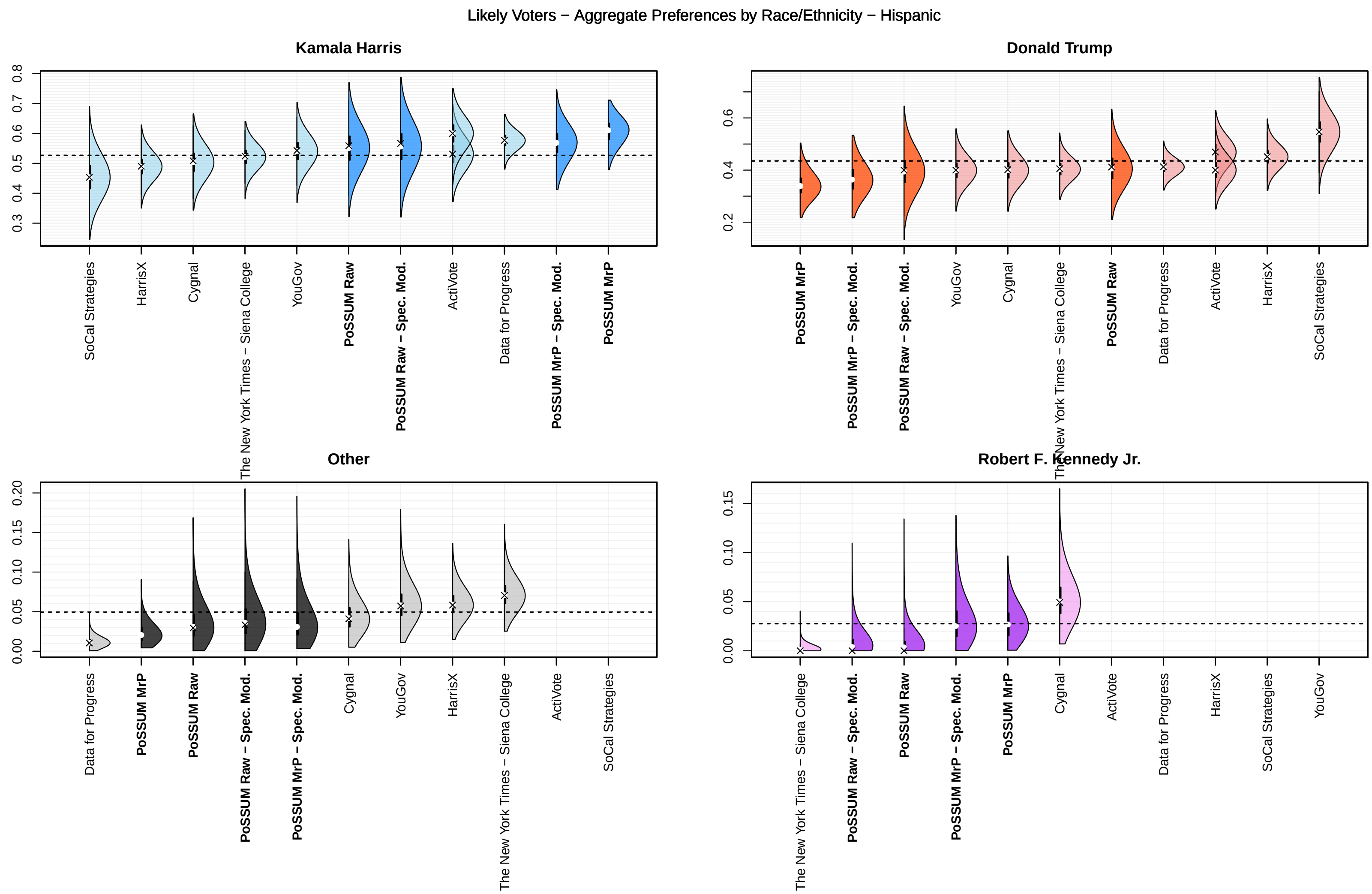


Figure 34: A comparison of the Hispanic distribution of 2024 vote choice across pollsters. The dotted line represents the simple arithmetic average of all polls. `PoSSUM MrP` is the primary estimate of interest. Large intervals for a given pollster correspond to small n for the given crosstab.

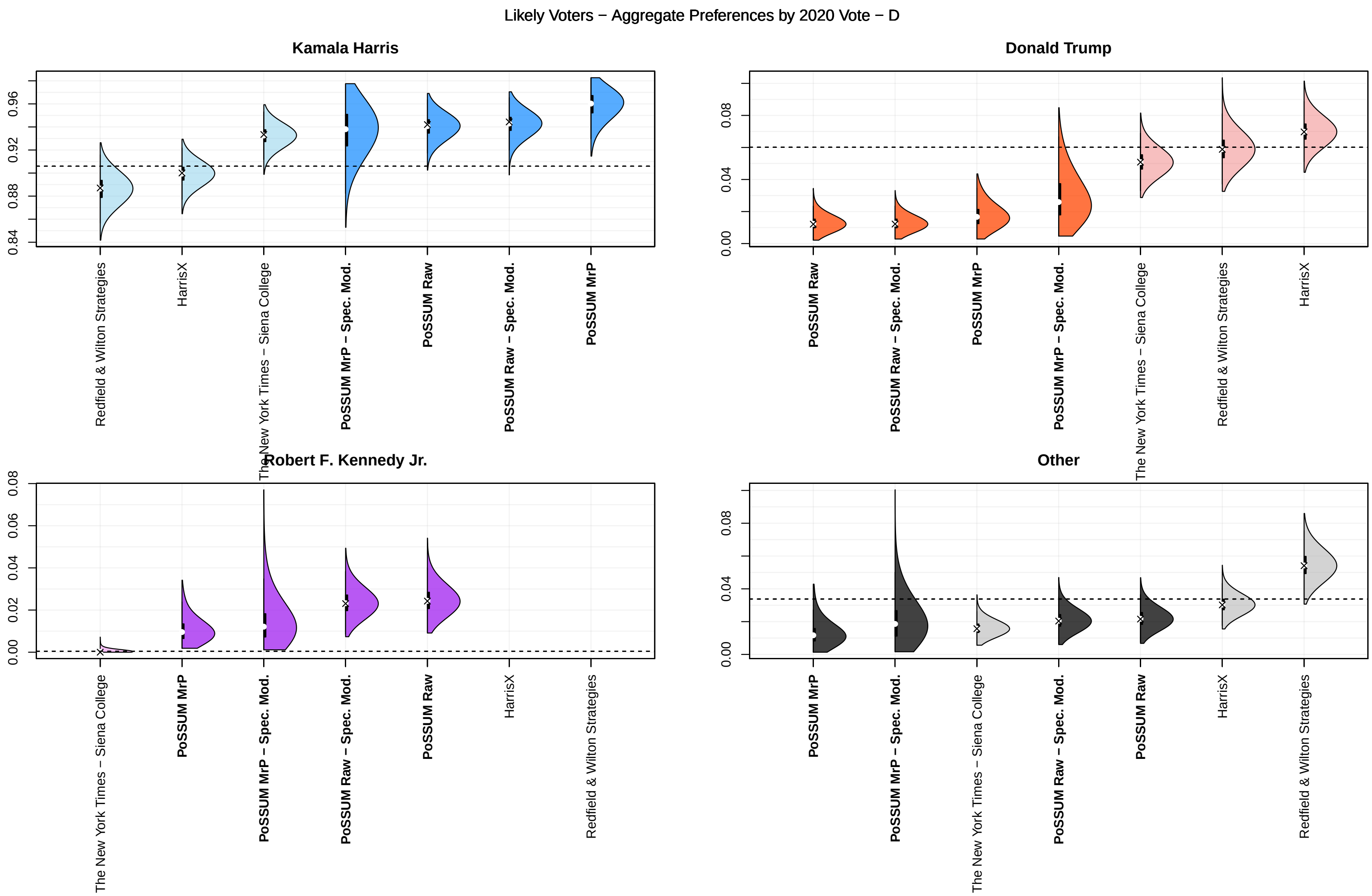


Figure 35: A comparison of the distribution of 2024 vote choice across pollsters, for those who voted for Joe Biden (D) in 2020. The dotted line represents the simple arithmetic average of all polls. `PoSSUM MrP` is the primary estimate of interest. Large intervals for a given pollster correspond to small n for the given crosstab.

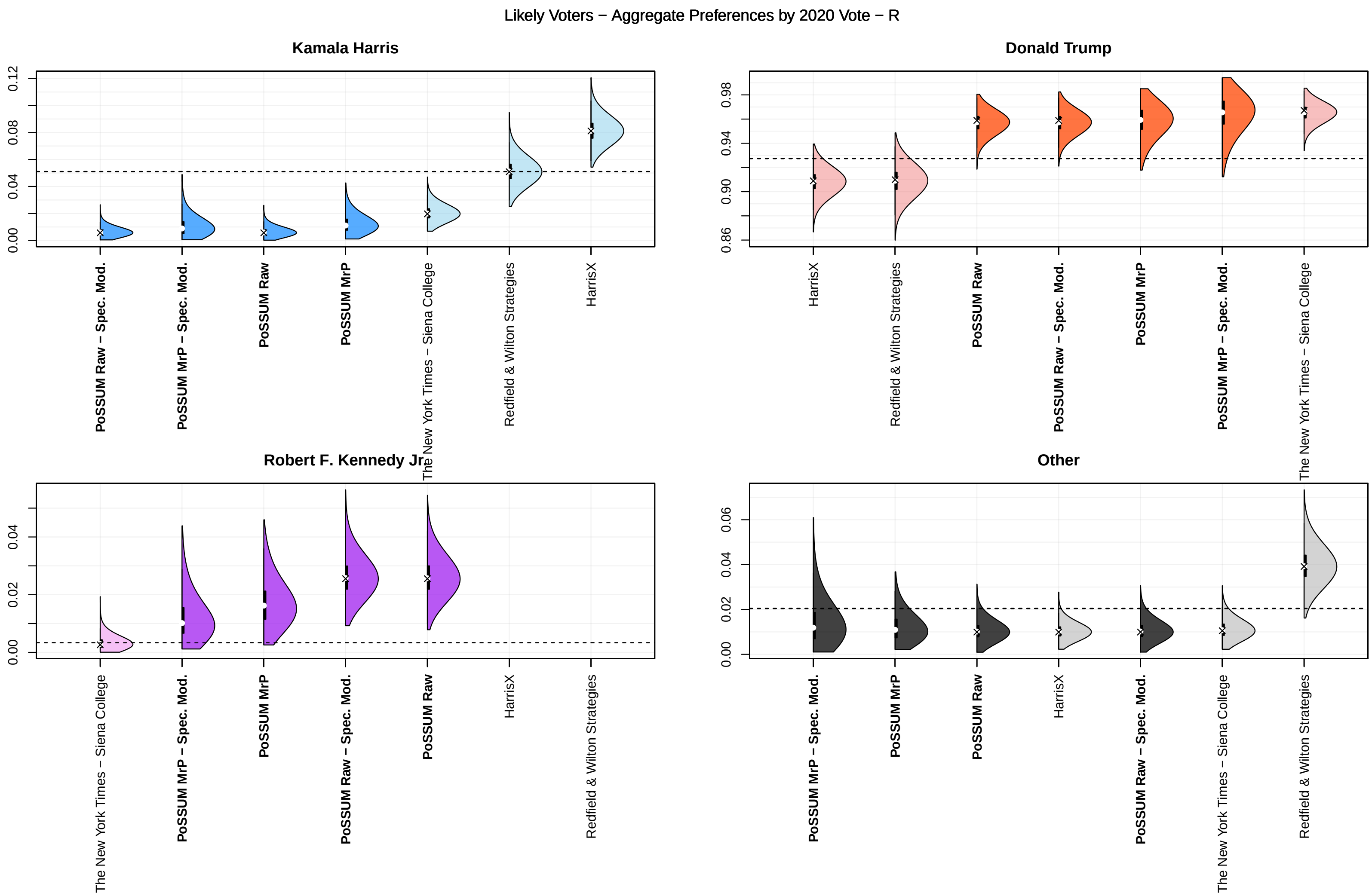


Figure 36: A comparison of the distribution of 2024 vote choice across pollsters, for those who voted for Donlad Trump (R) in 2020. The dotted line represents the simple arithmetic average of all polls. `PoSSUM MrP` is the primary estimate of interest. Large intervals for a given pollster correspond to small n for the given crosstab.

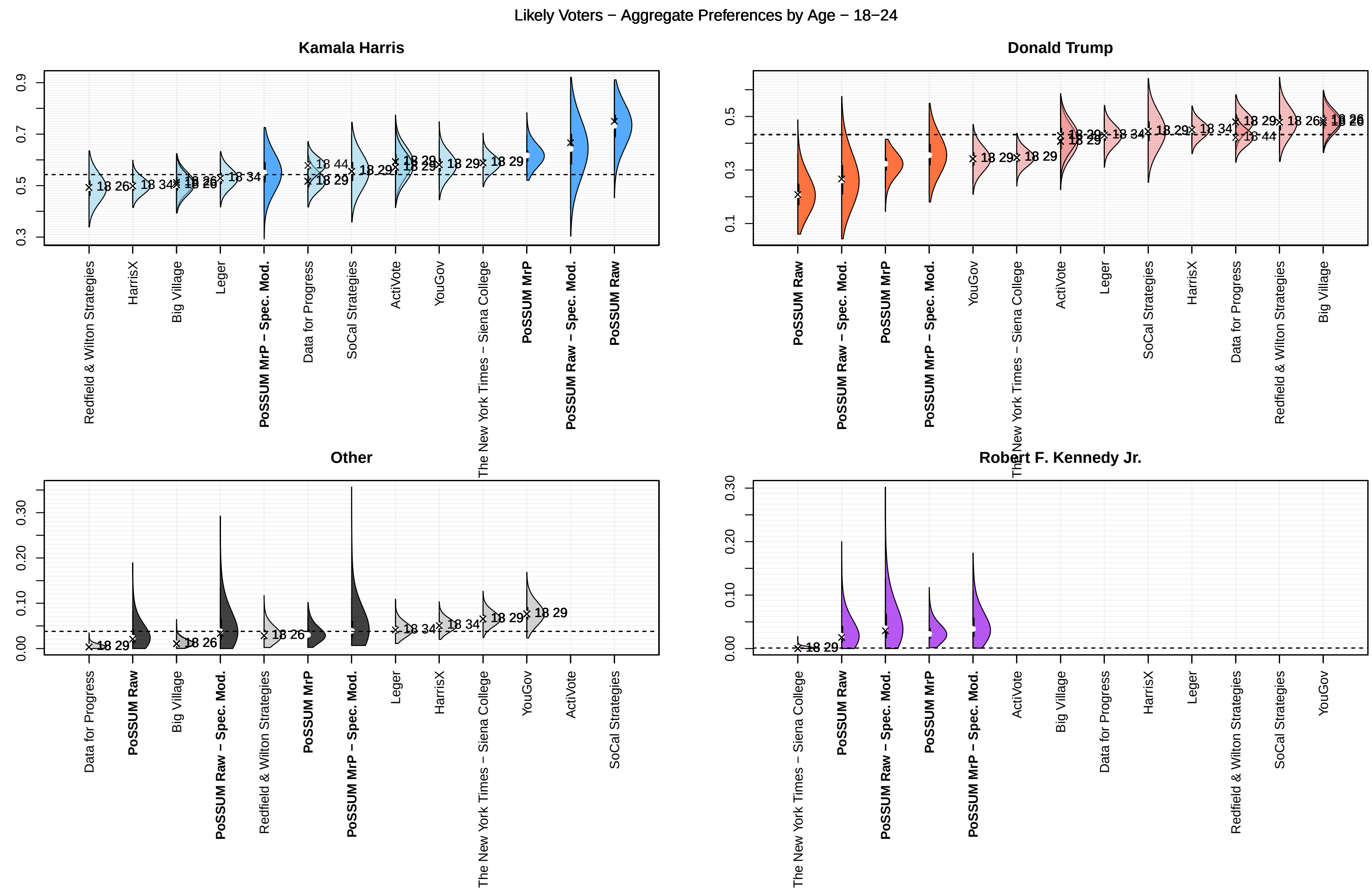


Figure 37: A comparison of the 18-24 years old distribution of 2024 vote choice across pollsters. The dotted line represents the simple arithmetic average of all polls. PoSSUM MrP is the primary estimate of interest. Large intervals for a given pollster correspond to small n for the given crosstab. To 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.

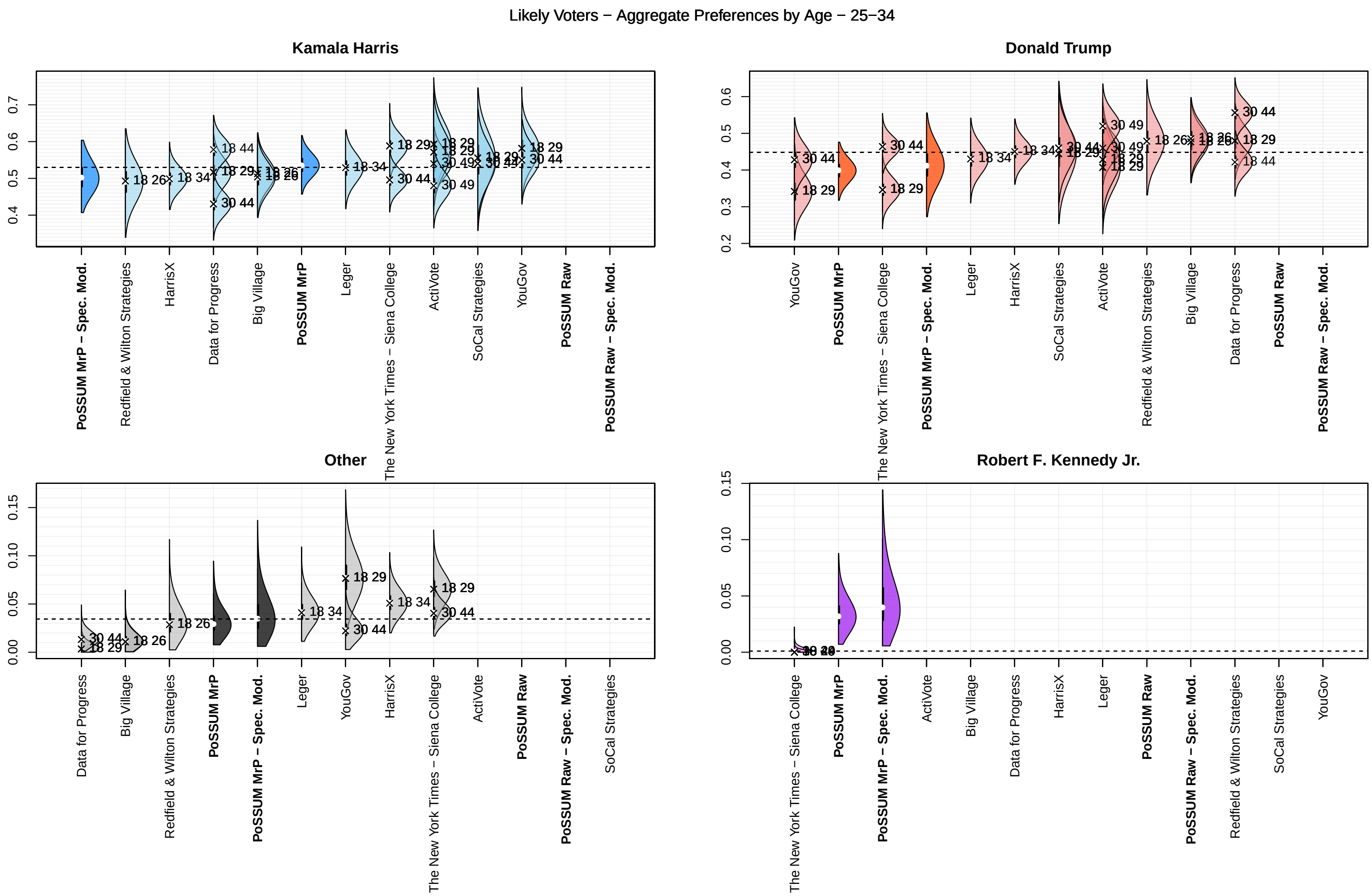


Figure 38: A comparison of the 25-34 years old distribution of 2024 vote choice across pollsters. The dotted line represents the simple arithmetic average of all polls. `PoSSUM MrP` is the primary estimate of interest. Large intervals for a given pollster correspond to small n for the given crosstab. 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.

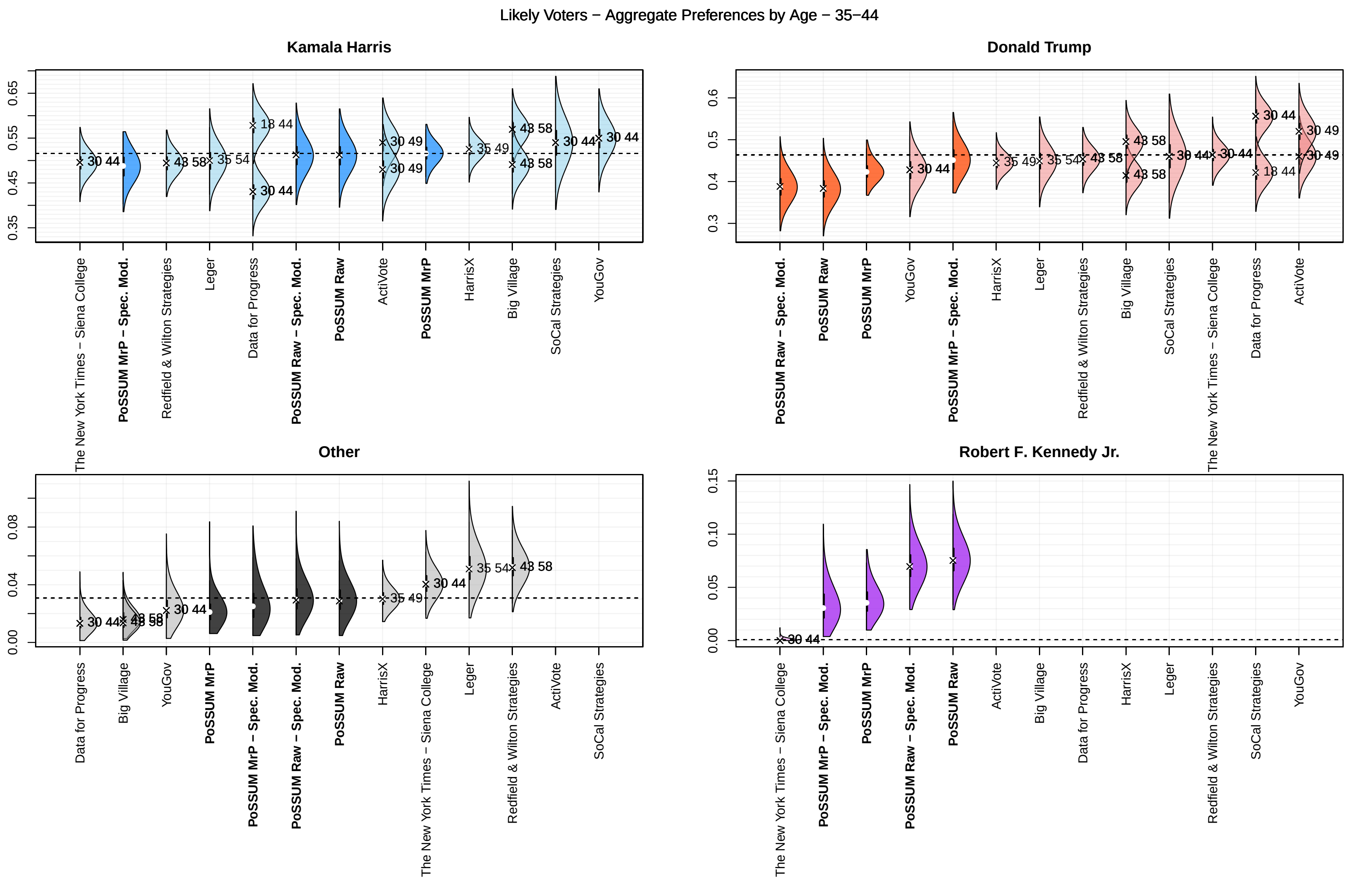


Figure 39: A comparison of the 35-44 years old distribution of 2024 vote choice across pollsters. The dotted line represents the simple arithmetic average of all polls. `PoSSUM MrP` is the primary estimate of interest. Large intervals for a given pollster correspond to small n for the given crosstab. 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.

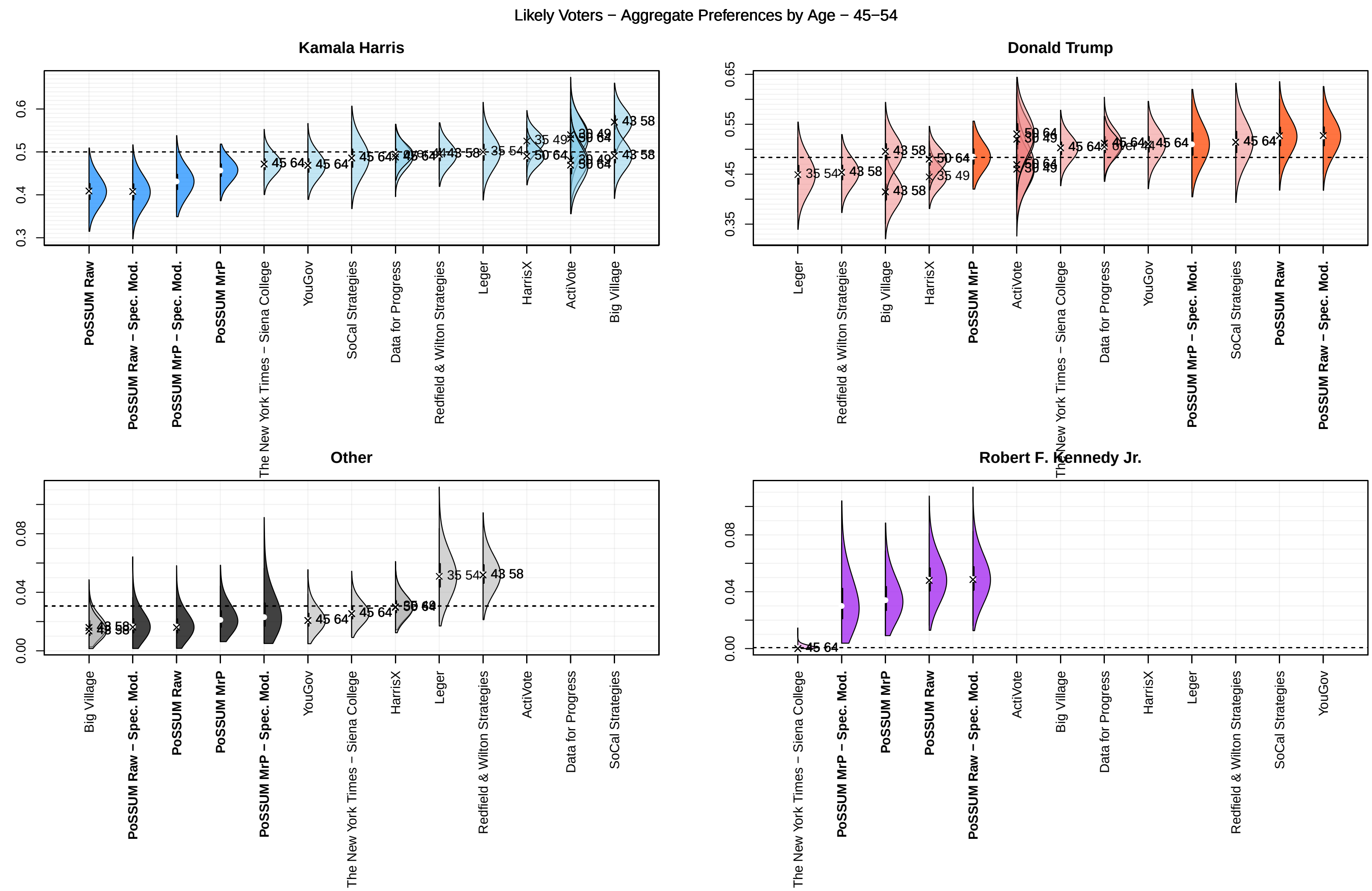


Figure 40: A comparison of the 45-54 years old distribution of 2024 vote choice across pollsters. The dotted line represents the simple arithmetic average of all polls. `PoSSUM MrP` is the primary estimate of interest. Large intervals for a given pollster correspond to small n for the given crosstab. 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.

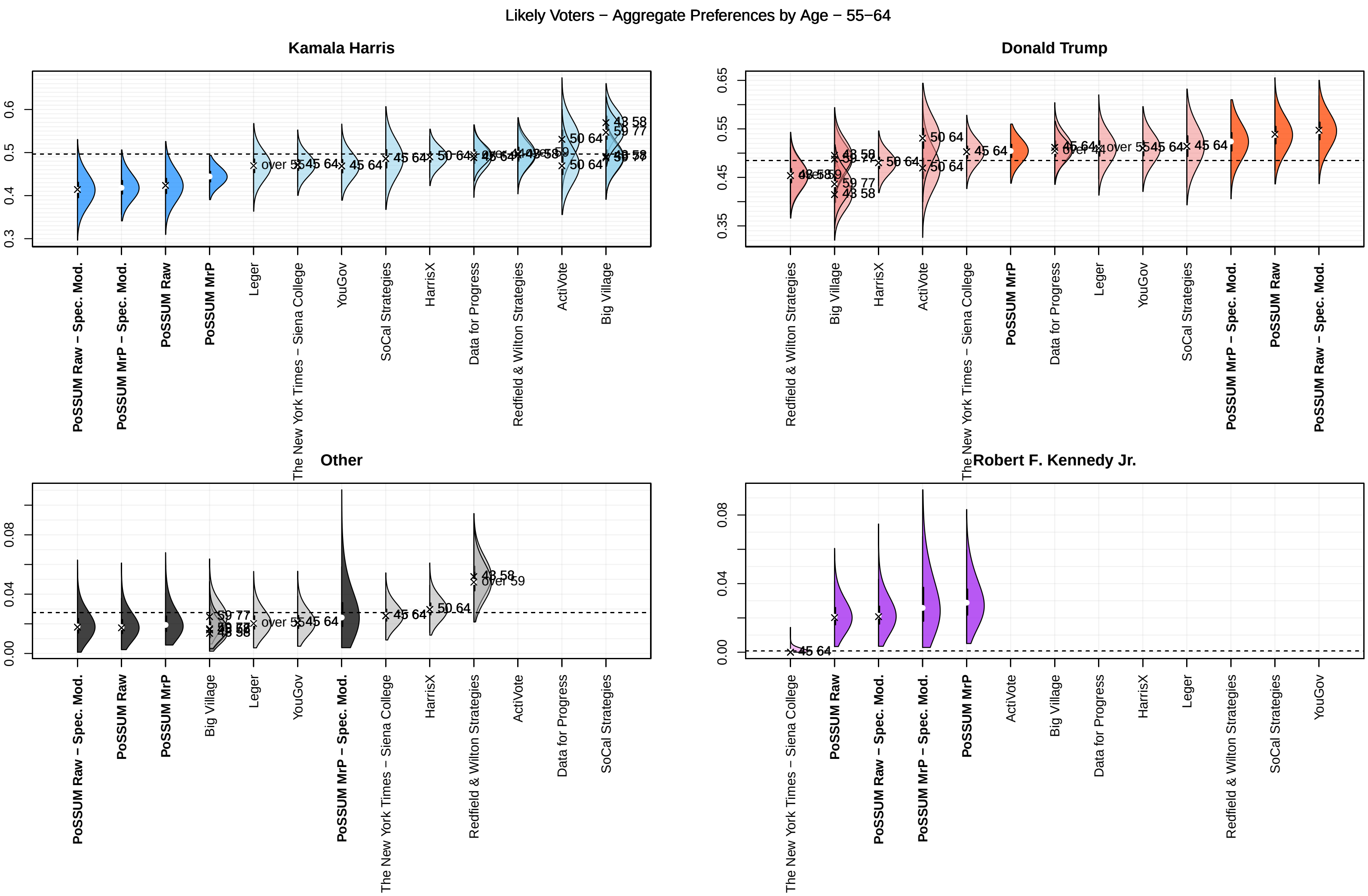


Figure 41: A comparison of the 55-64 years old distribution of 2024 vote choice across pollsters. The dotted line represents the simple arithmetic average of all polls. PoSSUM MrP is the primary estimate of interest. Large intervals for a given pollster correspond to small n for the given crosstab. To 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.

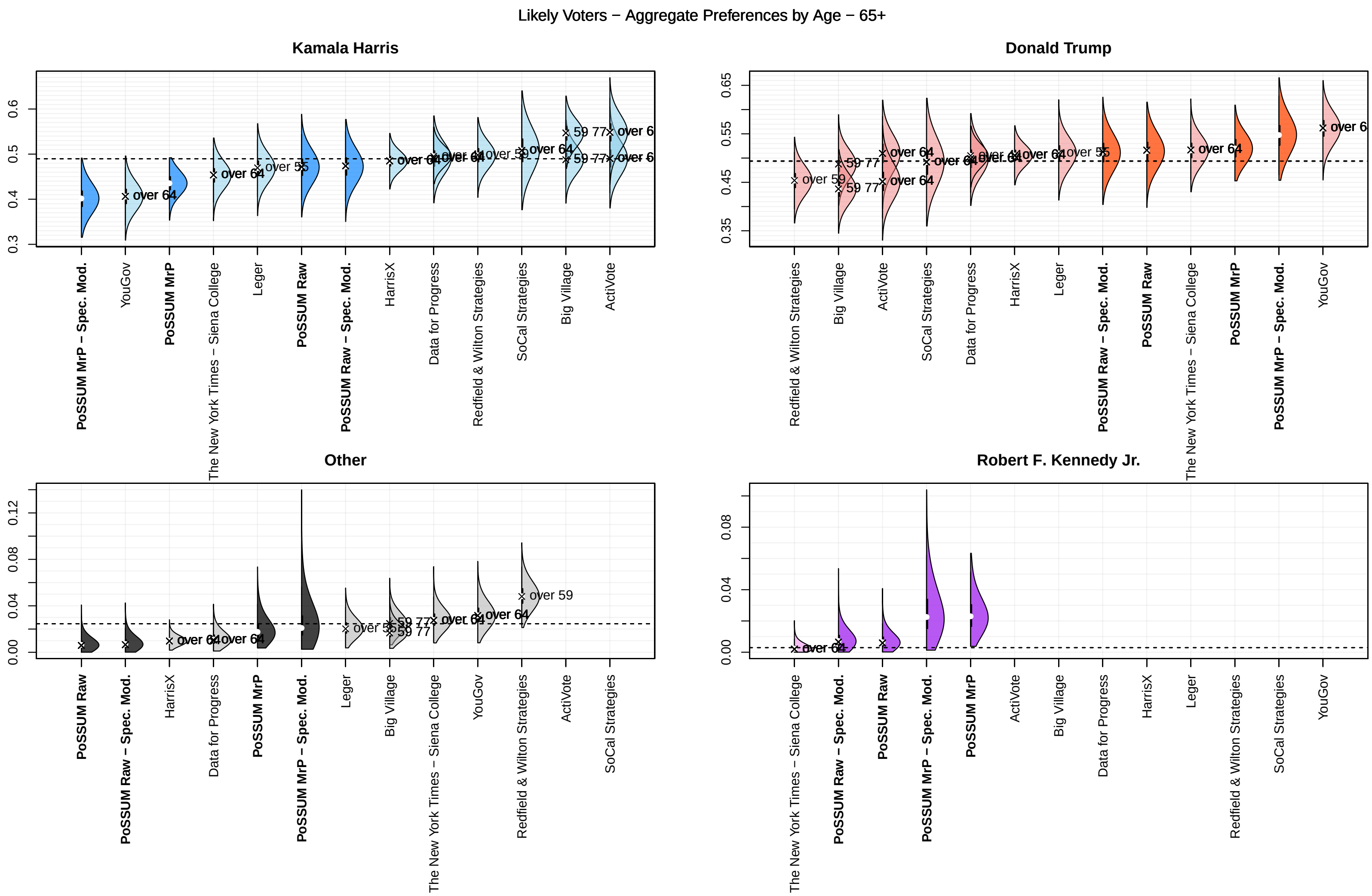


Figure 42: A comparison of the 65 years old and older distribution of 2024 vote choice across pollsters. The dotted line represents the simple arithmetic average of all polls. `PoSSUM MrP` is the primary estimate of interest. Large intervals for a given pollster correspond to small n for the given crosstab. 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.

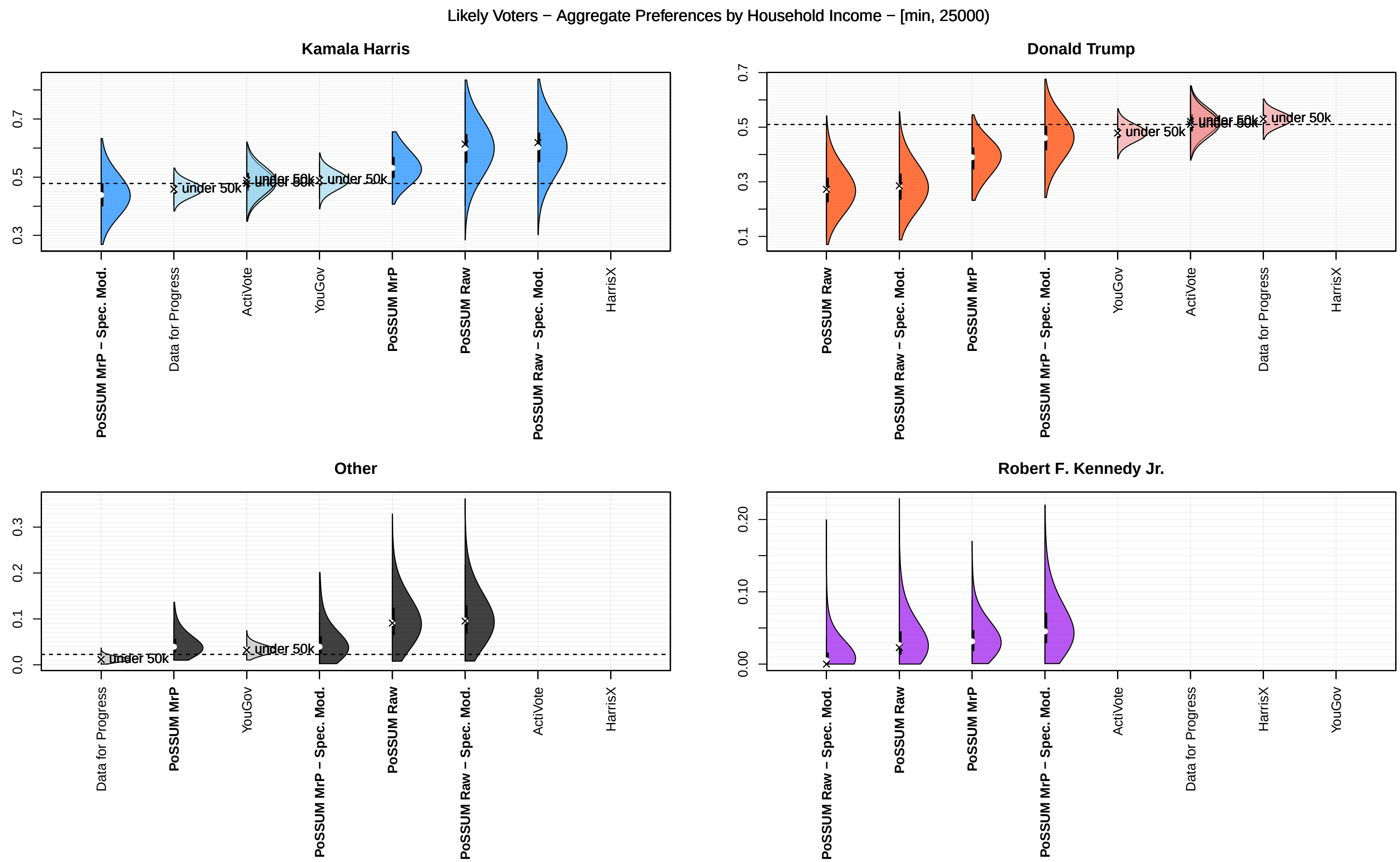


Figure 43: A comparison of the 2024 vote choice across pollsters, for voters with household income below $25k$. The dotted line represents the simple arithmetic average of all polls. `PoSSUM MrP` is the primary estimate of interest. Large intervals for a given pollster correspond to small n for the given 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.

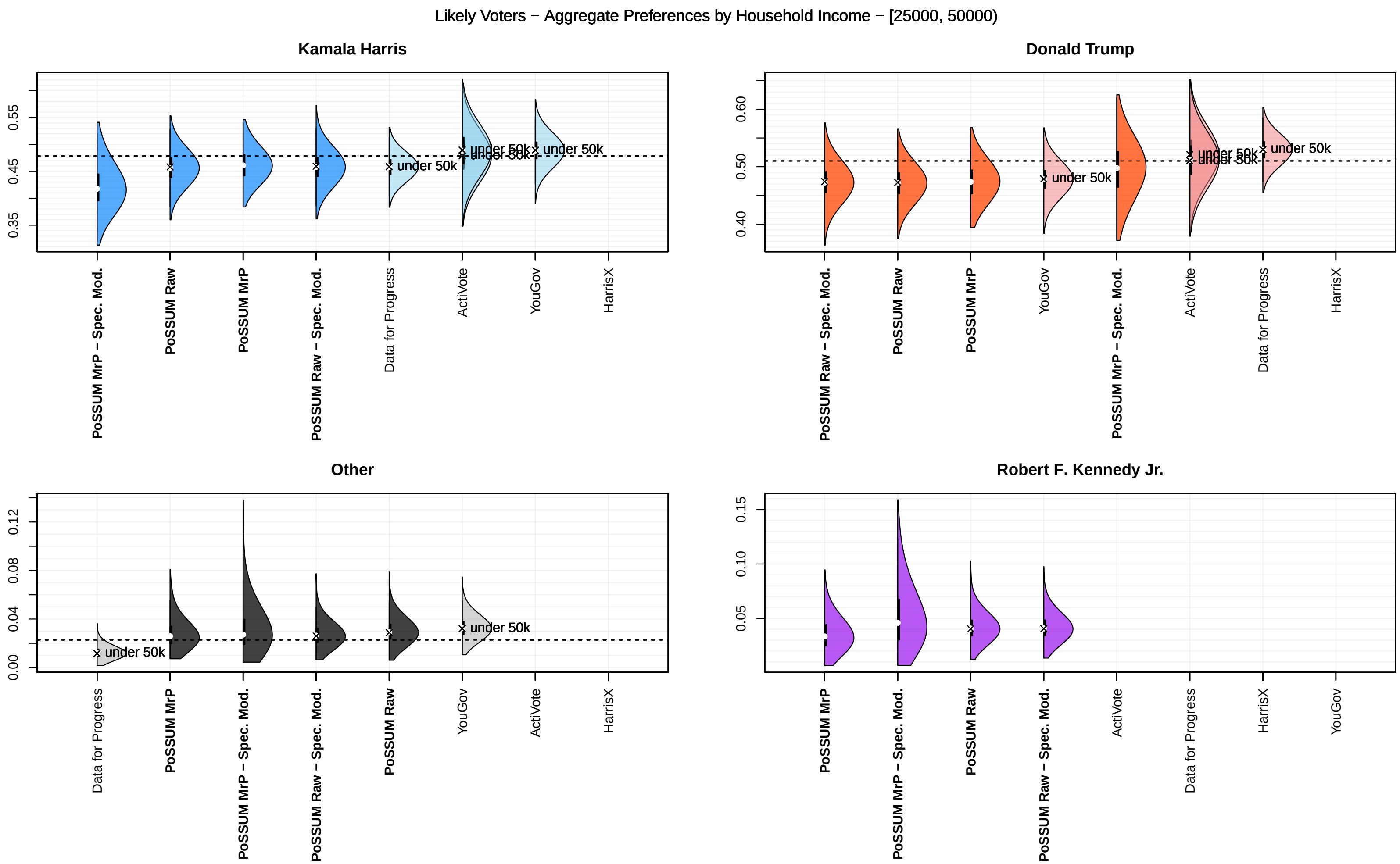


Figure 44: A comparison of the 2024 vote choice across pollsters, for voters with household income between $25k$ and $50k$. The dotted line represents the simple arithmetic average of all polls. `PoSSUM MrP` is the primary estimate of interest. Large intervals for a given pollster correspond to small 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.

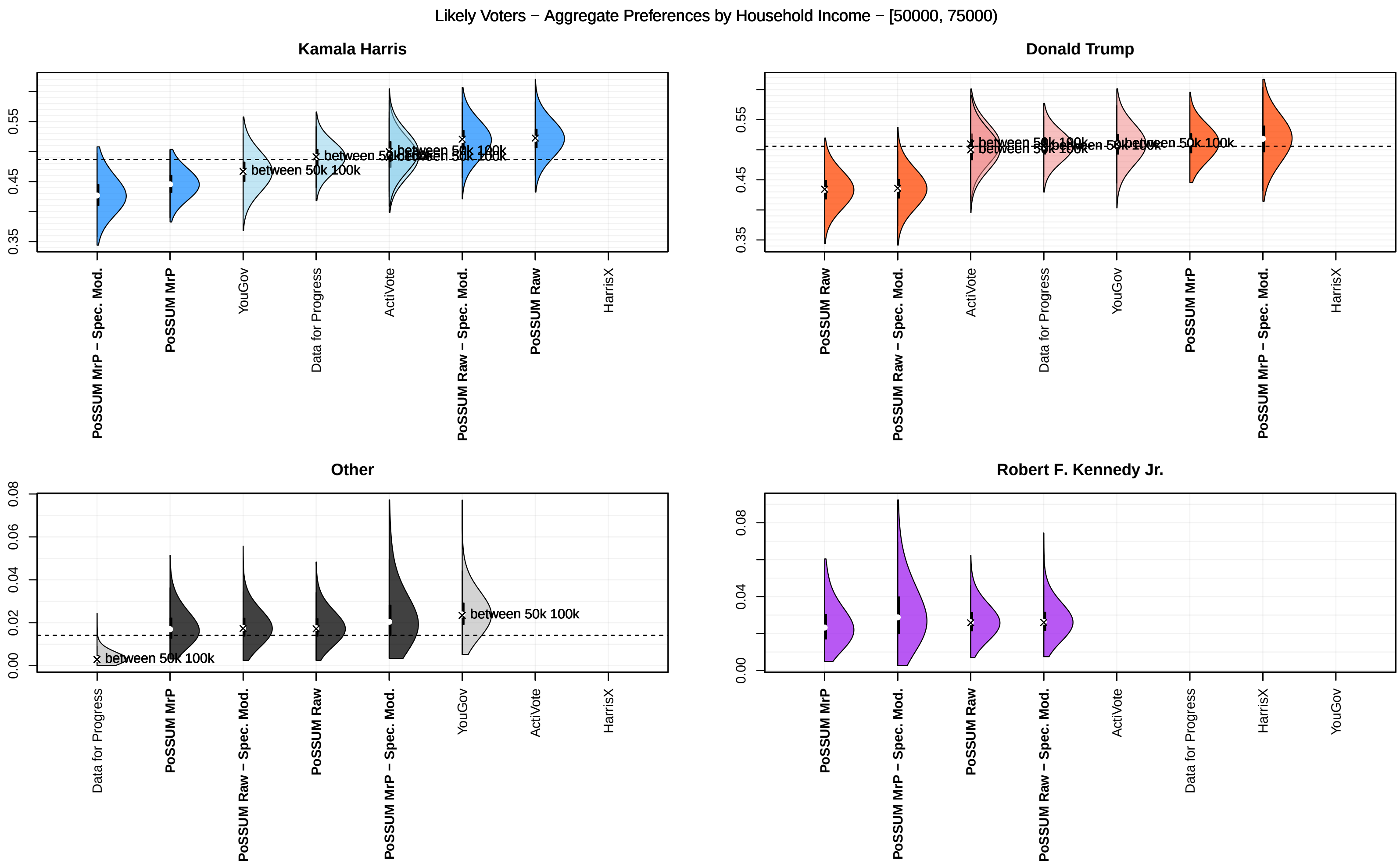


Figure 45: A comparison of the 2024 vote choice across pollsters, for voters with household income between $50k$ and $75k$. The dotted line represents the simple arithmetic average of all polls. `PoSSUM MrP` is the primary estimate of interest. Large intervals for a given pollster correspond to small 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.

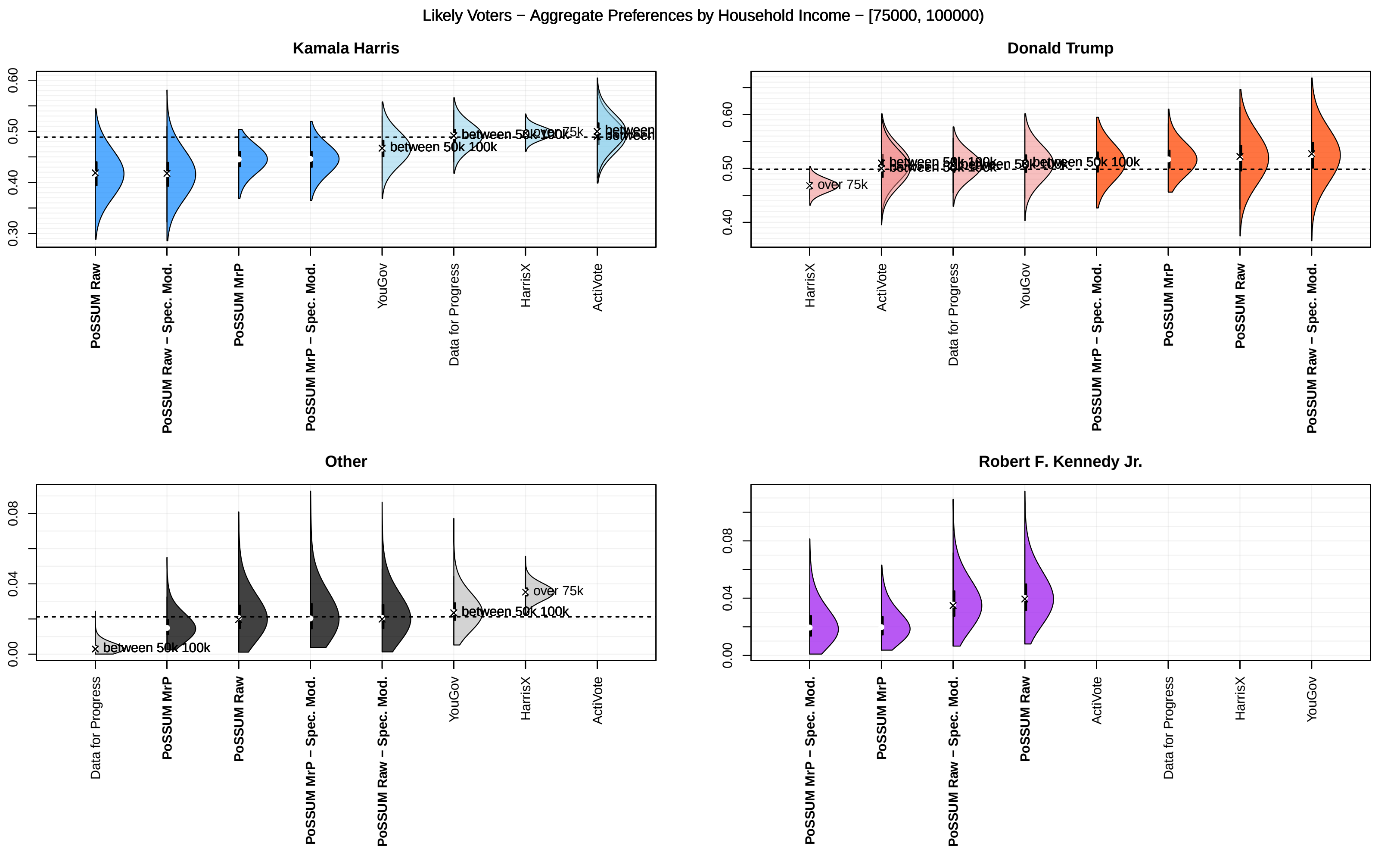


Figure 46: A comparison of the 2024 vote choice across pollsters, for voters with household income between $75k$ and $100k$. The dotted line represents the simple arithmetic average of all polls. `PoSSUM MrP` is the primary estimate of interest. Large intervals for a given pollster correspond to small 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.

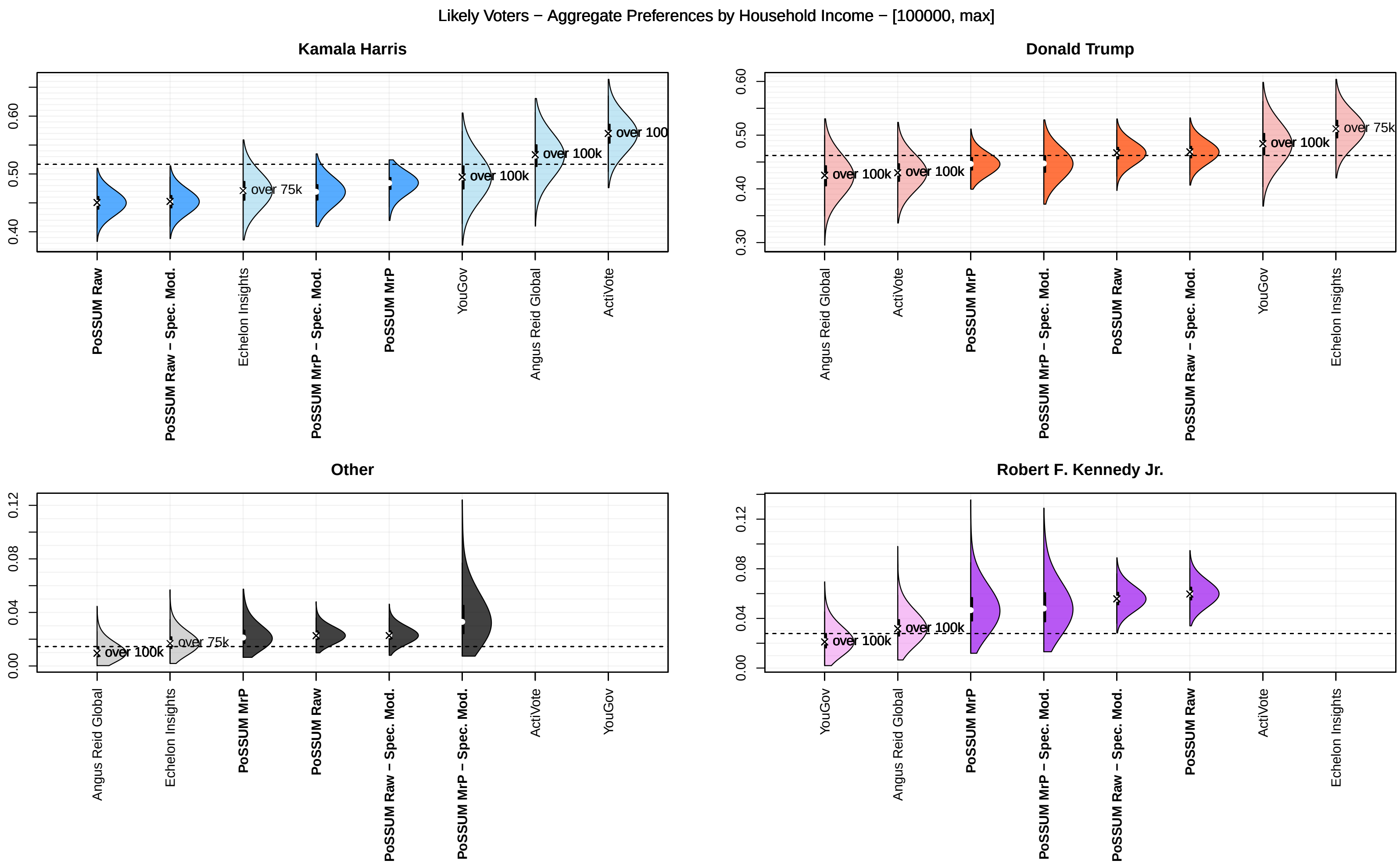


Figure 47: A comparison of the 2024 vote choice across pollsters, for voters with household income above $100k$. The dotted line represents the simple arithmetic average of all polls. `PoSSUM MrP` is the primary estimate of interest. Large intervals for a given pollster correspond to small n for the 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.

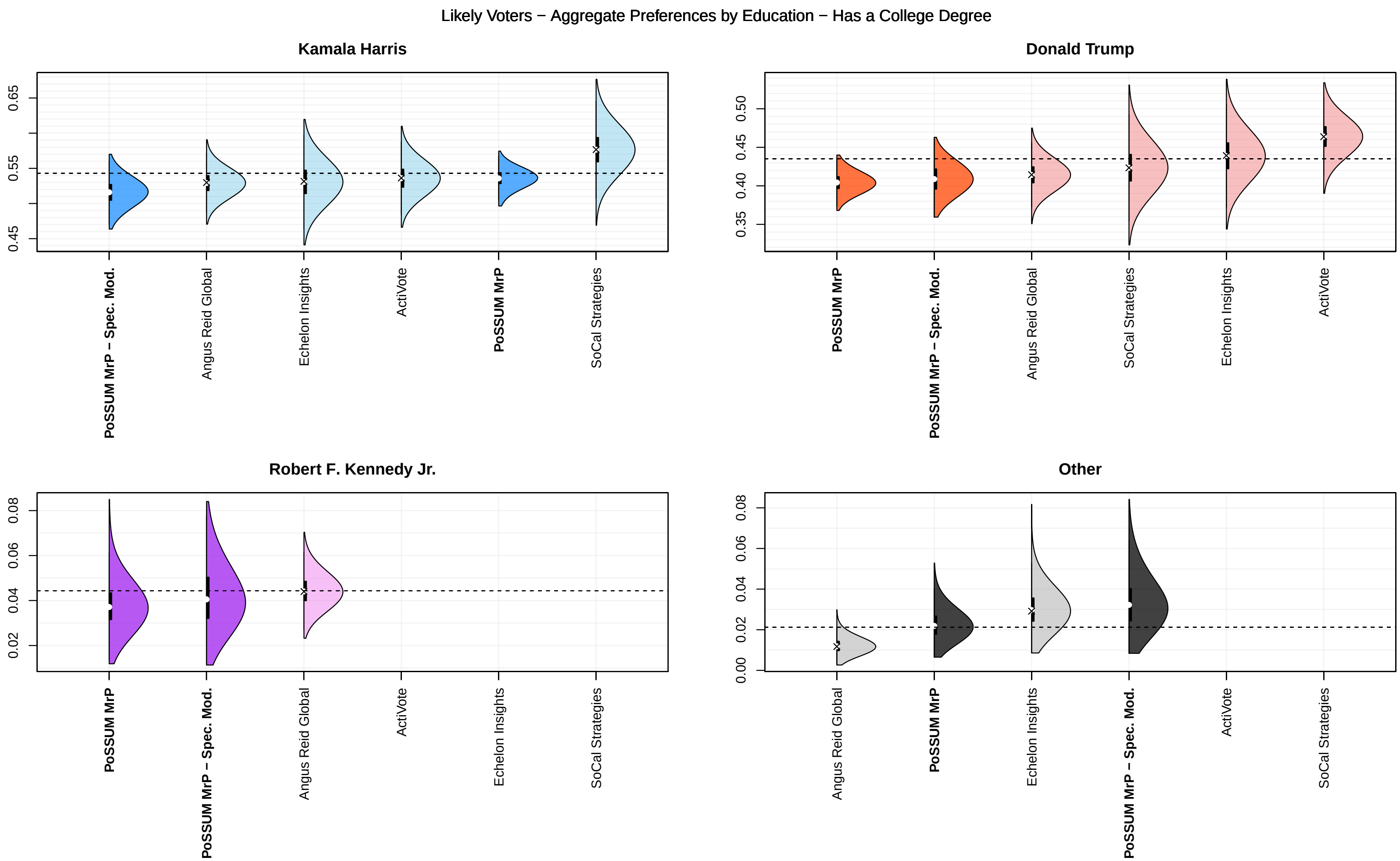


Figure 48: A comparison of the 2024 vote choice across pollsters, for voters with a college degree. The dotted line represents the simple arithmetic average of all polls. `PoSSUM MrP` is the primary estimate of interest. Large intervals for a given pollster correspond to small n for the given crosstab.

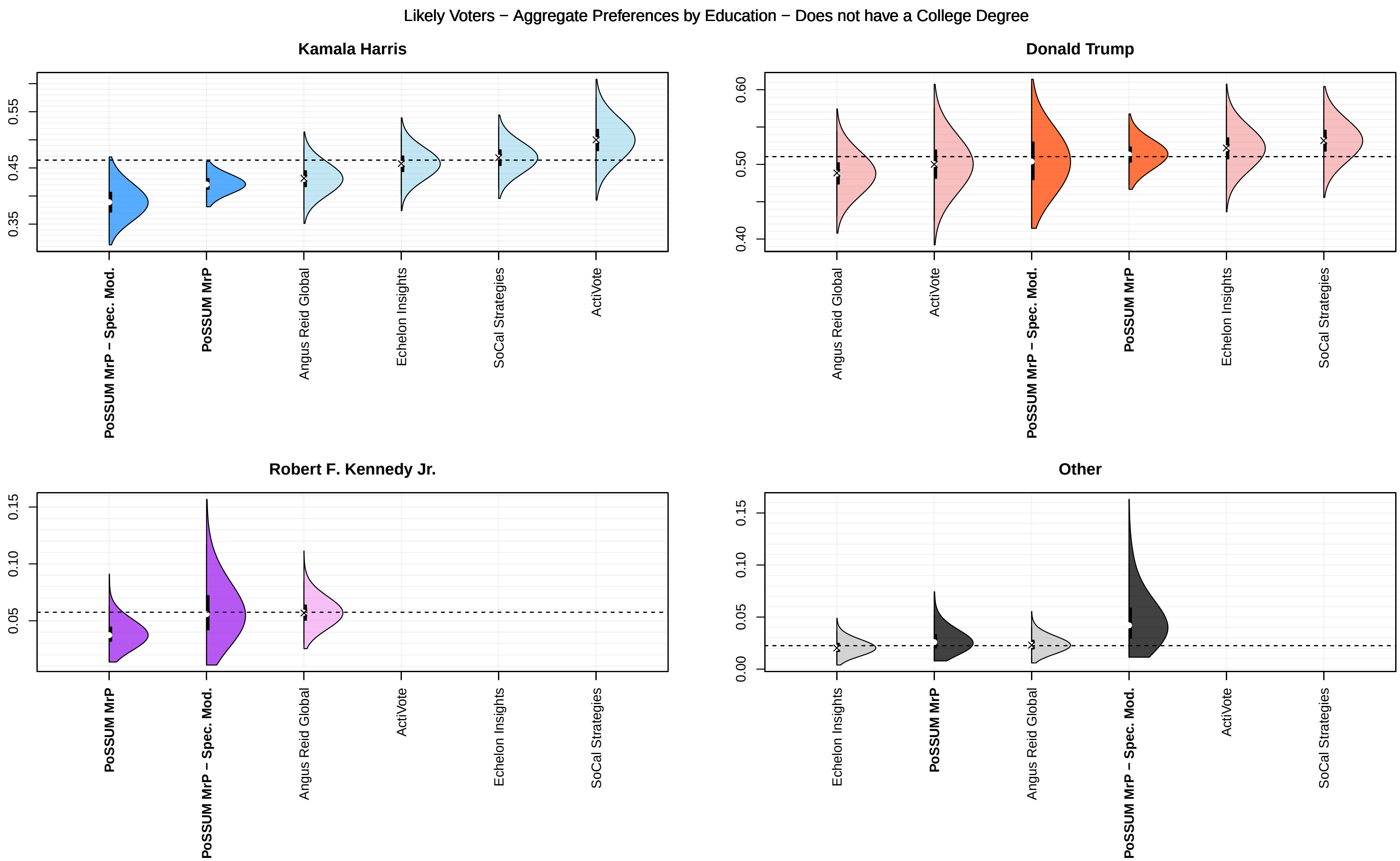


Figure 49: A comparison of the 2024 vote choice across pollsters, for voters without a college degree. The dotted line represents the simple arithmetic average of all polls. `PoSSUM MrP` is the primary estimate of interest. Large intervals for a given pollster correspond to small n for the given crosstab.

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.