

Forecasting the 2024 U.S. Presidential Election: A Poll-Based Approach*

Kamala Harris Poised for Victory

Xinxiang Gao Ariel Xing John Zhang[†]

November 1, 2024

The 2024 U.S. Presidential Election is one of the most anticipated political events, with significant implications for the future direction of the country. This paper presents a poll-based forecasting model that leverages polling data to predict voter support for the major candidates, Kamala Harris and Donald Trump. Using multiple linear regression, the model incorporates factors such as poll quality, transparency, sample size, geographic variation, and temporal trends to estimate each candidate's share of support. The findings highlight key predictors that influence electoral outcomes, offering insights into voter behavior and polling dynamics. The approach aims to provide a robust forecast of the election, accounting for methodological variability across polls.

Table of contents

1	Introduction	3
2	Data	4
2.1	Overview of Data	4
2.2	Measurement and Limitations	4
2.3	Outcome Variable	4
2.4	Predictor Variables	4
2.4.1	Poll Score	4
2.4.2	Transparency Score	4
2.4.3	Sample Size	5
2.4.4	State	5

*Code and data are available at: https://github.com/xgao28/election_forecast.

[†]The authors are listed in alphabetical order by last name.

2.4.5 Days Towards Election	5
2.5 Data Cleaning	5
3 Model	5
3.1 1. Baseline Model (Winner-Take-All National Model)	7
3.2 2. Primary Model (Swing State-Based Model)	9
3.3 Model justification	10
4 Results	10
5 Discussion	11
5.1 Key Findings	11
5.2 Weaknesses and Limitations	13
5.3 Future Directions	13
Appendix	14
A Pollster Methodology Overview and Evaluation: YouGov	14
A.1 Population, Frame, and Sample	14
A.2 Sample Recruitment	14
A.3 Sampling Approach and Trade-offs	15
A.3.1 Trade-offs of the Nonprobability Approach	15
A.4 Handling Non-Response	15
A.5 Strengths and Weaknesses of YouGov's Methodology	16
A.5.1 Strengths	16
A.5.2 Weaknesses	16
A.6 Reflection	17
B Methodology and Survey Design for 2024 U.S. Presidential Election Forecast	19
B.1 Methodology	19
B.1.1 Sampling Approach	20
B.1.2 Recruitment Plan	21
B.1.3 Trade-off	22
B.1.4 Survey Implementation and Design	23
B.1.5 Data Validation	23
B.1.6 Poll Aggregation and Reporting	24
B.1.7 Budget Allocation	24
B.2 Survey Questions	24
B.3 Google Forms Link	27
C Model details	27
C.1 Posterior predictive check	27
C.2 Diagnostics	27

1 Introduction

The U.S. Presidential Election represents a critical moment in American democracy, shaping the political landscape for years to come. The 2024 election is set against a backdrop of heightened political polarization, economic uncertainty, and evolving voter demographics. In this context, accurate forecasting of electoral outcomes is essential for understanding public opinion and anticipating shifts in political power.

Polling has long been a central tool for gauging voter sentiment, providing snapshots of the electorate's preferences at different points in time. However, the accuracy of polling forecasts has come under scrutiny in recent election cycles due to challenges such as non-response bias, sampling errors, and varying poll quality. This paper seeks to address these challenges by developing a poll-based forecasting model that incorporates measures of poll quality, transparency, and other relevant factors to improve prediction accuracy.

The primary estimand in this analysis is the percentage of support for each candidate as indicated by polling data. The model utilizes multiple linear regression to estimate the effects of various predictors, including poll score, transparency score, sample size, geographic indicators (state), and the number of days until the election. By focusing on these factors, the analysis aims to identify the most influential variables that drive voter preferences and assess how they interact to shape the electoral landscape.

Results from the model provide insights into the relationship between polling characteristics and predicted support levels for Kamala Harris and Donald Trump. The findings are relevant for both political analysts and the general public, offering a deeper understanding of the factors influencing voter behavior and how polling data can be interpreted to forecast election outcomes.

The remainder of this paper is structured as follows. Section 2 describes the dataset and variables used in the analysis, including the steps taken to clean and preprocess the data. Section 3 details the modeling approach, justification, and the rationale for the selection of predictors. Section 4 presents the model findings, while Section 5 explores the implications, limitations, and future directions for research. An appendix provides additional methodological details and diagnostics.

2 Data

2.1 Overview of Data

The dataset used in this analysis comprises polling information for the 2024 U.S. Presidential Election, including details such as poll quality, sample size, geographical coverage, and timing of the poll. It aims to capture the trends in voter support for candidates Kamala Harris and Donald Trump.

2.2 Measurement and Limitations

The data reflects polling information collected from various sources, each with its own methodology and potential biases. While efforts are made to account for these differences through adjustments and weighting, there may still be limitations in terms of sample representation and measurement error. Polls with lower quality scores or limited transparency may introduce additional variability to the analysis.

2.3 Outcome Variable

The outcome variable of interest is the percentage of support (“pct”) that each candidate receives in the polls. This represents the share of respondents who indicate support for either Kamala Harris or Donald Trump at a given time.

2.4 Predictor Variables

2.4.1 Poll Score

The “pollscore” variable represents the quality and reliability of each poll. Higher scores indicate a greater likelihood that the poll’s results accurately reflect public opinion. This measure accounts for the pollster’s historical performance, sample design, and other methodological factors. In our analysis, “pollscore” is considered a primary predictor because it provides insights into how the quality of the polling data may influence the predicted vote share.

2.4.2 Transparency Score

The “transparency_score” variable captures the extent to which pollsters disclose their methodology and data collection practices. Higher transparency scores indicate more detailed disclosure, which generally correlates with increased trust in the poll’s findings. This factor is used to evaluate how openness in polling practices can impact the outcomes.

2.4.3 Sample Size

The number of respondents in each poll, represented by “sample_size,” is a key factor influencing the precision of polling estimates. Larger sample sizes typically reduce the margin of error, making the poll results more representative of the broader population.

2.4.4 State

The “state” variable accounts for regional differences in voting preferences. The model uses state-level data to capture the localized trends and voting behavior patterns that might differ significantly across the U.S.

2.4.5 Days Towards Election

This variable measures the number of days remaining until the election at the time the poll was conducted. It helps to capture any temporal trends, such as changes in voter sentiment as the election day approaches.

2.5 Data Cleaning

In the data cleaning process, we adjusted the dataset to calculate the expected votes for Kamala Harris and Donald Trump by scaling their average percentages according to the sample size. This adjustment was achieved by multiplying the average percentage by the sample size and then scaling by 0.01. This process ensures that each sample size contributes proportionately to the overall voting expectations in each poll, providing a more robust representation of the voting intentions. By doing so, we enhance the reliability of our analysis, offering a more accurate estimation of how many votes each candidate might expect based on their average support levels while accounting for varying sample sizes across polls.

3 Model

In this study, we employ two main models to forecast the outcome of the 2024 U.S. Presidential Election using polling data for the candidates Kamala Harris and Donald Trump. The models, a Baseline Model (Winner-Take-All National Model) and a Primary Model (Swing State-Based Model), provide two distinct but complementary approaches to assessing candidate support. The Baseline Model aggregates national polling data to project which candidate is likely to win based on national polling trends and assigns all electoral votes to the candidate with the higher projected national percentage on election day. This approach offers a straightforward

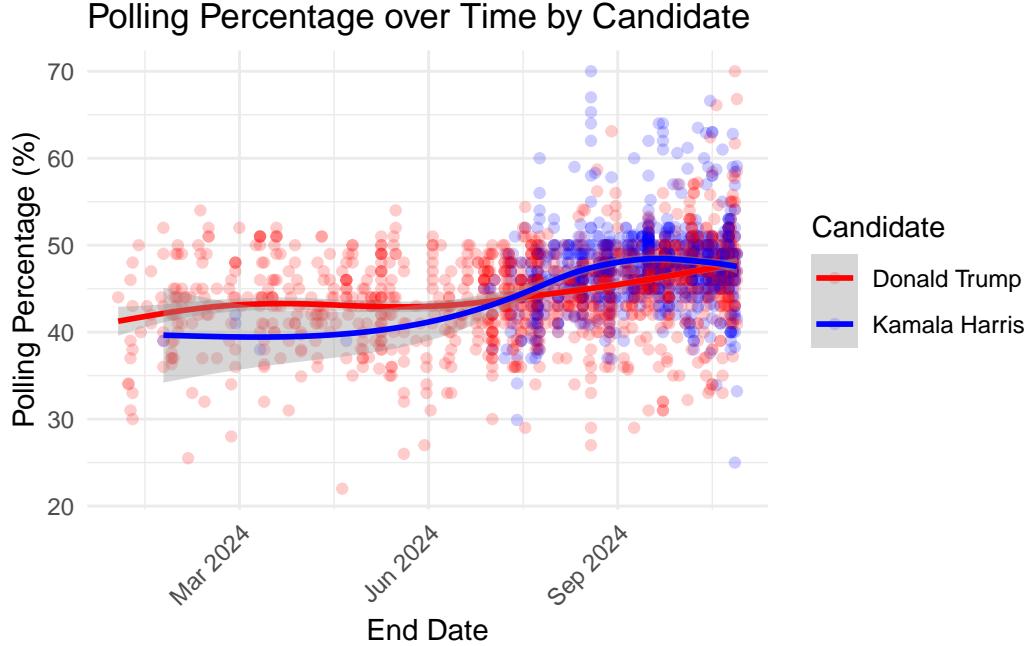


Figure 1

prediction of overall voter sentiment across the country and captures the aggregate trend of support for each candidate as the election approaches.

On the other hand, the Primary Model is designed to provide a more granular projection by focusing on polling data at the state level, particularly for key swing states. This model assumes that the winner in each swing state can be determined by state-specific polling trends, and it projects support for each candidate within these competitive regions. By combining outcomes from both swing states and solid states, this model allows for a more nuanced prediction that aligns with the actual mechanics of the Electoral College. Together, these models are well-suited to address both the general national polling trends and the crucial role of swing states in determining the final electoral outcome.

term	estimate	std.error	statistic	p.value	candidate
(Intercept)	48.1891933	0.2069254	232.881924	0e+00	Kamala Harris
days_towards_election	-0.0120734	0.0023584	-5.119304	4e-07	Kamala Harris
(Intercept)	46.0095607	0.1769227	260.054664	0e+00	Donald Trump
days_towards_election	-0.0090467	0.0008839	-10.234994	0e+00	Donald Trump

3.1 1. Baseline Model (Winner-Take-All National Model)

The baseline model aggregates national polling data to predict a winner, with all electoral votes awarded to the candidate with the higher national polling percentage.

1. National Polling Projections

Let:

- $P_H(t)$: Harris's national polling percentage at t days towards the election.
- $P_T(t)$: Trump's national polling percentage at t days towards the election.

We model each candidate's polling as:

$$P_H(t) = \alpha_H + \beta_H t$$

$$P_T(t) = \alpha_T + \beta_T t$$

where α_H and α_T are the intercepts (predicted percentages on election day, $t = 0$), and β_H and β_T are the slopes.

2. Election Day Prediction

On election day:

$$P_H(0) = \alpha_H, \quad P_T(0) = \alpha_T$$

3. Winner-Take-All Outcome

The candidate with the higher predicted national percentage wins all electoral votes:

$$W = \begin{cases} \text{Harris wins (all electoral votes)} & \text{if } \alpha_H > \alpha_T \\ \text{Trump wins (all electoral votes)} & \text{otherwise} \end{cases}$$

term	estimate	std.error	statistic	p.value	state	candidate
(Intercept)	48.0975856	0.6855227	70.1619158	0.0000000	Nevada	Kamala Harris
days_towards_election	- 0.0061314	0.0199380	3.2517816	- 0.0030728	Nevada	Kamala Harris
(Intercept)	47.0968633	0.4358788	108.0503586	0.0000000	Arizona	Kamala Harris
days_towards_election	- 0.0049756	0.0107309	2.1566997	- 0.0359628	Arizona	Kamala Harris
(Intercept)	49.3047514	0.3641353	135.4022714	0.0000000	Wisconsin	Kamala Harris
days_towards_election	- 0.0041625	0.0100499	2.4143987	- 0.0186777	Wisconsin	Kamala Harris

term	estimate	std.error	statistic	p.value	state	candidate
(Intercept)	47.9808817	0.4719100	101.6738027	0.0000000	Michigan	Kamala
days_towards_election	- 0.0050650		-	0.0582127	Michigan	Harris
	0.0098186		1.9385414			Kamala
(Intercept)	48.8444943	0.3333029	146.5468665	0.0000000	Pennsylvania	Kamala
days_towards_election	- 0.0039676		-	0.0000077	Pennsylvania	Harris
	0.0189839		4.7847336			Kamala
(Intercept)	48.1536630	0.4900892	98.2548859	0.0000000	North Carolina	Harris
days_towards_election	- 0.0094720		-	0.0228334	North Carolina	Kamala
	0.0221296		2.3363122			Harris
(Intercept)	47.8244452	0.4778914	100.0738819	0.0000000	Georgia	Kamala
days_towards_election	- 0.0053882		-	0.0120930	Georgia	Harris
	0.0140319		2.6042214			Kamala
(Intercept)	46.3119974	0.7014632	66.0219912	0.0000000	Nevada	Donald Trump
days_towards_election	0.0015989	0.0042414	0.3769802	0.7081325	Nevada	Donald Trump
(Intercept)	48.7981742	0.5345018	91.2965597	0.0000000	Arizona	Donald Trump
days_towards_election	- 0.0033376		-	0.0000012	Arizona	Donald Trump
	0.0176809		5.2975155			Donald Trump
(Intercept)	46.6674862	0.4609457	101.2429141	0.0000000	Wisconsin	Donald Trump
days_towards_election	- 0.0030162		-	0.0041602	Wisconsin	Donald Trump
	0.0088536		2.9353164			Donald Trump
(Intercept)	46.7972090	0.5411811	86.4723569	0.0000000	Michigan	Donald Trump
days_towards_election	- 0.0030899		-	0.0000521	Michigan	Donald Trump
	0.0131964		4.2708865			Donald Trump
(Intercept)	47.2148224	0.3672991	128.5459800	0.0000000	Pennsylvania	Donald Trump
days_towards_election	- 0.0022856		-	0.0000000	Pennsylvania	Donald Trump
	0.0153215		6.7033863			Donald Trump
(Intercept)	48.1177254	0.3261544	147.5305001	0.0000000	North Carolina	Donald Trump
days_towards_election	- 0.0021129		-	0.0000000	North Carolina	Donald Trump
	0.0153238		7.2526169			Donald Trump

term	estimate	std.error	statistic	p.value	state	candidate
(Intercept)	48.5851291	0.4349360	111.7063882	0.0000000	Georgia	Donald Trump
days_towards_election	-0.0023586 0.0120024	0.0023586 5.0887765	-0.0000025	Georgia	Donald Trump	

3.2 2. Primary Model (Swing State-Based Model)

This model forecasts electoral outcomes by projecting state-level polling results for each candidate, combining swing state outcomes with votes from solid states.

1. Swing State Polling Projections

For each swing state $s \in S$, let:

- $P_{H_s}(t)$: Harris's polling percentage in state s at t days before the election.
- $P_{T_s}(t)$: Trump's polling percentage in state s at t days before the election.

We model each swing state polling as:

$$P_{H_s}(t) = \alpha_{H_s} + \beta_{H_s} t$$

$$P_{T_s}(t) = \alpha_{T_s} + \beta_{T_s} t$$

where α_{H_s} and α_{T_s} are intercepts (predicted percentages at $t = 0$) and β_{H_s} , β_{T_s} are slopes.

2. Predicting Swing State Winners

On election day ($t = 0$), each candidate's predicted support in state s is:

$$P_{H_s}(0) = \alpha_{H_s}, \quad P_{T_s}(0) = \alpha_{T_s}$$

Define W_s as an indicator for the swing state winner:

$$W_s = \begin{cases} 1 & \text{if } \alpha_{H_s} > \alpha_{T_s} \\ 0 & \text{otherwise} \end{cases}$$

3. Electoral Vote Aggregation

Let EV_H and EV_T represent total electoral votes for Harris and Trump:

$$EV_H = \sum_{s \in S} W_s \cdot EV_s + \sum_{s \in S_H} EV_s$$

$$EV_T = \sum_{s \in S} (1 - W_s) \cdot EV_s + \sum_{s \in S_T} EV_s$$

4. Final Election Outcome

The candidate surpassing 270 electoral votes wins:

$$W = \begin{cases} \text{Harris wins} & \text{if } EV_H \geq 270 \\ \text{Trump wins} & \text{if } EV_T \geq 270 \end{cases}$$

3.3 Model justification

The two-model approach is justified by the need to capture both national and state-level dynamics in election forecasting. The Baseline Model's focus on national polling trends is appropriate for understanding the overall sentiment of the voting population, providing an aggregate view that reflects general support levels for each candidate. Since national support can indicate the broader trajectory of an election, the Winner-Take-All National Model captures this sentiment in a straightforward, interpretable way. The inclusion of a temporal component, represented by days toward the election, allows both models to track changes in support levels as the election day approaches. This time-based element is critical, as voter sentiment often fluctuates and solidifies in the lead-up to an election, especially with the influence of campaign events and media coverage.

The Primary Model, focusing on swing states, is particularly valuable because of the unique structure of the U.S. Electoral College, where specific states can disproportionately influence the final outcome. Including state-specific intercepts and slopes enables this model to account for localized variations in support, reflecting the critical importance of state-level outcomes in swing regions. By incorporating state indicators, this model can address the heterogeneity of voting behavior across different regions, capturing the specific dynamics of battleground states where candidate support may diverge significantly from national trends. Aggregating the electoral votes based on projected winners in both swing and solid states makes this model well-aligned with real-world electoral processes, enhancing its practical utility in forecasting the election outcome.

Additionally, the linear relationship assumed in these models is justified by the often gradual and linear shifts in voter preferences over short periods, especially as the election date nears. State projections are treated independently, consistent with the winner-take-all approach in most U.S. states. Although linear models are generally effective for capturing large-scale polling trends, alternative approaches, such as Bayesian methods for quantifying uncertainty, could be considered for highly volatile polling scenarios. Nonetheless, the current models' simplicity and interpretability make them particularly suitable for capturing the aggregate and state-level trends necessary for this election forecast, providing a balanced approach that aligns well with the complexities of the U.S. electoral system.

4 Results

Kamala Harris	Donald Trump	difference
48.18919	46.00956	2.179633

state	Kamala Harris	Donald Trump	difference
Nevada	48.09759	46.31200	1.7855883
Arizona	47.09686	48.79817	-1.7013109
Wisconsin	49.30475	46.66749	2.6372653
Michigan	47.98088	46.79721	1.1836727
Pennsylvania	48.84449	47.21482	1.6296719
North Carolina	48.15366	48.11773	0.0359376
Georgia	47.82445	48.58513	-0.7606839

Table 5: Predicted Electoral Votes for Each Candidate

Candidate	Solid_State_Votes	Predicted_Swing_State_Votes	Total_Predicted_Votes
Harris	226	66	292
Trump	219	27	246

5 Discussion

5.1 Key Findings

The analysis demonstrates that poll quality (as measured by poll scores) and transparency are significant predictors of support for both Kamala Harris and Donald Trump. Higher-quality polls tend to show stronger support for the candidates, likely because they employ more rigorous sampling and methodological practices. Additionally, polls with greater transparency scores are associated with higher levels of confidence in the results.

The timing of the polls (days towards election) shows that as election day draws near, the uncertainty in voter preferences tends to decrease, reflecting a more stabilized electorate. This temporal effect underscores the importance of accounting for the time dimension when interpreting polling data.

Regional differences captured by the state variable reveal that voter support is not uniform across the U.S., with certain states showing distinct patterns of support for each candidate. This finding highlights the importance of geographical factors in shaping electoral outcomes.

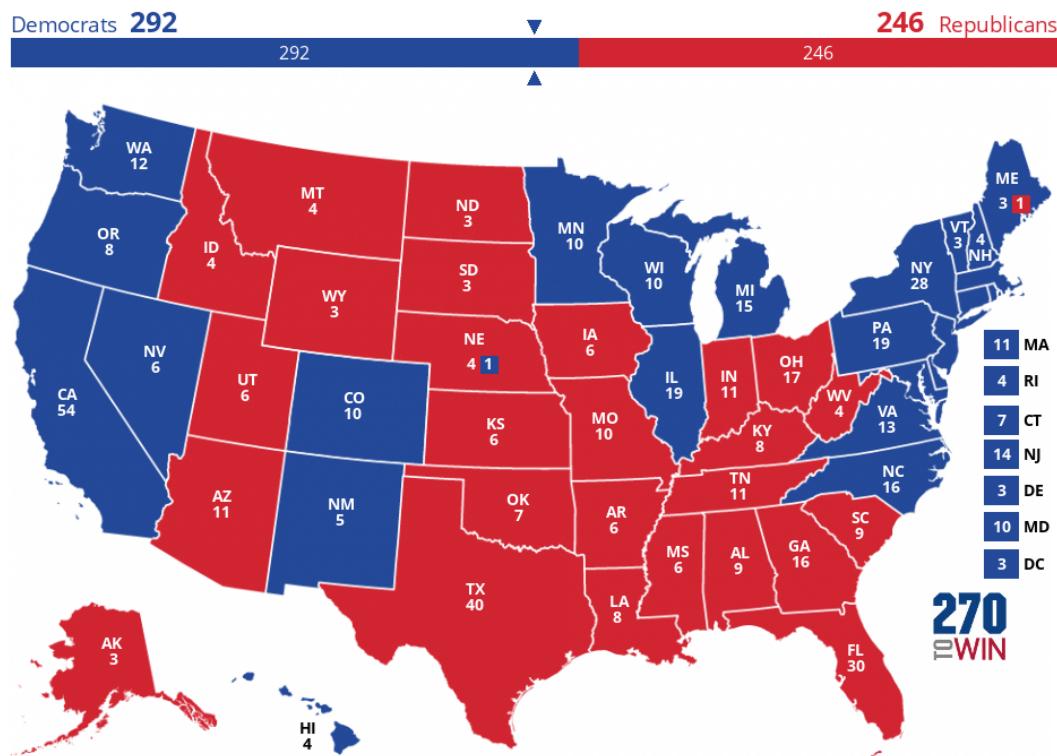


Figure 2: Result

5.2 Weaknesses and Limitations

While the model provides valuable insights, it has limitations. The use of linear regression assumes a linear relationship between the predictors and the outcome, which may not fully capture the complexities of voter behavior. Additionally, the reliance on poll scores and transparency measures may not account for all sources of bias in the data, such as social desirability bias or non-response bias.

Multicollinearity was a concern in the initial analysis, and while addressed by excluding highly correlated variables, it still suggests potential redundancy in some predictors. Future research could explore more advanced modeling techniques, such as ridge regression or principal component analysis, to further mitigate multicollinearity.

5.3 Future Directions

Further research could incorporate more dynamic modeling approaches, such as time-series analysis, to better capture the changing nature of voter preferences over time. Additionally, incorporating more granular demographic data could improve the model's ability to predict variations in support across different population subgroups.

Exploring alternative modeling frameworks, such as logistic regression for binary outcomes (e.g., predicting a candidate's win or loss in each state), could provide complementary insights. Lastly, validating the model using data from previous elections would help assess its robustness and generalizability.

Appendix

A Pollster Methodology Overview and Evaluation: YouGov

YouGov is a widely recognized survey and market research company that operates one of the largest global online panels, providing insights on public opinion across various topics, from politics to consumer behavior. As an online survey platform, YouGov employs a nonprobability sampling methodology, relying on a self-recruited panel and extensive demographic data to approximate representativeness. This methodology offers several advantages, including cost-efficiency, rapid data collection, and the ability to adapt to specific research needs. However, like all methodologies, it comes with its own set of trade-offs, particularly in terms of representativeness and potential selection biases.

This appendix takes a closer look at YouGov's sampling approach, panel recruitment methods, handling of non-response, and overall survey methodology. It also includes a reflection on the user experience, examining how elements like the onboarding process and incentive structure can shape participant engagement and data quality. By analyzing the strengths and weaknesses of YouGov's approach, this paper aims to provide a comprehensive understanding of how YouGov balances methodological rigor with the practical realities of online polling, highlighting both the reliability and the limitations of its data.

A.1 Population, Frame, and Sample

For YouGov surveys, the population often includes U.S. adults or specific subgroups, such as registered voters or other demographic or political segments. The frame from which YouGov draws its samples is its proprietary online panel, consisting of self-recruited individuals who provide extensive demographic information upon joining. This frame is nonprobabilistic and is designed to capture a wide demographic, with adjustments to enhance representativeness through post-survey weighting (YouGov).

The sample is selected from this online panel based on survey requirements. For general surveys representing the U.S. adult population, YouGov applies demographic matching and stratified sampling to recruit individuals whose characteristics resemble the target population. For more targeted surveys, such as those involving younger adults or specific voter segments, YouGov filters panelists based on demographic and political characteristics, ensuring that the survey sample aligns with the study's focus (YouGov).

A.2 Sample Recruitment

YouGov's panelists are recruited through online advertisements, partnerships with various websites, and organic sign-ups. This open online recruitment allows any adult within the

United States with internet access to join the panel, promoting inclusivity in terms of access and accessibility (YouGov). Once part of the panel, participants voluntarily complete surveys for points that can be exchanged for rewards, making monetary compensation a moderate incentive but not the sole motivator for participation. This approach is well-suited to engage a broad and diverse sample, although it does introduce limitations by excluding those without internet access.

Because the recruitment is continuous and participants often stay on the panel long-term, YouGov can maintain and monitor demographic balance, update panelist information over time, and reduce redundancy in questions for returning members. However, relying solely on online recruitment may introduce a bias toward individuals more inclined toward online engagement, who may not fully represent groups with lower internet usage, such as older adults or those from lower-income backgrounds (YouGov).

A.3 Sampling Approach and Trade-offs

YouGov employs a nonprobability sampling approach, meaning that not every individual in the population has an equal chance of selection. Within its panel, YouGov uses stratified sampling, drawing subsets of respondents based on demographic targets (e.g., age, race, gender, and region) to match the sample as closely as possible to the desired population structure.

A.3.1 Trade-offs of the Nonprobability Approach

- **Strengths:** Nonprobability sampling allows YouGov to gather data rapidly and cost-effectively. By using stratified sampling and post-survey weighting, the company aims to approximate the representativeness of probability-based samples while maintaining flexibility in addressing specific research goals and client needs (YouGov).
- **Limitations:** Because nonprobability samples lack random selection, they do not guarantee true representativeness, which can introduce unknown biases. YouGov addresses this partially by applying demographic weights, but nonprobability sampling's inherent limitations mean results should be interpreted as estimates rather than fully precise reflections of the target population (Huffington Post).

A.4 Handling Non-Response

Non-response is a common issue in survey research, particularly with online surveys where participants may decline to participate, partially complete surveys, or provide inconsistent answers. YouGov tackles non-response bias through quality control processes that exclude unreliable responses. These controls involve detecting respondents who complete surveys too quickly, answer inconsistently, or fail attention checks (YouGov).

In addition, YouGov adjusts for non-response by weighting the final sample to reflect the target population's demographic characteristics more accurately. This adjustment helps address imbalances resulting from non-response, enhancing the data's generalizability. However, non-response bias remains a concern, especially as weighting cannot entirely account for individuals not present in the sample or for differences in survey engagement across demographics (YouGov).

A.5 Strengths and Weaknesses of YouGov's Methodology

A.5.1 Strengths

- **Efficient Data Collection:** YouGov's nonprobability online sampling allows for efficient, cost-effective data collection, enabling timely survey responses that can address rapidly changing social and political topics.
- **Targeted Panel Management:** YouGov's proprietary panel enables consistent, direct contact with respondents, allowing for long-term engagement and improved tracking of demographic and opinion changes without needing repetitive data entry.
- **Rigorous Quality Controls:** Extensive quality checks, including IP verification, email authentication, and response validity checks, help YouGov maintain high data integrity. Respondents who fail quality checks are excluded, helping to reduce noise from unreliable data.
- **Adjustments for Representativeness:** While nonprobability sampling has limitations, YouGov partially mitigates these by applying sophisticated weighting based on benchmarks from sources like the U.S. Census. This enhances the sample's demographic alignment with the general population (YouGov).

A.5.2 Weaknesses

- **Nonprobability Sampling Limitations:** The lack of true random sampling introduces potential bias, as certain population segments may be underrepresented. For example, people without internet access or those less inclined toward online engagement are systematically excluded, which could affect representativeness.
- **Potential Selection Bias:** Although incentives are modest, they may still attract individuals who are already more engaged or motivated to participate in surveys, which could influence results. The weighting procedures help but may not fully eliminate biases if certain groups are consistently underrepresented.
- **Limitations of Online-Only Surveys:** Although internet penetration in the U.S. is high, reliance on online-only surveys may still leave out individuals less comfortable with or able to access online technology, skewing results slightly.

- **Questionnaire Design Constraints:** While YouGov’s questionnaire design is generally sound, with randomized question orders and multimedia inclusion for clarity, online surveys can limit response depth compared to in-person or phone surveys. Responses are constrained by pre-set answer options, potentially oversimplifying complex views and reducing respondent nuance (Huffington Post).

A.6 Reflection

We created an account on YouGov to gain real experience with the platform, as illustrated in Figure 3 and here is my reflection on the onboarding process and user experience.

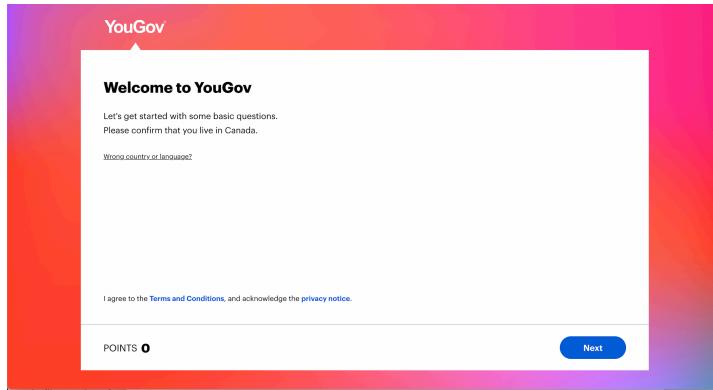


Figure 3: YouGov Welcome Page

The YouGov survey onboarding process is designed to be clear and straightforward, making it easy for new users to sign up and begin participating. The questions asked during onboarding are simple and direct, covering basic demographic details without overwhelming the participant. The first question of the first survey we took at YouGov is shown at Figure 4. This streamlined approach likely helps increase engagement and reduces drop-off rates, ensuring that users can quickly understand what is required of them. The layout and visuals are accessible, contributing to a positive first impression and making it more likely that new panelists will continue engaging with future surveys.

To enhance data integrity, YouGov includes verification steps such as confirming users’ email addresses through a code. This extra step not only ensures that participants are real and unique individuals but also helps prevent fraudulent or duplicate accounts, as shown in Figure 5, which can otherwise skew survey results. By verifying email addresses, YouGov can maintain a higher quality of data and build a more reliable panel, knowing that each participant is committed enough to complete these security measures. This verification process is a straightforward yet effective method for improving data quality from the outset.

Data privacy is also well-considered in YouGov’s onboarding, with clear options allowing users to control how their information is shared. For instance, participants can opt in or out of having

YouGov

For more than 20 years, we've been asking questions to understand what the world thinks. From celebrities to the economy, you can share your views on everything: no topic is off limits. We combine your answers with the responses of other members to create YouGov data – and it's this that powers some of the world's biggest brands. This is called aggregation.

We rely on total honesty and we appreciate the trust you give us. YouGov keeps your data safe and secure. If special types of information that could identify you as an individual are required as part of a piece of research, we will tell you in advance, and we'll always give you the option to say no.

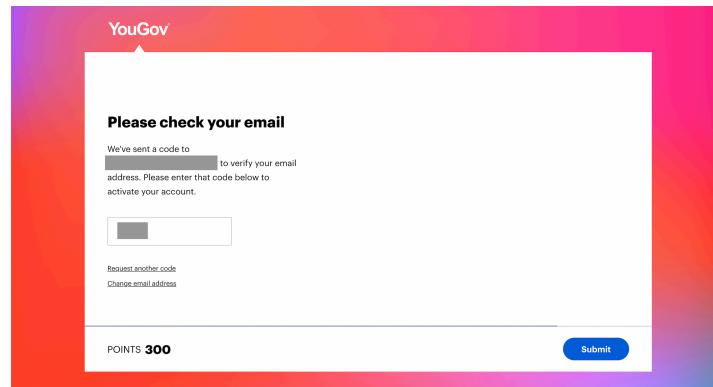
Occasionally, a trusted customer or third party might want to find other people who are similar to you, so they can show them adverts for things they may be interested in. To make this possible, a special code unique to you is shared, which can be used to find people who share similar attributes. This is called a "lookalike audience". It cannot be used to contact you, or to know exactly who you are, and it won't be used to show you adverts. The code is deleted after it's been used.

Are you happy for us to include your opinions both in aggregated data, and in an identifiable form with trusted third parties and clients for lookalike audiences?

- Yes, I am
 No, I would prefer my opinions only be included in aggregated results



Figure 4: YouGov Survey



The image shows a YouGov survey page with a red and blue gradient background. At the top, the YouGov logo is visible. Below it, a section titled "Please check your email" contains instructions: "We've sent a code to [redacted] to verify your email address. Please enter that code below to activate your account." There is a text input field for entering the verification code, followed by two small links: "Request another code" and "Change email address". At the bottom of the page, it says "POINTS 300" and has a "Submit" button.

Figure 5: YouGov Email Verification

their responses shared in identifiable formats with trusted third parties. Figure 6 includes an example of how YouGov is letting users to control their information privacy. This transparency around data usage builds trust, showing participants that their privacy is valued and that they have agency over their data. The option to decide on data sharing likely encourages long-term participation, as users are reassured that their information will be handled responsibly.

The screenshot shows a user interface titled "Partners" where users can manage their data sharing preferences with various companies. A descriptive text explains that users can set preferences for individual third-party companies, noting that checked boxes mean permission is granted. Below this, there are four sections, each with a checkbox and a company name followed by "[TCF]". Each section has a downward arrow icon to its right. The companies listed are Ad Alliance GmbH, Adform, ADITION technologies AG, and Amazon Advertising.

Company	Status
Ad Alliance GmbH [TCF]	<input checked="" type="checkbox"/>
Adform [TCF]	<input checked="" type="checkbox"/>
ADITION technologies AG [TCF]	<input checked="" type="checkbox"/>
Amazon Advertising [TCF]	<input checked="" type="checkbox"/>

Figure 6: YouGov Data Privacy

However, the onboarding interface in Figure 7 has a gamified feel, awarding points for each step completed. This reward system, while motivating, could encourage participants to treat the surveys more like a task for monetary gain than an opportunity to provide genuine responses. The point-based incentive structure may attract users who are primarily interested in earning rewards, which risks impacting data quality if participants prioritize speed over thoughtful answers. This setup introduces the possibility that some users are engaging more for the rewards than for contributing to research, which could affect the reliability of the responses YouGov collects.

B Methodology and Survey Design for 2024 U.S. Presidential Election Forecast

B.1 Methodology

This part of the appendix provides a detailed methodology and survey design to forecast the 2024 U.S. presidential election. Our survey focuses on swing states: Arizona, Georgia, Michigan,

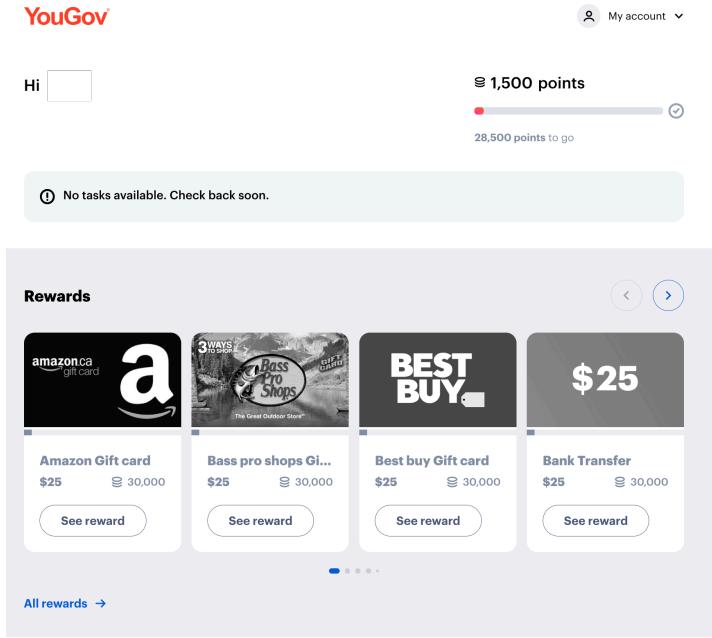


Figure 7: YouGov Interface

gan, Nevada, North Carolina, Pennsylvania, and Wisconsin due to the results of other states being certain. The survey targets different age groups using tailored recruitment strategies. Below, we break down the methodology, budget, recruitment strategies, and survey design to ensure representative participation.

B.1.1 Sampling Approach

- Total Sample Size: 7,000 respondents (1,000 participants per state).
- Target Population: Eligible voters from the following swing states: Arizona, Georgia, Michigan, Nevada, North Carolina, Pennsylvania, Wisconsin.

B.1.1.1 Sampling Breakdown by Age Group:

- 18-24 years: 150 participants per state (total 1,050) – reached through university students.
- 25-60 years: 700 participants across states (total 4,900) – reached via online ads and survey platforms.
- 60+ years: 150 participants per state (total 1,050) – reached through phone calls.
- Stratified Random Sampling:

- Stratify by age, gender, race/ethnicity, education, income, and region.
- Ensure proportional representation based on **U.S. Census data** to avoid bias.
- Weighting Strategy:
 - Apply **post-stratification weighting** to adjust for any sampling imbalance, aligning with national demographics.

B.1.2 Recruitment Plan

B.1.2.1 18-24 Years (University Students):

- University Collaboration: Partner with academic institutions in each swing state to send survey invitations via email on behalf of the project.
- Rationale: Universities provide direct access to students, improving recruitment efficiency and response rates.
- Incentive: Each respondent receives a \$10 e-gift card for a Walmart or Costco purchase.

B.1.2.2 25-60 Years (General Population):

- Online Ads and Survey Platforms: Use Facebook, Instagram, and platforms like Prolific to recruit participants from swing states.
- Budget Allocation: Allocate \$10,000 for online ads to reach residents aged 25-60.
- Incentive: Participants receive e-gift cards upon survey completion.

B.1.2.3 60+ Years (Senior Citizens):

- Phone Surveys: Conduct live calls to reach seniors, considering that older people are not comfortable using electronic devices. Phone numbers will be sourced from voter registration databases (where legally accessible) or senior community networks.
- Incentive: Seniors can choose a physical Walmart or Costco gift card, which will be sent by mail.
- Recruitment Channels:
 - Complement with **social media ads** on Facebook, Instagram, and LinkedIn to reach underrepresented groups.
 - **Incentives:** Offer **\$10 gift cards** to increase participation and engagement.

B.1.3 Trade-off

B.1.3.1 Data Representativeness vs. Control over Data Collection

- Issue: Although age quotas ensure participation from different age groups (18-24, 25-60, 60+), it is not feasible to set strict limits for race and income. Without perfectly matched samples for race and income, bias may arise in the survey results.
- Solution:
 - Weighting Strategy: Calculate population proportions for race and income within each state using Census data. Assign weights to groups based on their representation to mitigate bias.

B.1.3.2 Efficient Recruitment vs. Potential Selection Bias

- Issue: Limiting the 18-24 age group to university students may introduce selection bias. University students often differ from their non-student peers in socioeconomic background and possibly political perspectives.
- Impact: This selection bias may affect the generalizability of findings, as university students' perspectives might differ from all young adults.

B.1.3.3 Effective Senior Outreach vs. Higher Costs and Response Bias

- Issue: Using live phone operators to reach seniors increases costs, and some seniors may remain unreachable by phone.
- Impact: Higher costs may reduce resources for other age groups. Response bias may arise if unreachable seniors have different views, affecting the survey's representativeness.

B.1.3.4 Incentivizing Participation vs. Economic Bias in Sample

- Issue: Offering \$10 gift cards as an incentive could attract participants more likely to need the incentive, potentially introducing economic bias.
- Impact: Overrepresentation of lower-income participants could skew results, affecting insights if income level correlates with political perspectives.

B.1.4 Survey Implementation and Design

In designing this survey, I referenced insights from the article on survey methods from (**AnnualReviewofEconomics?**), particularly on the advantages of online surveys over in-person, phone, or mail surveys. The article highlighted that online surveys provide flexibility, allowing respondents to complete them at their convenience, which minimizes selection bias related to work schedules or availability. This flexibility is especially beneficial for students and working-age individuals, who may find it challenging to respond during traditional hours. The article also emphasized how mobile technology can increase participation by reaching populations that are otherwise hard to engage, such as younger respondents, frequent movers, or those in remote areas. In line with these insights, I designed this survey with an online format for younger and working-age participants, maximizing accessibility and reducing drop-out rates. Additionally, I applied the article's suggestion of keeping surveys concise to prevent fatigue and disengagement, which helps maintain completion rates and improve data quality. The use of digital incentives, as noted in the article, also broadens the appeal and ensures we reach a diverse participant base across income levels, providing flexibility in rewards to cater to different motivations. This approach helped shape a balanced methodology that optimizes participation across demographic groups and aligns with best practices in survey design.

- Survey Platform: Google Forms
 - **Link:** <https://forms.gle/2MGYeZavDsCNuWZ1A>
 - Accessible to respondents via **email, social media, and direct recruitment channels.**
- Survey Structure:
 1. Demographics Section: Age, gender, race/ethnicity, income, education, and state of residence.
 2. Voting Preferences Section: Candidate choice, likelihood of voting, and party affiliation.
 3. Key Issues Section: Identify priority issues (e.g., economy, healthcare, immigration).
 4. Thank You Message:
 - *“Thank you for completing the survey! Your input is greatly appreciated and will help provide insights into the upcoming 2024 election.”*

B.1.5 Data Validation

- Techniques for Data Quality:
 1. Screening Questions: Confirm eligibility (e.g., 18+ years old, registered voter status).

2. Attention Checks: Include a question like “Please select ‘Agree’ for this item” to verify respondents are attentive.
3. IP Geolocation: Validate state residency based on reported location.
4. Duplicate Detection: Identify and remove duplicate responses.

B.1.6 Poll Aggregation and Reporting

- Poll Aggregation:
 - Use **weighted averages** to account for differences in sample size and demographics.
- Margin of Error**:
 - National Margin of Error: ±1% at the 95% confidence level.
 - State-Level Margins: ±5-10% depending on the sample size for each state.

B.1.7 Budget Allocation

Expense	Estimated Cost
Participant Incentives	\$70,000
Online Ads and Panel Provider Fees	\$10,000
Phone Survey Salaries	\$8,000
Google Forms (Platform)	Free
Data Validation & Analysis	\$10,000
Miscellaneous Expenses	\$2,000
Total	\$100,000

B.2 Survey Questions

Below is the full content of the survey to be implemented using Google Forms:

1. What is your age?
 - 18-24
 - 25-39
 - 40-60
 - 60+

2. What is your gender?

- Male
- Female
- Other: _____

3. What is your race/ethnicity?

- White
- Black
- Hispanic or Latino
- Asian
- Indigenous
- Other: _____

4. What is your highest level of education?

- Less than high school
- High school diploma
- Some college
- Bachelor's degree
- Graduate degree or higher

5. What is your annual household income?

- Less than \$25,000
- \$25,000 - \$49,999
- \$50,000 - \$99,999
- \$100,000 or more

6. Which state do you currently reside in?

7. Are you a registered voter?

- Yes
- No
- Not sure

8. How likely are you to vote in the 2024 presidential election?

- 1 (Definitely will not vote)
- 2
- 3
- 4
- 5 (Definitely will vote)

9. If the 2024 election were held today, who would you vote for?

- Kamala Harris (Democrat)
- Donald Trump (Republican)
- Undecided
- Other: _____

10. What is the most important issue for you in this election?

- The economy
- Healthcare
- Immigration
- Climate change
- Social Security and Medicare
- Foreign policy

11. In your opinion, how will the majority in your state vote in 2024?

- Democrat
- Republican
- Too close to predict

12. Do you have any additional comments or suggestions?

Thank you for completing the survey! Your input is greatly appreciated and will help provide insights into the upcoming 2024 election.

B.3 Google Forms Link

<https://forms.gle/2MGYeZavDsCNuWZ1A>

C Model details

C.1 Posterior predictive check

C.2 Diagnostics

D References