Evaluating the Effectiveness of Persona Simulation in Opinion Prediction with GPT-4.1

Sarah Li
McLean High School
McLean, VA, USA
sarahyl2018@gmail.com

Ziyu Yao
Department of Computer Science
George Mason University, VA, USA
ziyuyao@gmu.edu

2025 IEEE International Conference on Data Mining Workshops (ICDMW) | 979-8-3315-8132-9/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICDMW69685.2025.00377

***Abstract*—Persona simulation involves utilizing large language models (LLMs) to anticipate human choices or interactions based on specific characteristic information. To further understand current limitations and future directions, we tested persona simulation in opinion prediction with GPT-4.1 (knowledge cutoff by June 2024). Using personas from nine U.S. states provided by Columbia University's Personas dataset, GPT-4.1 accurately predicted 2024 election outcomes in eight out of the nine states, only failing in one of the swing states. We then focused on opinions related to medicine and healthcare. With the American Trends Panel Wave 123 dataset from Pew Research Center, GPT-4.1 was able to anticipate beliefs about childhood vaccines with an accuracy of up to 0.94. Furthermore, we applied GPT-4.1 to generate conversations among personas and observed that the simulated dialogues and opinions adhered well to personas' personalities and backgrounds, albeit lacking natural human-like flow. Persona simulation proves to be a promising application of artificial intelligence as long as biases are addressed. In the near future, it will be beneficial to apply it to opinion analysis and reaction prediction in diverse fields ranging from public health to lawmaking to economics.**

***Index Terms*—persona simulation, large language models, generative artificial intelligence, social sciences.**

I. Introduction

Persona simulation involves utilizing large language models (LLMs) to anticipate human choices or interactions based on characteristic information about demographics, background, or personality. Generated samples can be used as additional data points in surveys, addressing the challenges that come with data collection, such as imbalanced sampling or non-response bias. Applications include marketing [1], political science [2], social science [3], and more. Promise has also been shown by Park et al. in simulating human conversations and interactions in a virtual town environment [4] and Yue et al. in an educational setting [5]. This creates possibilities for improved dialogue personalization with chatbots in a variety of use cases.

However, persona simulation has been noted to show bias, overestimate homophily, and assume monolithic qualities for groups [3], [6]. Concerns have also been raised about the difficulty of generating personas that can truly represent the population of interest in terms of intersectional demographics; census data, for instance, only reports marginal distributions without showing the joint distributions across different attributes [6]. Representing personality traits is also challenging, and current experiments rely on tests such as the Big Five to capture characteristics [6], [7].

In this work, we seek to understand the current promises and limitations of LLM-based persona simulation by evaluating GPT-4.1 [8] in opinion prediction: election forecasting, medical opinion prediction, and dialogue simulation. For election prediction, we looked at overall outcomes as well as voting distributions for nine states. Medical opinion prediction involved assessing accuracies for four pertinent questions relating to vaccines and healthcare. For dialogue simulation, we generated conversations between three personas to evaluate how closely dialogues adhered to the given personas.

We found that opinion prediction with GPT-4.1 can reach high accuracies and align closely with the ground truth, but the underlying bias is undeniable, especially with election forecasting. GPT-4.1 relied heavily on over-generalizations of demographic groups when predicting votes. With dialogue generation, personas' personalities were not captured well, and extracted opinions relied on over-generalizations as well. Additionally, generated conversations sounded unrealistic and had a manufactured cheerfulness, limiting the model's potential use for simulations using current methods.

Yet persona simulation still opens many possibilities aside from politics, healthcare, and conversations, including but not limited to forecasting commerce choices [9] and predicting effectiveness ratings of advertisements or public service announcements [10]. We hope to see persona simulation grow as existing techniques and frameworks are refined.

II. Related Work

As LLMs have become more capable, more research has turned to persona simulation as a viable application. Personas can come in various forms that are used in different situations; as outlined in Chen et al.'s work [11], persona types include demographic personas, character personas, and individualized personas. In our project, our focus is on demographic personas. Similar simulations have used personas as "silicon samples" to create realistic population representations in contexts of marketing [1], politics [2], [6], social network simulation [3], and more.

Previous work has focused on persona simulation in election prediction. Argyle et al. used the American National Election Studies (ANES) dataset to predict the results of the 2012,

2375-9259/25/$31.00 ©2025 IEEE
DOI 10.1109/ICDMW69685.2025.00377

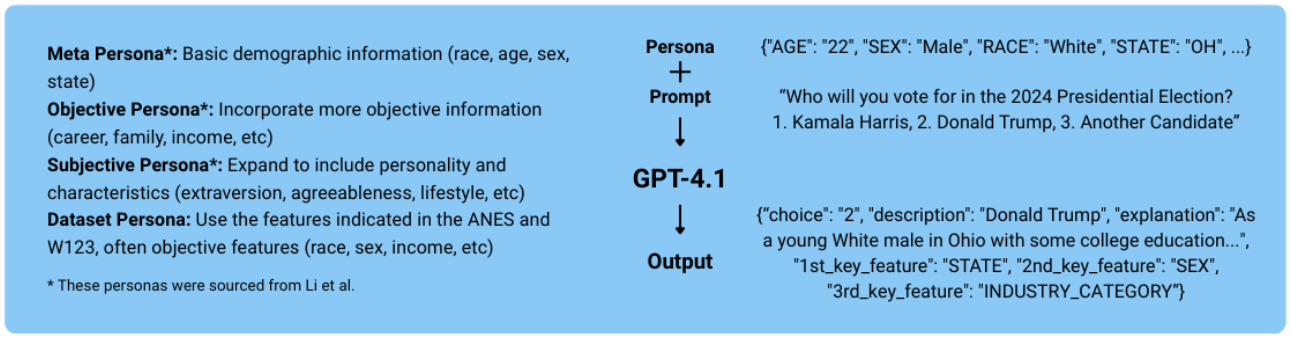

Fig. 1. The persona simulation framework.

2016, and 2020 elections [2]. Li et al., on the other hand, created personas for each state using census data and used those to predict election results for 2016, 2020, and 2024 [6]. However, both did not highlight state-by-state voting distributions, which is a key focus in our work. We also noted important features that impacted GPT-4.1's predictions, providing insight into the reasoning behind decisions.

In addition, the existing literature has had a limited focus on anticipating medical or health-related opinions, instead focusing on simulating patient-doctor interactions [12]. In this project, one of our approaches focused specifically on opinions related to vaccines and healthcare in the United States; however, our other approaches study predictions related to political opinions and dialogues, which thus provide a more comprehensive understanding of GPT-4.1's capability for persona simulation.

Dialogue generation has been a highly tested field with the emergence and popularization of LLMs. Chatbots designed to behave like famous celebrities or fictional characters are a common example [13]. When communicating with LLMs impersonating personas across ages and expertises, Salewski et al. have noted that impersonation is effective for making conversations more realistic, but may introduce race and gender bias [14]. In this project, we evaluate the ability of GPT-4.1 to represent the personality and beliefs of a persona in the context of a conversation, and evaluate whether any biases exist.

III. Persona Simulation Framework

Effective persona simulation relies on representing the population using diverse characteristics. The personas we used were extracted from the Personas dataset (PERSONAS) from Columbia University [15], the 2024 American National Election Studies (ANES) dataset [16], and the American Trends Panel Wave 123 (W123) dataset from Pew Research Center [17].

Overall, we used four types of personas: meta, objective, subjective, and dataset. The first three descriptions below are built upon Li et al.'s definitions, methods, and the PERSONAS dataset they provided [6], [15].

Meta Personas: These provide characteristics that can be found in census data — race, age, and sex — and were developed using realistic census distributions from the 48 states of the United States mainland. Each state has 1000 meta personas.

Objective Personas: These extend meta personas to include more information about demographics and lifestyle. They include characteristics such as income, marital status, education level, household language, and more.

Subjective Personas: In addition to the information from objective personas, these add information about personality, such as Big Five scores, political views, religion, and ability to speak English.

Dataset Personas: These personas use the features provided by a given dataset; in this study, the additional datasets we used were ANES and W123, which had 5521 and 10,701 randomly sampled individuals, respectively. ANES included 2024 election information and W123 included opinions on healthcare. Most of the information provided by those datasets would be in the realm of objective personas, with features such as income, education level, gender, race, and more.

The additional information in objective and subjective personas was generated by Llama-3.1-70B [18] given existing meta personas. For the objective personas, the prompt gave multiple-choice options for each feature. For subjective personas, however, most features were open-ended, with instructions to remain "reasonable and succinct". We refer readers to Li et al. [6] for more details.

For opinion simulations, personas were fed one by one in their existing json format to GPT-4.1 using the same prompt for all. All questions were treated as forced multiple-choice to prevent GPT from declining to answer. Outputs consisted of the predicted choice, an explanation, and the extracted top three features most contributing to the selected choice. This framework can be seen in Fig. 1.

IV. Experimental Setup

A. Election Forecasting

Election forecasting was conducted using two different approaches: one using personas from PERSONAS [15], and the other using personas from ANES [16].

1000 meta personas and 1000 objective personas from nine states — three red states (Alabama, Texas, Wyoming), three blue states (Virginia, Maryland, California), and three swing states (Pennsylvania, North Carolina, Nevada) — were extracted from PERSONAS. As PERSONAS lacked a ground truth vote for each persona, we relied on state results and voting distributions [19] as evaluation markers.

ANES provided 2024 voting preferences and related information for 5521 adult United States citizens. We extracted 11 columns: sex, gender, race, race of spouse, education level, household income, place of birth, sexual orientation, transgender status, number of children, and vote. We then removed any samples that had missing values for columns, ending up with 2882 people. Of the 2882 people remaining, 1532 voted for Kamala Harris, 1289 for Donald Trump, 15 for Jill Stein, 9 for Cornel West, and 37 for alternate candidates. Using these samples, we then performed persona simulation using multinomial logistic regression and GPT-4.1.

For feature importance analysis with logistic regression, we used absolute values of coefficients to sort importance. For GPT-4.1, as part of the output of each generated prediction, we asked for the vote, an explanation, and the top three key features contributing to the final voting decision.

GPT-4.1 has a knowledge cutoff of June 2024 [8], making it suitable for the prediction of the 2024 election. Despite some literature suggesting that using GPT to predict earlier elections does not impact accuracy or reliability, it may still introduce unseen effects [2], [6]. Thus, we decided to focus only on the 2024 election and ignore elections that occurred earlier.

B. Healthcare Opinion Prediction

For opinion prediction, we used W123 [17], which included various questions about government, healthcare, and emerging technologies. The original dataset included 10,701 patients and 145 questions. After removing questions where greater than 50% of patients declined to answer and then removing any patients who declined to answer questions, we were left with 7992 patients and 110 questions. We isolated four major questions that we wanted to focus on, pertaining to vaccines and the healthcare system. See Table I for the questions.

For the features of the dataset personas, we extracted 33 objective features for each persona. We then randomly selected 100 persona samples to perform the simulation. The "% of 1" column refers to the percentage of individuals who responded with a choice of "1" to the question. For evaluation, in addition to accuracy, we used F1 score to account for class imbalance as seen in some questions in Table I.

When scores remained low, we incorporated an additional 30 features for each persona, making a total of 63 features. These features provided more information about personas' healthcare habits and beliefs, improving GPT-4.1's performance.

TABLE I
Four highlighted questions from W123 to test healthcare opinion simulation

Label	Question	Options	% of 1
CHILD	Overall, do you think...	1: "The benefits of childhood vaccines for measles, mumps, and rubella outweigh the risks"; 2: "The risks outweigh the benefits"	90%
COVID	Overall, do you think...	1: "The benefits of COVID-19 vaccines outweigh the risks"; 2: "The risks outweigh the benefits"	68%
CHOICE	Which comes closest to your point of view?	1: "Parents should be able to decide NOT to vaccinate their children"; 2: "Healthy children should be required to be vaccinated in order to attend public schools"	30%
HEALTH	Which comes closest to your point of view?	1: "Medical treatments these days are worth the costs because they allow people to live longer and better quality lives"; 2: "Medical treatments these days often create as many problems as they solve"	51%

C. Persona Conversation Simulation

For conversation generation, we randomly selected 3 subjective personas from PERSONAS and prompted GPT-4.1 to generate a realistic conversation given a topic. The topics we used were education in schools, vaccines, and the treatment of transgender people, as well as more common and unstructured topics: daily routines, hobbies, and "no topic."

Knowing that we did not specifically prompt GPT-4.1 to sound human, our evaluations mostly focused on the following two questions:

1) Is this conversation realistic? Would humans interact in the way that these personas do?
2) Are the personas' personalities and backgrounds reflected in their simulated words and responses to others?

V. Experimental Results

A. Election Forecasting

With samples from PERSONAS, we used GPT-4.1 to predict votes for nine different states in the 2024 election (Trump vs. Harris), the distributions of which are shown in Fig. 2. Using meta personas, eight of the nine state outcomes were predicted correctly, with the exception of Nevada, a swing state. Using objective personas, however, only five out of nine state outcomes were predicted correctly. In all states, the predicted objective persona state distribution had a higher percentage of blue votes. A similar phenomenon was observed by Li et al., so this may be due to bias in the generation of objective personas or bias in the voting simulation [6]. This may suggest that including too much information in personas may lead to misleading results, and balance is needed.

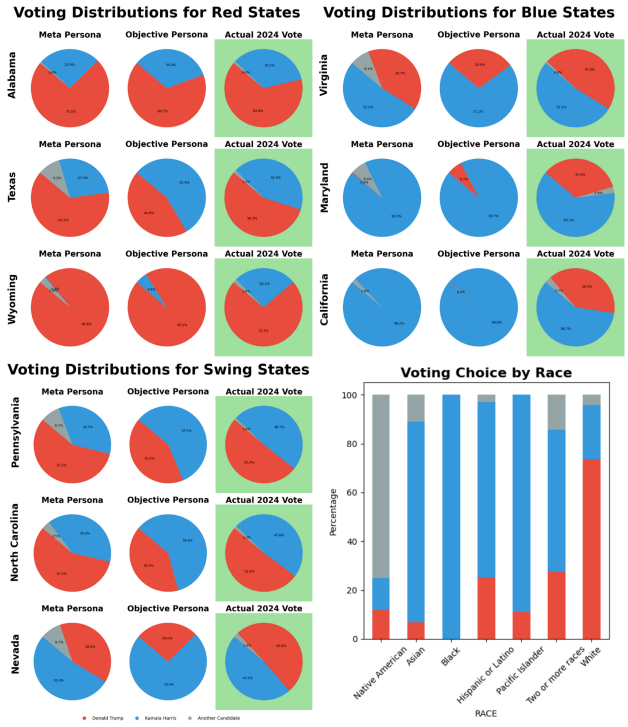

Fig. 2. Voting distributions for nine states and across races.

Despite successfully predicting voting outcomes in eight states using meta personas, voting distributions differ from the ground truth. In states that have historically voted red or blue, the difference was especially apparent. In California, for example, 98.2% of personas were predicted to vote blue, with no red votes. In the actual 2024 election, 58.7% of votes went to Harris, and 38.5% voted for Trump. A similar trend occurs with Maryland, another blue state, and Wyoming, a red state. In general, predictions also overestimate the proportion of votes going to third-party candidates.

Digging deeper into the distribution of votes by race using meta personas, more trends become clear. All Black voters were predicted to support Harris, in comparison to the ground truth of 83% [20]. Additionally, 75% of Native Americans were predicted to vote for third parties, and 73.5% of Whites for Trump, both of which are overestimates. These show the biases that impact the true performance of persona simulation, underneath the state-by-state results.

Persona simulation with ANES followed a different format, as voting truth labels were given for each sample. Our main goal was to test logistic regression and GPT-4.1 as viable options for persona simulation and also analyze the important features used by each model. Overall, logistic regression had an accuracy of 0.648, and GPT-4.1 had an accuracy of 0.610, showing room for improvement. Confusion matrices and important features can be seen in Fig. 3.

The four most important features for logistic regression were household income, sexual orientation, gender, and education level. On the other hand, the top four features for GPT-4.1 were education level, household income, race, and place of childhood (referring to where the persona grew up). For logistic regression, the features mostly follow a smooth decline of importance, but for GPT-4.1, the top four features are considered very heavily compared to the rest. Since GPT-4.1 performs worse than logistic regression, this may indicate that GPT is focusing on the wrong features, causing it to struggle. These reveal hidden influences underlying every prediction, and emphasize the importance of bias analysis for all models to increase transparency.

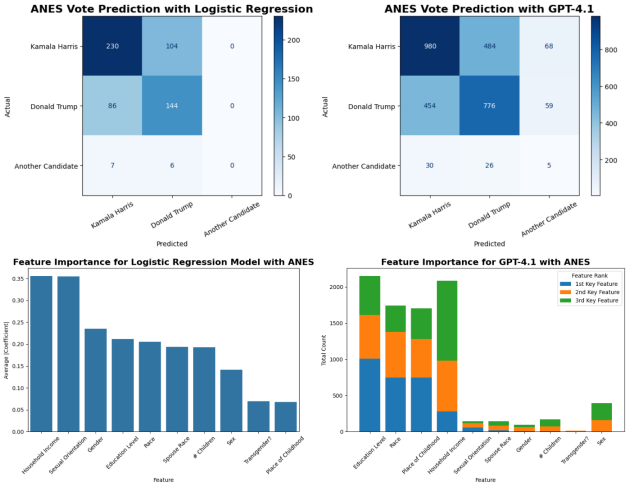

Fig. 3. Confusion matrices and feature importances for ANES using logistic regression and GPT-4.1.

B. Healthcare Opinion Prediction

When using GPT-4.1 to predict healthcare-related opinions from W123, scores were difficult to improve, especially for questions with split responses, such as HEALTH (refer to Table I). After adding more information about each persona's vaccination habits and other healthcare-related beliefs, scores skyrocketed, as shown in Table II. At its peak, GPT-4.1 was able to anticipate beliefs about childhood vaccines with an accuracy of up to 0.94 and an F1 score of up to 0.8468.

Across Table II, it is clear that the HEALTH question (referring to whether or not medical treatments are worth the costs or not) causes GPT-4.1 to perform the worst. It is a very divisive question, with 51% believing treatments are worth the costs, while 49% disagree. It is also affected by

TABLE II
THE ACCURACIES AND F1 SCORES OF OUR W123 QUESTIONS BEFORE AND AFTER ADDING ADDITIONAL HEALTHCARE INFORMATION

	Before		After	
Label	Acc	F1	Acc	F1
CHILD	0.89	0.6036	0.94	0.8468
COVID	0.82	0.7532	0.88	0.8698
CHOICE	0.75	0.6710	0.81	0.7759
HEALTH	0.59	0.5890	0.75	0.7442

personal experiences, which are difficult to capture through personas. Thus, we realize that GPT-4.1 struggles most with controversial questions where most people are evenly split.

C. Persona Conversation Simulation

Dialogue generation involved randomly selecting three subjective personas from PERSONAS and prompting them to have a conversation around a provided topic. We used a combination of structured and unstructured topics, ranging from the education system to hobbies.

In general, we noticed that the simulated conversations were very rigid, with each persona taking turns to speak without variation in the order of speaking. For one particular topic, which was about the treatment of transgender people, the conversation read like a scripted refute of transphobia, with one persona raising concerns and the other two swiftly batting them down. Other similar dialogues contributed to decreasing the realism of generated conversations.

The beliefs that the conversations showed adhered to the personas, and each persona showed unique hobbies and occupations. GPT-4.1 was acceptable at portraying objective traits, but struggled with more subjective ones. No matter what personality each persona had, the dialogue read the same. Rarely was there an emotion other than optimism, and introverted and extroverted personas showed no difference in their manner of speaking.

VI. Conclusion and Future Work

Persona simulation can reach high accuracies, but a closer look into those results reveals bias and overgeneralization. For election forecasting, we observed biases by state and race, and GPT-4.1 focused on different persona features than the more impartial logistic regression. Additionally, current models have room for improvement in generating conversations that sound more realistic and adhere to given personalities.

Park et al. has an effective framework in which individuals are interviewed on their personal experiences, which are then used to predict responses to questions [7]. That approach eliminates bias that may occur when LLMs generate subjective traits for personas; indeed, Li et al. noted that as LLMs had more influence in the generation of personas, those personas tended to lean left, have skewed opinions, and include more positivity [6]. Park et al.'s approach also incorporates more life experiences and personality compared to more tabular approaches for capturing information about a persona [7]. We would like to perform future work with that framework, given more time and resources.

Developing sophisticated metrics to assess generated dialogues is also important, as our current methods remain quite subjective. More objective analyses would bring credence to our findings and allow for scientific growth.

In the future, more work needs to be done in bias analysis, especially looking at different opinion distributions for each demographic group. Increasing accuracy and other evaluation metrics is important and exciting, but should not come at the cost of transparency or fairness.

Overall, persona simulation is a promising use of artificial intelligence as long as biases are addressed. It will be beneficial to apply it to opinion and reaction prediction in diverse fields, from lawmaking to economics to quality assurance, enabling more informed decisions and personalized interventions.

Acknowledgment

We gratefully acknowledge the administrative and financial support of George Mason University's Aspiring Scientists Summer Internship Program (ASSIP), which facilitates hosting high school students in Mason research groups.

References

[1] M. Sarstedt, S. Adler, L. Rau, and B. Schmitt, "Using large language models to generate silicon samples in consumer and marketing research: Challenges, opportunities, and guidelines," in *Psychology & Marketing*, 2024, pp. 1254–1270.

[2] L. P. Argyle, E. C. Busby, N. Fulda, J. Gubler, C. Rytting, and D. Wingate, "Out of One, Many: Using Language Models to Simulate Human Samples," in *Political Analysis*, 2023, pp. 337–351.

[3] S. Chang, A. Chaszczewicz, E. Wang, M. Josifovska, E. Pierson, and J. Leskovec, "LLMs Generate Structurally Realistic Social Networksbut Overestimate Political Homophily," in *Proceedings of the Nineteenth International AAAI Conference on Web and Social Media*, 2025, pp. 341–371.

[4] J. S. Park, J. O'Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein, "Generative agents: Interactive simulacra of human behavior," in *Proceedings of the 36th annual acm symposium on user interface software and technology*, 2023, pp. 1–22.

[5] M. Yue, W. Lyu, W. Mifdal, J. Suh, Y. Zhang, and Z. Yao, "MathVC: An LLM-Simulated Multi-Character Virtual Classroom for Mathematics Education," *arXiv preprint arXiv:2404.06711*, 2024.

[6] A. Li, H. Chen, H. Namkoong, and T. Peng, "LLM Generated Persona is a Promise with a Catch," in *arXiv preprint arXiv:2503.16527*, 2025.

[7] J. S. Park *et al.*, "Generative Agent Simulations of 1,000 People," in *arXiv preprint arXiv:2411.10109*, 2024.

[8] OpenAI. (2025) Introducing GPT-4.1 in the API. [Online]. Available: https://openai.com/index/gpt-4-1/

[9] S. Mansour, L. Perelli, L. Mainetti, G. Davidson, and S. D'Amato, "PAARS: Persona Aligned Agentic Retail Shoppers," in *Proceedings of the 1st Workshop for Research on Agent Language Models (REALM 2025)*, 2025, pp. 143–159.

[10] P. Sheeran *et al.*, "Artificial Intelligence Simulation of Adolescents' Responses to Vaping-Prevention Messages," in *JAMA Pediatrics*, 2024.

[11] J. Chen *et al.*, "From Persona to Personalization: A Survey on Role-Playing Language Agents," *Transactions on Machine Learning Research*, 2024.

[12] D. Kyung *et al.*, "PATIENTSIM: A Persona-Driven Simulator for Realistic Doctor-Patient Interactions," in *arXiv preprint arXiv:2505.17818*, 2025.

[13] N. Shazeer and D. de Freitas, "Character.AI," https://character.ai/, 2025.

[14] L. Salewski, S. Alaniz, I. Rio-Torto, E. Schulz, and Z. Akata, "In-Context Impersonation Reveals Large Language Models' Strengths and Biases," in *Proceedings of the 37th International Conference on Neural Information Processing Systems*, 2023, pp. 72 044 – 72 057.

[15] Tianyi-Lab. (2025) Personas. [Online]. Available: https://huggingface.co/datasets/Tianyi-Lab/Personas

[16] American National Election Studies. (2025) ANES 2024 Time Series Study Full Release [dataset and documentation]. [Online]. Available: https://electionstudies.org/data-center/2024-time-series-study/

[17] Pew Research Center. (2023) American Trends Panel Wave 123. [Online]. Available: https://www.pewresearch.org/dataset/american-trends-panel-wave-123/

[18] Meta AI, "Llama-3.1-70b large language model," https://huggingface.co/meta-llama/Llama-3.1-70B, 2024, released July 23, 2024.

[19] CNN Politics. (2024) Election 2024: Presidential results. [Online]. Available: https://www.cnn.com/election/2024/results/president?election-data

[20] H. Hartig, S. Keeter, A. Daniller, and T. Van Green, "Voting patterns in the 2024 election," in *Pew Research Center*, 2025.