Will Trump Win in 2024? Predicting the US Presidential Election via Multi-step Reasoning with Large Language Models

Chenxiao Yu[1], Zhaotian Weng[1], Zheng Li[2], Xiyang Hu[3†], Yue Zhao[1†]
[1]University of Southern California [2]Arima [3]Carnegie Mellon University
{cyu96374, wengzhao, yzhao010}@usc.edu, winston@arimadata.com, xiyanghu@cmu.edu
[†]Corresponding authors

Abstract

Can Large Language Models (LLMs) accurately predict election outcomes? While LLMs have demonstrated impressive performance in various domains, including healthcare, legal analysis, and creative tasks, their ability to forecast elections remains unknown. Election prediction poses unique challenges, such as limited voter-level data, rapidly changing political landscapes, and the need to model complex human behavior. To address these challenges, we introduce a multi-step reasoning framework designed for political analysis. Our approach is validated on real-world data from the American National Election Studies (ANES) 2016 and 2020, as well as synthetic personas generated by the leading machine learning framework, offering scalable datasets for voter behavior modeling. To capture temporal dynamics, we incorporate candidates' policy positions and biographical details, ensuring that the model adapts to evolving political contexts. Drawing on Chain of Thought prompting, our multi-step reasoning pipeline systematically integrates demographic, ideological, and time-dependent factors, enhancing the model's predictive power. Also, we apply our framework to predict the outcome of the *2024 U.S. presidential election in advance*, demonstrating the adaptability of LLMs to unseen political data.

Important Notice

- **Ongoing Work:** This research is ongoing as of October 22, 2024.
- **Research Integrity:** This research is conducted independently w/o any funding.
- **Content Warning:** This paper may contain some offensive content generated by LLMs.
- **Disclaimer:** This study explores LLMs' capacity in election forecasting. Predictions do not reflect the authors' views and should not be interpreted as definitive forecasts.

1 Introduction

Large Language Models (LLMs) have demonstrated remarkable capabilities across various domains, including natural language understanding, content generation, etc. (Brown et al., 2020). Their potential extends far beyond mere text processing, including a broad spectrum of applications from medical diagnostics (Zhang et al., 2023) to legal analysis (Chalkidis et al., 2022) and creative domains (Yang et al., 2022). This versatility stems not only from LLMs' ability to understand and generate text but also from their capacity to leverage large amounts of common knowledge (Roberts et al., 2020), simulate diverse personas (Hu and Collier, 2024), and effectively model human behavior in complex social science tasks (Bommasani et al., 2021). Specifically, LLMs have shown promising results in capturing human-like common sense reasoning (Zhou et al., 2020; AlKhamissi et al., 2022) and have been successfully applied to simulate human decision-making in various contexts (Zhou et al., 2023; Ziems et al., 2024). These multifaceted capabilities position LLMs as potential tools for simulating human decision-making processes in complex contexts. Recent research has begun exploring LLMs' political science applications, analyzing policy documents, campaign speeches, and public sentiment (Xu, 2022; Haq et al., 2023). While the text-based nature of political data certainly aligns with LLMs' strengths, it is the models' holistic combination of language understanding, knowledge integration, and human-like reasoning that truly underscores their potential for simulating complicated dynamics of political-related decision-making (Argyle et al., 2023; Bisbee et al., 2024).

Motivation. Despite LLMs' success in the above straightforward political science tasks, their capacity to handle more complex tasks like election prediction remains uncertain (Lerer et al., 2022). In-

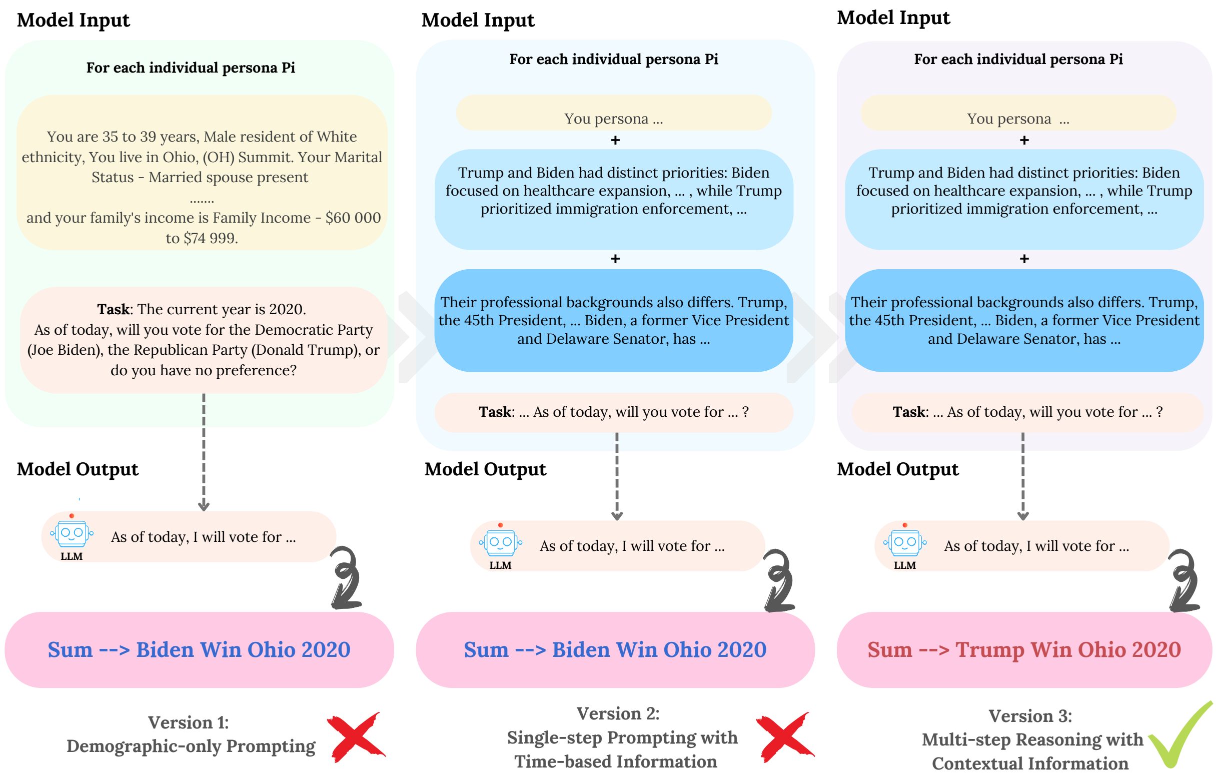

Figure 1: Demonstration of three prompt designs in §3.2. V1 is the direct prompt on voter demographic information, while V2 introduces time-dependent information to capture candidates' agenda and V3 also uses multi-step reasoning. In this example for 2020 Ohio result prediction, only V3 can accurately predict the results, demonstrating the importance of leveraging both time-dependent information and multi-step reasoning for election result prediction.

deed, the potential for LLMs to accurately predict election results is an intriguing prospect, given their ability to process vast amounts of historical information and their success in other predictive tasks. However, election forecasting presents unique challenges that test the limits of LLM capabilities. First, the high cost of acquiring voter-level data makes conducting experiments and verifying models in election prediction research challenging. Second, unlike many other predictive tasks, election forecasting requires modeling individual voter behavior as well as the candidates', which is inherently difficult and shifting with time. It remains uncertain whether text-based data alone can capture this complexity (Graefe, 2014). Third, accurate election forecasting requires reasoning beyond simple inference, integrating multiple factors such as economic trends, political events, and demographic changes (Holbrook, 2016). The capacity of LLMs to perform sophisticated reasoning for accurate election predictions is an open question (Wei et al., 2022).

Our Solution. To address the challenges of using LLMs for election predictions, we propose a novel approach that leverages their strengths while mitigating limitations in data availability, time-varying factors, and complex political dynamics. First, to overcome the scarcity of detailed voter-level data, we employ the *Sync* synthetic data generation framework (Li et al., 2020b), which probabilistically reconstructs individual-level demographic and behavioral profiles from aggregated public datasets. We complement this synthetic data with real-world datasets, such as the American National Election Studies (ANES) 2020 Time Series (Studies, 2022), ensuring our approach reflects real voting behaviors. Second, our solution adapts to evolving political contexts by incorporating time-dependent factors. Specifically, we aggregate information from presidential campaign data, such as candidates' policy agendas and biographical backgrounds, to align our model with changing political landscapes (Holbrook, 2016). Third, we introduce a multi-step reasoning framework tailored for election prediction. Inspired by Chain of Thought prompting (Wei et al., 2022), this framework decomposes the prediction process into intermediate steps, enabling the model to systematically integrate demographic information, ideological alignment, and time-sensitive factors. This multi-step reasoning improves the model's accuracy by addressing biases and overfitting issues observed with simpler approaches. Each component of our framework builds progressively based on observations and refinements. As shown in Figs. 1 and 2, we iterate through multiple pipeline versions to develop our final pipeline. This final version demonstrates significant improvements in both predictive accuracy and alignment with real-world results (Figs. 3, 4, and 5), outperforming other pipelines across all.

Our technical contributions include:

1. **First Large-Scale LLM-based Election Prediction Framework**. This work establishes a new frontier in election forecasting by demonstrating how LLMs can model voter behavior using a combination of real-world data and synthetic datasets, capturing voter-level dynamics with significant scale and detail.

2. **Novel Multi-Step Reasoning Framework**. We introduce a novel multi-step reasoning process tailored specifically for political forecasting. This framework enhances the model's ability to integrate and analyze over 11 critical, time-sensitive features—such as policy agendas and candidates' backgrounds—allowing for more complex, context-aware predictions.

3. **New Insights and Future Directions**. Our analysis uncovers essential insights into the strengths and limitations of LLMs in election prediction, including potential ideological biases and the challenges of temporal modeling. Future research will explore the integration of multiple LLMs for comparative analysis and further refinement of prompting to improve prediction reliability and robustness.

2 Related Work

2.1 LLMs in Political Science: A New and Emerging Field

The application of LLMs in political science represents a new and rapidly evolving field, with a limited but growing body of research. While LLMs have revolutionized natural language processing and various other domains, their potential in political science remains largely untapped. Initial studies have demonstrated promising results in areas such as election forecasting, policy analysis, and public opinion simulation (Smith and Doe, 2023; Johnson and Lee, 2024). However, political science often involves complex social dynamics and multi-layered causal relationships, posing significant challenges for effectively utilizing LLMs in this context (Brown, 2023).

Recent work by Chen and Rodriguez (2024) highlights the potential of LLMs in decoding political speeches and policy documents, emphasizing the need for domain-specific fine-tuning to capture the subtleties of political language. Future research is likely to focus on designing models that can handle the intricacies of political discourse while ensuring robustness against biases and misleading inferences (Wilson et al., 2024; Thompson, 2023).

2.2 Political Election Research: Classical and Agent-Based Approaches

Traditional political science literature has long relied on survey data and statistical models to analyze voter behavior and predict election outcomes. Classical models like the Downsian spatial model and the Median Voter Theorem explain election results by assuming voters' positions in policy space (Downs, 1957; Black, 1948). Time-series regression models also play a key role in long-term election forecasting, analyzing the relationships between economic indicators, political events, and voter sentiment (Lewis-Beck and Rice, 1992; Erikson and Wlezien, 2016).

In recent years, Agent-Based Models (ABMs) have gained traction in election research. These bottom-up approaches simulate individual voter decisions and social interactions, providing a more granular view of electoral dynamics. For instance, Gao et al. (2022a) introduced an agent-based election prediction model that captures social networks and voter-candidate interactions to offer more accurate election forecasts. This approach builds upon earlier work by Lemos et al. (2019), who demonstrated the effectiveness of ABMs in modeling voter turnout and preference formation.

ABMs excel at modeling heterogeneity and dynamic processes in real-world voting scenarios, offering greater flexibility and granularity than traditional statistical models. The work of Collins and Martinez (2023) further illustrates how ABMs can incorporate complex factors such as media influence and peer effects in voter decision-making processes. As the field evolves, future election studies may integrate ABMs with LLMs to balance large-scale data analysis with individual voter behavior modeling, as proposed by Zhang et al. (2024) in their hybrid forecasting framework.

2.3 LLMs in Political Science: Current Research and Our Contributions

Although LLMs have seen rapid advancements, their application in political science remains limited. Only a small number of studies have explored how LLMs can be used for tasks like election prediction, policy analysis, and public opinion tracking. Political text analysis is an important area, and some early benchmark datasets are starting to emerge. However, political language is often complex, with nuanced meanings and context, which presents a significant challenge for LLMs (Anderson and Taylor, 2023; Williams et al., 2024).

One notable example is the "Political Campus" project by Roberts et al. (2023), which developed a benchmark dataset specifically for election prediction and evaluated LLM performance on various election-related questions. This work has been instrumental in highlighting both the potential and limitations of LLMs in political forecasting. Similarly, the research by Kim and Patel (2024) on using LLMs for policy sentiment analysis demonstrates the models' capacity to process large volumes of public opinion data, while also underscoring the need for careful interpretation of results.

In contrast to the existing work, our research provides a more comprehensive approach. In 2024, we introduced an integrated framework that combines LLMs with Agent-Based Models to predict election outcomes and analyze voter behavior. Unlike previous studies, our approach not only focuses on semantic analysis of political texts but also incorporates individual voter behavior modeling, offering more granular and accurate election predictions. This integration addresses the limitations identified by Lopez and Singh (2023) regarding the disconnect between macro-level language models and micro-level voter behavior.

3 Using LLMs for Election Result Prediction

How can we effectively leverage LLMs to predict election results? In this work, we simulate *each voter's decision-making process* by providing LLMs with detailed voter information and asking them to predict voter preferences based on that data. To achieve this, we focus on two key aspects: (1) establishing an evaluation framework with appropriate datasets that contain voter-level information, and (2) designing an LLM-based pipeline for accurate election predictions. In §3.1, we introduce the datasets used in this study and describe the details of our pipeline evaluation process. We then provide an overview of our design approach in §3.2, with a discussion of three progressive pipelines in §3.2.1, §3.2.2, and §3.2.3, ranging from simple prompting to multi-step reasoning based on observations. Finally, we evaluate the performance of these three pipelines on two datasets in §3.3.

3.1 Datasets, Evaluation, and Settings

Before presenting the pipelines for election prediction, we first describe the datasets, establish the evaluation framework, and introduce the experimental settings. In this work, we use two data sources: (1) real-world American National Election Studies (ANES) 2016 and 2020 Time Series data (Studies, 2019, 2022), and (2) voter-level synthetic data generated using advanced machine learning techniques based on aggregated information (Li et al., 2020b). Both datasets provide non-personally identifiable voter-level information. The following sections offer detailed descriptions of these datasets, explain their role in the evaluation framework, and outline the experimental settings used for testing the pipelines.

3.1.1 Real-world Data by American National Election Studies (ANES)

For evaluation, we use pre-election data from the ANES 2016 and 2020 Time Series Studies (Argyle et al., 2022; Studies, 2022), which provide 4,270 and 8,280 real-world samples, respectively, from individuals who participated in the 2016 and 2020 elections. The dataset includes a wide range of variables: (1) racial/ethnic self-identification, (2) gender, (3) age, (4) ideological self-placement on a conservative-liberal scale, (5) party identification, (6) political interest, (7) church attendance, (8) frequency of discussing politics with family and friends, (9) patriotic feelings associated with the American flag (unavailable in 2020), and (10) state of residence (unavailable in 2020). Additionally, the dataset records how individuals voted in both the 2016 and 2020 elections. Previous studies, such as Argyle et al. (2023), have evaluated GPT-3 using this dataset. We apply our method directly to this established benchmark to assess its effectiveness and performance.

3.1.2 Synthetic Data for the US Population

In addition to the medium-sized benchmark dataset, we utilize synthetic demographic data derived from a 1:1 synthetic population dataset of the United States (Li et al., 2020b). Synthetic data plays a crucial role in social and applied sciences, with recent applications in water quality estimation (Chia et al., 2023), financial modeling (Potluru et al., 2023a), tourist profiling (Merinov et al., 2023a), and measuring the social impact of engineered products (Stevenson et al., 2023). High-quality synthetic datasets provide researchers with large-scale data at a lower cost while maintaining privacy, making them a reliable resource.

For our purposes, the synthetic data enables the creation of a cost-effective, large-scale virtual panel of respondents that is both “wide" (each respondent has over 50k modeled features) and “long" (enough samples to reflect a national dataset). However, running LLM inference on the entire U.S. population would be prohibitively expensive, so we employ a sampling strategy. Given the pivotal role of swing states in determining election outcomes, we focus on simulating voter behavior in these states while including representative samples from red and blue states for comparison.

Synthetic Data Generation: The synthetic data used here is generated using the SynC framework (Li et al., 2020b), which reconstructs individual-level data from aggregated sources where collecting real-world individual data is impractical due to privacy, time, or financial constraints. SynC is widely recognized and applied across multiple fields to support research and overcome data limitations. For instance, it has been used in outlier detection (Li et al., 2020a), finance (Potluru et al., 2023b), tabular data modeling (Borisov et al., 2022), healthcare (Sichani et al., 2024), and tourism (Merinov et al., 2023b), demonstrating its effectiveness and importance in various domains.

SynC leverages publicly available data, such as the 2023 American Community Survey (ACS), which provides data on 242,338 census block groups, including population statistics and response proportions for each block. Using *Data Downscaling*, SynC probabilistically recreates the 340 million residents represented in the aggregated census data. For our simulation, the synthetic population includes variables relevant to election predictions: (1) age, (2) gender, (3) ethnicity, (4) marital status, (5) household size, (6) presence of children, (7) education level, (8) occupation, (9) individual income, (10) family income, and (11) place of residence.

SynC addresses the challenge of reconstructing individual data $\{x^d_{m,1}, \dots, x^d_{m,n_m}\}$ from aggregated observations $X^d_m = \sum_{k=1}^{n_m} x^d_{m,k}/n_m$, where X^d is the d-th survey question of interest, m is the census block id and n is the number of individuals in m. A *Gaussian copula* is employed to model dependencies between survey questions. Given a $d \times d$ covariance matrix Σ of the d sruvey questions, the synthetic individuals are drawn as:

$$Z^d_m \sim N(0, \Sigma), \quad u^d_m = \Phi(Z^d_m), \quad X^d_m = F_d^{-1}(u^d_m), \tag{1}$$

where $Z^d_m \sim N(0, \Sigma)$ denotes a random seed from a multivariate normal distribution, Φ is the cumulative distribution function (CDF) of the standard normal distribution, and F_d^{-1} is the inverse CDF of the marginal distribution for feature d, which is estimated based on census block level data. To maintain alignment with aggregated data, SynC uses *marginal scaling*. For categorical variables, it applies a multinomial distribution:

$$X^d \sim \text{Multi}(1, c^d, p^d_{m,k}), \tag{2}$$

where $p^d_{m,k}$ is the probability distribution over c^d categories for question d and individual k. Marginal constraints are adjusted iteratively if discrepancies arise between sampled and target proportions.

The multi-phase SynC framework ensures that: (1) marginal distributions of individual features align with real-world expectations, (2) feature correlations are consistent with aggregated data, and (3) aggregated results match the input data. For further details on SynC's methodology and algorithms, please see the original paper (Li et al., 2020b).

Partition Design and State Categorization: The synthetic dataset evaluation will operate at the state level, where we sample synthetic individuals from each state to simulate voter behavior and aggregate their votes to compare the simulated outcomes with actual election results. Given the critical role of swing states and tipping-point states in determining election outcomes, our primary focus is on these states, which include Florida (FL), Wisconsin (WI), Michigan (MI), Nevada (NV), North Carolina (NC), Pennsylvania (PA), Georgia (GA), Texas (TX), Minnesota (MN), Arizona (AZ), and New Hampshire (NH). For broader comparison in the following evaluations, we also sample from several reliably “red states,” such as Alabama (AL), Arkansas (AR), Idaho (ID), Ohio (OH), and South Carolina (SC), as well as from “blue states,” such as California (CA), Illinois (IL), New York (NY), New Jersey (NJ), and Washington (WA). These classifications are based on the 2020 election results as described by Wikipedia (contributors, 2024).

Regarding the nationwide 2024 prediction discussed in Section 4, we will run a comprehensive

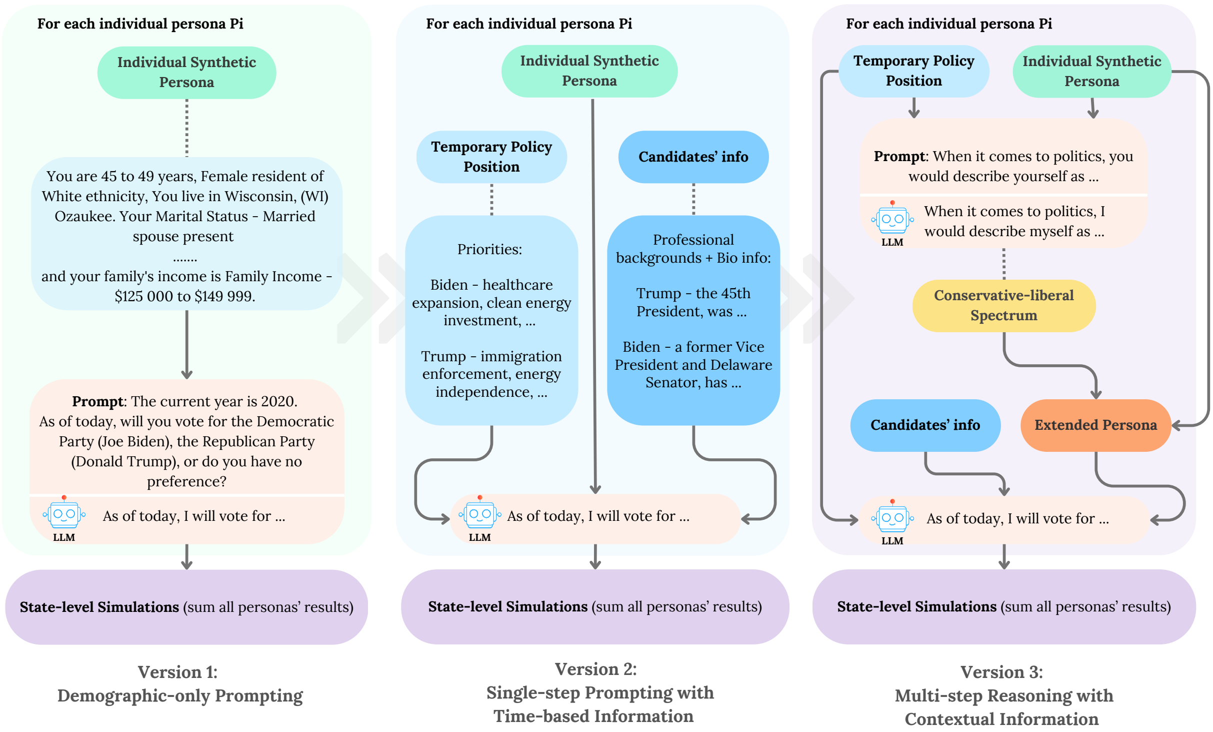

Figure 2: Progressive design of LLM pipelines for election predictions. **V1: Direct Prompt on Demographic** (§3.2.1) uses static demographic personas but lacks temporal context. **V2: Time-dependent Prompts** (§3.2.2) incorporates election-year policy shifts and candidate information, but struggles with overloaded prompts that limit prediction accuracy. **V3: Multi-step Reasoning** (§3.2.3) structures the decision-making process into sequential steps, allowing for more nuanced reasoning and yielding unbiased results that align closely with real-world outcomes. Each version aggregates individual results through state-level simulations to reflect broader election trends.

simulation across all states. This allows us to evaluate better the LLM's ability to generate results on previously unseen datasets.

Sampling Method: Running LLM inference on the entire synthetic population is computationally prohibitive, so we adopt a random sampling approach. Each state serves as a sampling unit, with sample sizes ranging between 1/100 and 1/2000 of the synthetic population, depending on the state's population size. For example, a 1/2000 sampling ratio is applied to highly populated states like California, while a 1/100 ratio is used for smaller states such as New Hampshire. This approach ensures a minimum sample size of 4269 individuals per state, corresponding to a 1.5% margin of error at a 95% confidence level, to maintain sufficient representation. Although our primary focus is on swing states due to their critical influence on election outcomes, we apply the same sampling method to red and blue states included in our simulations to ensure consistency across the analysis.

3.1.3 Evaluation Using Real-world and Synthetic Datasets

We employ two evaluation methods to assess our proposed approaches. First, for the ANES 2016 and 2020 benchmarks (Studies, 2019, 2022), we follow the methodology of Argyle et al. (2023). We compare the average voting probabilities:

$$\text{Probability} = \frac{\text{Republican Votes}}{\text{Republican Votes} + \text{Democratic Votes}} \quad (3)$$

calculated across the entire sample. Accuracy is assessed by comparing the predicted winning party with the actual election outcome.

Second, for the synthetic dataset, we treat each state as an independent validation unit. We compare the predicted results—both in terms of the winning candidate and vote share percentages—against the actual 2020 election results for each state. Accuracy is evaluated based on: (1) the agreement between the predicted and actual winning candidate for each state, and (2) the aggregate performance across all states, ensuring that the model reflects overall election trends. This state-level evaluation uses voter-level information processed through LLMs to predict outcomes accurately.

3.1.4 Hardware and LLM Settings

Our experiments are deployed on a GPU server equipped with an AMD EPYC Milan 7763 processor, 1 TB (64x16 GB) DDR4 memory, 15 TB SSD storage, and 6 NVIDIA RTX A6000 Ada GPUs. For the LLM component, we primarily utilize OpenAI's GPT-4o model for election predictions. Additionally, Meta's LLaMA 3.1 405B model is used in intermediate steps to provide neutral summarization of time-dependent information, enhancing certain pipelines by capturing temporal dynamics more effectively.

3.2 Our Progressive Design of LLM Pipelines

In this section, we present our progressive design for making voter-level election predictions using LLMs. As illustrated in Fig. 2, the methodology evolves through three distinct versions, each addressing limitations of the previous approach and incorporating more advanced techniques.

V1: Demographic-only Prompting (§3.2.1): This version uses static demographic personas to prompt the LLM for voter-level predictions. While straightforward, it cannot account for temporal shifts in candidates' political focus over time, limiting its predictive power.

V2: Single-step Prompting with Time-based Information (§3.2.2): To address temporal factors, this version enriches the prompts with election-year-specific information, such as candidates' policy positions and campaign focuses. However, packing all relevant info. into a single prompt creates cognitive overload, which can hinder effective reasoning and reduce prediction accuracy.

V3: Multi-step Reasoning with Contextual Information (§3.2.3): This version breaks down the prediction process into sequential steps to improve reasoning. Structuring the decision-making process allows the model to effectively incorporate voter information, candidates' profiles, and political context. Our experiments across all datasets show that this approach produces improved predictions closely aligned with real-world outcomes.

The subsequent sections provide detailed descriptions of each version and its development. The quantitative evaluation of the three pipelines is presented in §3.3.

3.2.1 Version 1: Demographic-only Prompting

Building on prior research demonstrating LLMs' ability to simulate human behavior (Xie et al., 2024), this initial version directly prompts the LLM with a persona and asks how that persona would vote (Argyle et al., 2023). This method provides all relevant information simultaneously, making it the simplest approach for voter-level prediction.

> **Task:** You are persona [age, gender, ethnicity, marital status, household size, presence of children, education level, occupation, individual income, family income, and place of residence.] The current year is [year].
>
> Please answer the following question as if you were the resident:
>
> 1. As of today, will you vote for the Democratic Party (Joe Biden), the Republican Party (Donald Trump), or do you have no preference?
>   - Democratic
>   - Republican
>   - No Preference

In our prompt, we explicitly specify the year as 2020 to avoid confusion, since the LLM used—GPT-4o—has knowledge only up to 2023. Without this clarification, the model might assume the present year is 2023, impacting its predictions. The structure of the voting options follows the style used in Pew Research Center's 2014 Political Polarization and Typology Survey (Pew Research Center, 2014).

Limitations: While simple and intuitive, this approach is limited by its inability to account for temporal changes in candidates' political agendas and public opinion. As a result, predictions for different years (e.g., 2020 vs. 2024) may not reflect meaningful variation, reducing the method's effectiveness in dynamic election contexts.

3.2.2 Version 2: Single-step Prompting with Time-based Information

Capturing macro-level and time-dependent variables is essential for bottom-up agent-based modeling in election prediction (Gao et al., 2022b). To enhance the contextual relevance of our simulations, we extended our pipeline by integrating election-year data sourced from Ballotpedia, a well-regarded political information platform. It includes campaign agendas, key policy stances, and candidates' biographical and professional backgrounds.

Delivering this time-based information neutrally to LLMs is crucial to avoid skewed predictions. Given the documented political bias in LLMs (Feng et al., 2023), we experimented with both GPT-4o and LLaMA3-405B to summarize the information neutrally. Our preliminary findings indicate that LLaMA3-405B offers more balanced expressions These unbiased summaries were integrated into the prompts as follows:

> **Task:** You are persona [demographics]. The current year is [year]. [Two parties' policy agenda]. [Presidential candidates' biographical and professional backgrounds].
>
> Please answer the following question as if you were the resident:
>
> 1. As of today, will you vote for the Democratic Party (Joe Biden), the Republican Party (Donald Trump), or do you have no preference?
>   - Democratic
>   - Republican
>   - No Preference

Limitations: While this version creates more dynamic and contextually grounded simulations, it introduces a skew in the predictions for certain states. Specifically, when tested across five "deep red" states, five "deep blue" states, and all 11 swing and tipping-point states, we observed a pronounced skewness towards the Democratic Party, even in historically red states such as Alabama and South Carolina, as well as swing states like Texas and Florida. This aligns with prior research suggesting that GPT-4o tends towards liberal ideologies (Feng et al., 2023). However, this skewness was less pronounced when tested on the ANES 2020 dataset (Studies, 2022). In that case, the predicted share for Trump was 46.7%, slightly higher than the ground truth of 41.2%. Further analysis revealed that the ANES dataset contains more features than standard demographic datasets, including ideological self-placement along the conservative-liberal spectrum.

Through additional experimentation, we found that removing the ideological self-placement feature from the ANES data caused the predictions to shift significantly toward demographic factors. This suggests that ideological self-placement is critical in mitigating political skew, showing its importance as a corrective feature in election prediction.

3.2.3 Version 3: Multi-step Reasoning with Contextual Information

To overcome the limitations of Version 2 and leverage insights from the ANES dataset analysis, we developed a multi-step prompting pipeline. Inspired by the Chain of Thought prompting strategy (Wei et al., 2022), this approach divides the task into intermediate steps, allowing the model to process information more systematically and accurately.

The process consists of two main steps: (**1**) Conservative-Liberal Spectrum Placement: First, the LLM is provided with a specific persona along with the current policy positions of both parties. The model is then asked to place the persona on the conservative-liberal spectrum based on the provided information. (**2**). Extended Persona and Voting Simulation: The conservative-liberal spectrum placement is incorporated into the persona to create an extended persona. This extended persona, along with the time-based information, is used in the second step to simulate voting behavior. The overall prompts are structured as follows:

> **Step 1:** You are a persona with [demographics]. The current year is [year]. [Two parties' policy agenda]. When it comes to politics, would you describe yourself as:
>
> | | |
> |---|---|
> | No answer | Very liberal |
> | Somewhat liberal | Closer to liberal |
> | Moderate | Closer to conservative |
> | Somewhat conservative | Very conservative |
>
> **Step 2:** You are a persona with [demographics]. Your [conservative-liberal spectrum]. The current year is [year]. [Two parties' policy agenda]. [Presidential candidates' biographical and professional backgrounds].
> Please answer the following question as if you were the resident:
>
> 1. As of today, will you vote for the Democratic Party (Joe Biden), the Republican Party (Donald Trump), or do you have no preference?
>   - Democratic
>   - Republican
>   - No Preference

We applied this multi-step method to simulate the 2020 election at the state level, covering various types of states (e.g., "deep red", "deep blue", swing, and tipping-point states). This approach resulted in significantly improved performance, with outcomes closely aligning with real-world scenarios. Specifically, the voting distributions reflected expected patterns in deep red and deep blue states such as Alabama and California, with vote counts closely matching ground-truth distributions. For swing states, the results were also accurate, except for Arizona and North Carolina, where the predicted outcomes flipped. Nonetheless, in all other swing states, the simulated vote distributions

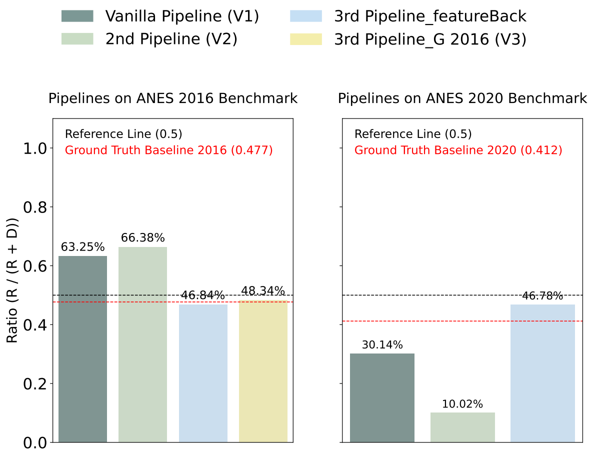

Figure 3: Comparison of the three pipelines on ANES 2016 and 2020 benchmarks. The y-axis shows the predicted Republican vote ratio (R / (R + D)), with 0.5 indicating a balanced outcome. V1 (Vanilla Pipeline) and V2 (Single-step Time-based Prompting) overestimate Republican support, particularly in 2016. V3 (Multi-step Reasoning) achieves the most accurate results, closely matching the ground truth ratios: 48.34% vs. 47.7% (2016) and 46.78% vs. 41.2% (2020). These results highlight the improved accuracy of V3.

closely mirrored the real-world outcomes, capturing these states' balanced and marginal dynamics.

Based on this approach's improved accuracy and alignment with real-world results, we choose Version 3 as our **final pipeline** for election prediction. The structured, multi-step reasoning process not only mitigates skewness but also effectively captures the complex dynamics of voter behavior across different states.

3.3 Validation on the Proposed Pipelines

3.3.1 Evaluations on Real-world Data (ANES)

Settings. We evaluate our pipelines using the public ANES 2016 and 2020 Time Series datasets (Studies, 2019) (Studies, 2022) to: (1) assess the overall performance of each pipeline and (2) validate the V3 pipeline's ability to generate the critical Conservative-Liberal Spectrum feature.

For the first two pipelines (V1: Demographic-only Prompting and V2: Single-step Prompting with Time-based Information), we manually excluded the Conservative-Liberal Spectrum feature during evaluation to simulate how these pipelines would perform based purely on demographic data. This step mimics the limitations of simpler prompts that lack deeper ideological alignment.

The V3 pipeline, as outlined in §3.2.3, addresses the limitations of earlier versions by using a multi-step reasoning approach inspired by Chain of Thought prompting (Wei et al., 2022). This design involves two primary steps. First, the LLM places a persona on the Conservative-Liberal Spectrum based on the persona's demographics and the two parties' policy positions. Second, this placement is incorporated into an extended persona, which, along with policy agendas and candidates' biographical information, is used to simulate voting behavior. The evaluation of the V3 pipeline involved slightly different strategies for the two datasets. The 2020 ANES dataset lacks key demographic features such as 'state of residence' and 'patriotic feelings associated with the American flag,' making it more challenging for the LLM to generate the Conservative-Liberal Spectrum. To address this, we restored the original spectrum in the 2020 dataset for evaluation. For the 2016 dataset, which contains all relevant features, we tested two variations of the V3 pipeline. In the first test, labeled *3rd Pipeline_G 2016*, the LLM generated the Conservative-Liberal Spectrum using the available demographic information. In the second test, labeled *3rd Pipeline_featureBack*, we restored the original spectrum to simulate a comparison. The results for these evaluations are shown in Figure 3.

Validation Results. Our results reveal a consistent trend: when the Conservative-Liberal Spectrum is removed, the predictions shift toward the winning party of the respective election year. For instance, in 2016, the LLM predicted 63.25% support for Trump, favoring the Republican Party. In 2020, the prediction shifted to 30.14% support for Trump, favoring the Democratic Party. This shift becomes even more pronounced in the V2 pipeline, where additional time-based information introduces further skew in the predictions.

Despite the missing features in the 2020 dataset, the V3 pipeline still demonstrated noticeable improvements over the simpler V1 and V2 pipelines. On the more feature-complete 2016 dataset, the V3 pipeline performed well, achieving 46.84% support for Trump when using the original feature and 48.38% with the generated feature—both closely aligning with the ground truth baseline of 47.7%. Notably, the generated spectrum in the *3rd Pipeline_G 2016* version produced results even closer to the ground truth than the restored spectrum. The performance improvements observed underscore the importance of incorporating ideological alignment in voter simulations.

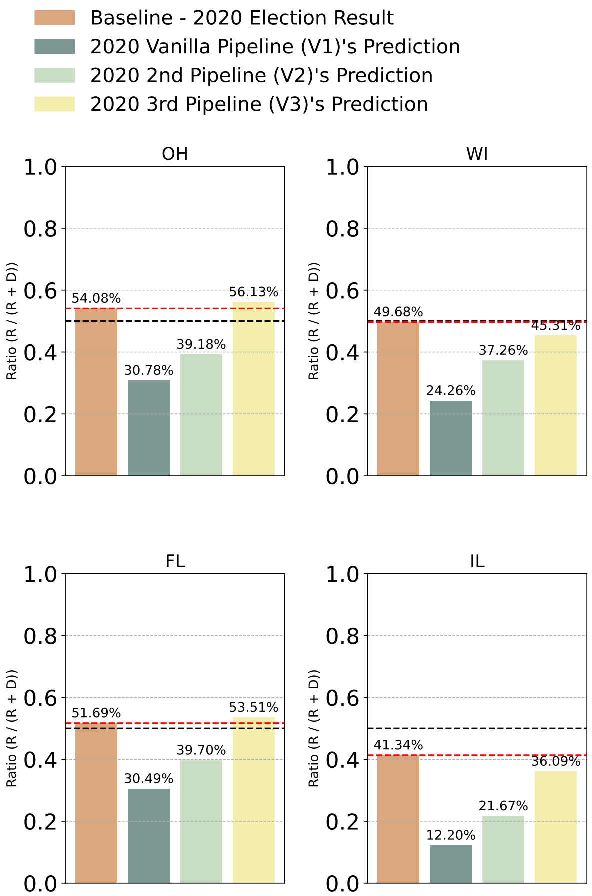

Figure 4: LLM's predictions for four states in the 2020 election compared with Ground Truth results. The figure presents results for one red state (Ohio, OH), one blue state (Illinois, IL), one swing state (Wisconsin, WI), and one tipping-point state (Florida, FL). V1 and V2 pipelines tend to underestimate Republican support, while V3 (Multi-step Reasoning) provides the closest alignment with actual outcomes, especially in swing and tipping-point states.

3.3.2 Evaluations on Synthetic Data for the 2020 US Population

In addition to the nationwide evaluation on the ANES datasets, we conducted state-level simulations using synthetic data to compare predictions with actual 2020 election outcomes. For each state, we performed random sampling based on population size to ensure a statistically meaningful number of personas. The simulation outcomes were then benchmarked against official 2020 Presidential General Election Results from the Federal Election Commission (FEC). As in the benchmark evaluations, we calculated the average voting probabilities to assess the alignment of predictions with real-world outcomes. We evaluated five red states, five blue states, and 11 swing and tipping-point states. Figure 4 highlights representative results from these categories, providing insights into the model's performance in different electoral contexts.

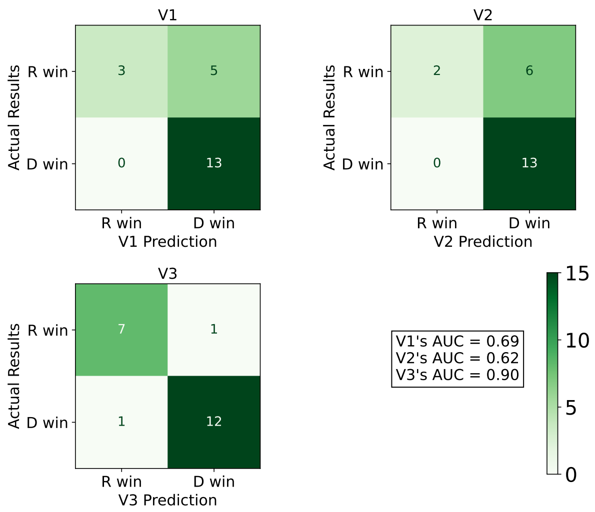

Figure 5: Aggregated results of the three pipelines (V1, V2, V3) on state-level simulations. Each confusion matrix presents the number of states where predictions align with or deviate from actual outcomes. V1 (AUC = 0.69) and V2 (AUC = 0.62) show lower accuracy, while V3 (AUC = 0.90) performs best, effectively capturing Republican victories without compromising Democratic predictions. It is worth noting that, so far, we have only tested the pipelines in 21 states. If the scope is expanded to include all states, the AUC of V3 is expected to improve further, while the AUC of V1 and V2 are expected to decline.

Consistent with the ANES dataset evaluations, the V1 pipeline (Demographic-only Prompt) exhibited a skew toward the Democratic Party, even in traditionally Republican-leaning states like South Carolina (SC), Alabama (AL), and Ohio (OH), with predictions diverging significantly from actual results. This illustrates the limitations of using demographic data alone without time-sensitive context. The V2 pipeline (Time-dependent Prompt) introduced election-year-specific information, which partially reduced the skew in the state-level simulations. However, the model still struggled to eliminate prediction biases, particularly in polarized states. Interestingly, this differed from the ANES evaluations, where including time-dependent information amplified the bias. The V3 pipeline (Multi-step Reasoning) demonstrated the most accurate performance, effectively mitigating skewness across deep red and blue states. In these polarized states, the predictions closely mirrored the actual voting outcomes, reflecting the model's improved ability to incorporate ideological alignment through multi-step reasoning.

For swing and tipping-point states, the V3 pipeline achieved robust results, correctly predicting the outcomes in 9 out of 11 states. Minor deviations were observed in North Carolina (NC) and Arizona (AZ), where the predictions were slightly misaligned with the real results. Nonetheless, the V3 pipeline provided balanced predictions that accurately captured the competitive dynamics typical of swing states, further validating its effectiveness.

In summary, the comparative performance of the three pipelines across different state categories is shown in Figure 4. The V3 pipeline consistently outperformed the other two, delivering more stable and accurate predictions. Aggregate results for all pipelines on all 21 chosen states is shown in the below figure 5. And all state-level simulation results can be found in the appendix.

4 Prediction for the 2024 US Election

4.1 Experiment Settings

We extend our multi-step reasoning pipeline to predict the outcome of the 2024 US Presidential Election, featuring Donald Trump and Kamala Harris as the primary candidates, which thus introduce differences from the 2020 results. Same as in the 2020 simulations, we incorporate only the candidates' policy positions and backgrounds as time-sensitive inputs, excluding real-time data from social media or news sources to maintain consistency and control within the model's input space. We follow the same simulation setup used in §3.3, utilizing identical hardware and LLM APIs. Since GPT-4o's training corpus extends only until October 2023, this experiment evaluates the model's ability to make predictions on future, unseen data. After the election, these predictions can be compared with the actual results, offering a unique opportunity to assess the LLM's predictive capabilities.

Consistent with our prior methods, we simulate individual votes cast exclusively for the Democratic and Republican parties, omitting No-Preference votes from the analysis. The preference for Trump (Republican Party) in each state is calculated using Eq. (3). We applied the V3 pipeline to all 50 states. As in previous experiments, we employed random sampling with proportional extraction of synthetic personas for each state, ensuring statistical significance while optimizing computational efficiency. This state-level analysis tests the adaptability of our multi-step pipeline to unseen election data and provides a scalable framework for forecasting election outcomes with synthetic personas.

4.2 Prediction Results

The LLM's 2024 forecast reveals notable shifts from the 2020 election. A key change is in Wisconsin (WI), where Trump is now projected to win with 54.90% of the vote, marking a significant shift. Trump also shows gains in other swing states, including Pennsylvania (PA) with 47.85%, Michigan (MI) with 48.87%, and New Hampshire (NH) with 46.39%, though he is still forecasted to narrowly lose these states to Harris.

In other battlegrounds, Trump is predicted to carry Arizona (AZ) with 51.09%, flipping it back to the Republicans, but lose North Carolina (NC) with 45.69%—consistent with the 2020 forecast, despite the incorrect outcome in that election. As shown in Figure 6, Trump is expected to secure Florida (FL) with 53.62% and Texas (TX) with 56.36%, further strengthening his position. However, Harris holds an edge in Nevada (NV) with 34.77%, Georgia (GA) with 44.36%, and Minnesota (MN) with 44.23%. These results indicate highly competitive dynamics in key swing states, where narrow margins could determine the final outcome.

In other traditional red and blue states, Trump retains strong support in red states such as Arkansas (AR) with 69.69% and Alabama (AL) with 67.20%. Meanwhile, Harris is projected to maintain dominance in Democratic strongholds like California (CA) with 19.18%, New York (NY) with 22.41%, and Illinois (IL) with 30.36%. Notably, in Alaska (AK)—a state that has remained red for the past 52 years—the LLM predicts a narrow Democratic victory with 49.39% of the vote in 2024. It is worth mentioning that our preliminary test of the LLM on the 2020 Alaska election yielded an accurate Republican win with 52.42%, making the predicted 2024 shift even more intriguing, as it could subtly tip the scales toward the Democrats.

Based on the LLM's predictions and the winner-takes-all electoral vote allocation, Harris (Democratic Party) is projected to narrowly defeat Trump (Republican Party) with 270 (267 + 3 in Washington, D.C.) electoral votes to Trump's 268. Trump's strength in major Republican states like Texas and Florida underscores his resilience, while Harris's performance in strongholds such as California and New York reaffirms Democratic dominance. The tight margins in swing states like Pennsylvania, Georgia, and Arizona emphasize how slight shifts in voter preferences can still alter the outcome, demonstrating the fine balance between the two.

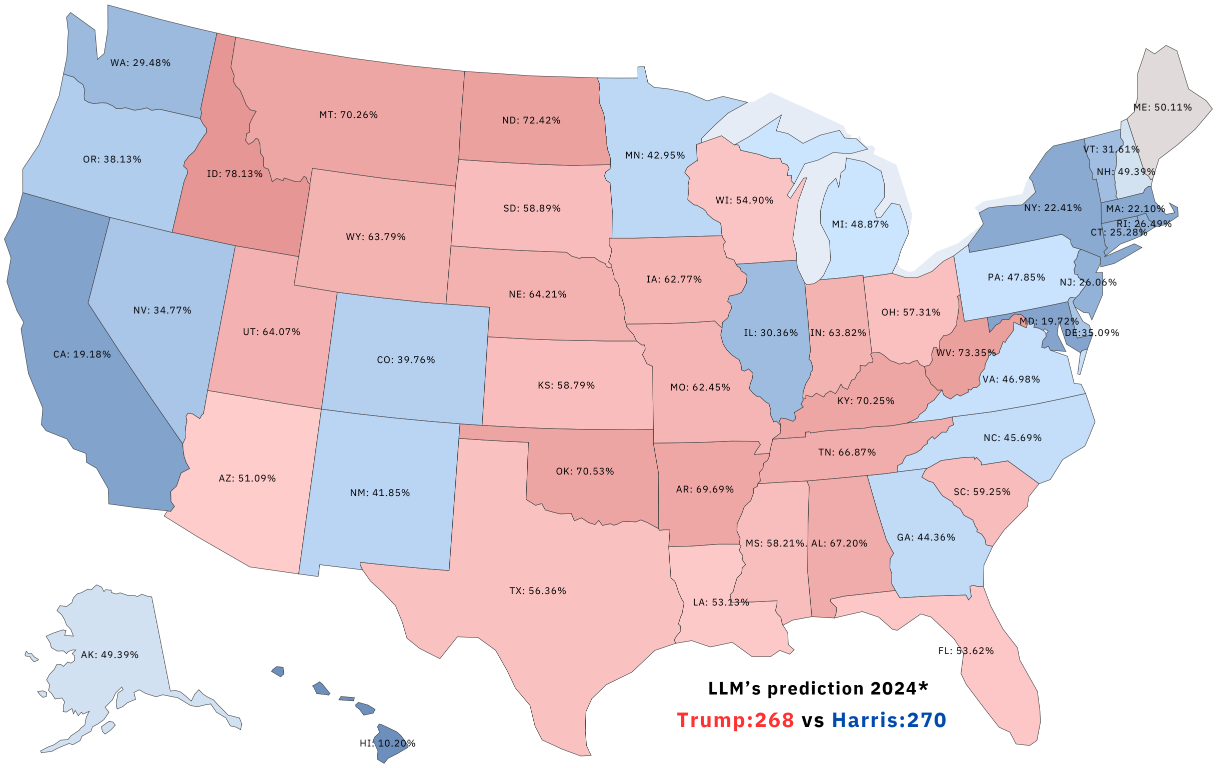

Figure 6: LLM's prediction for the 2024 election. The predicted probability for each state is calculated as $P = \frac{R}{R+D}$, where R and D represent the total votes for the Republican and Democratic candidates, respectively. States shaded in red (■) indicate a predicted Republican win, while states shaded in blue (■) indicate a predicted Democratic win. For Maine (ME) and Nebraska (NE), we used the winner-takes-all method instead of the congressional district method to calculate electoral votes, ensuring consistency across states. Based on the LLM's predictions, Kamala Harris (Democratic) is projected to secure a narrow victory over Donald Trump (Republican), with a final electoral vote tally of Harris: 270 vs. Trump: 268. Trump's predicted wins in key states like Texas, Florida, and Ohio demonstrate his strength in traditionally Republican regions. Meanwhile, Harris is expected to perform well in Democratic strongholds such as California, New York, and Illinois. The results emphasize the competitive dynamics in swing states like Pennsylvania, Georgia, and Arizona, where slight shifts in voter behavior can alter the outcome.

Limitations. The 2024 prediction pipeline relies on synthetic persona demographics, party policy positions, and candidate biographical information. While this controlled framework ensures consistency, it introduces certain limitations. *First*, the model does not account for shifts in public opinion, media narratives, or unexpected events—such as economic shocks or political scandals—that could alter voter behavior. *Second*, the simulation assumes static demographics from 2020 to 2024, neglecting any population shifts or migration patterns, even if small, that could affect state-level outcomes.

Note that *this paper's objective is not to provide definitive forecasts but rather to showcase the potential of LLMs to generalize on unseen data*. This experimental setup highlights how well these models align with real-world patterns, serving as a demonstration of their predictive capacity rather than a source of fully reliable election predictions.

5 Conclusion and Future Directions

In this work, we present a novel framework for election prediction using large language models (LLMs) with a focus on multi-step reasoning. By leveraging both synthetic personas and real-world datasets, we demonstrated the potential of LLMs to capture individual voting behaviors and state-level election outcomes. Our iterative design highlights the importance of integrating temporal information and complex reasoning for accurate predictions. The 2020 and 2024 simulations reveal both the strengths and limitations of using LLMs in dynamic political environments, emphasizing the model's ability to generalize on unseen data while showing the challenges associated with static demographic assumptions and limited real-time data inputs.

Future research can extend this work by incorporating multiple LLMs to better understand their internal political tendencies, enhancing temporal modeling with public opinion data and real-time trends for improved accuracy, and developing stronger multi-step reasoning pipelines through refined Chain of Thought (CoT) designs to further enhance prediction performance and mitigate biases. This study lays the foundation for future applications of LLMs in political forecasting, offering promising directions for further development in both election prediction and the broader study of LLM behavior in social contexts.

References

Badr AlKhamissi, Millicent Li, Asli Celikyilmaz, Mona Diab, and Marjan Ghazvininejad. 2022. A review on language models as knowledge bases. *arXiv preprint arXiv:2204.06031*.

Emma Anderson and Michael Taylor. 2023. Navigating the nuances: Challenges of political language for llms. *Computational Linguistics in Politics*, 7(3):412–435.

Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. 2022. Replication Data for: "Out of One, Many: Using Language Models to Simulate Human Samples".

Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351.

James Bisbee, Joshua D. Clinton, Cassy Dorff, Brenton Kenkel, and Jennifer M. Larson. 2024. Synthetic replacements for human survey data? the perils of large language models. *Political Analysis*, page 1–16.

Duncan Black. 1948. On the rationale of group decision-making. *Journal of Political Economy*, 56(1):23–34.

Rishi Bommasani et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Vadim Borisov, Thomas Leemann, Katharina Seßler, Jonas Haug, Martin Pawelczyk, and Gjergji Kasneci. 2022. Deep neural networks and tabular data: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 34(4):1686–1711.

Robert Brown. 2023. Challenges in applying llms to complex political systems. *Computational Politics Review*, 8(3):301–320.

Tom Brown et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Ilias Chalkidis et al. 2022. Lexglue: A benchmark dataset for legal language understanding in english. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 4310–4330.

Wei Chen and Maria Rodriguez. 2024. Decoding political rhetoric: An llm-based approach. *AI in Governance*, 5(1):45–67.

Min Yan Chia, Chai Hoon Koo, Yuk Feng Huang, Wei Di Chan, and Jia Yin Pang. 2023. Artificial intelligence generated synthetic datasets as the remedy for data scarcity in water quality index estimation. *Water Resources Management*, 37(15):6183–6198.

Megan Collins and Juan Martinez. 2023. Media influence and peer effects in voter decision-making: An agent-based approach. *Political Communication Quarterly*, 31(2):178–201.

Wikipedia contributors. 2024. Swing state. Accessed: 2024-10-12.

Anthony Downs. 1957. *An Economic Theory of Democracy*. Harper & Row.

Robert S Erikson and Christopher Wlezien. 2016. Forecasting us presidential elections using economic and noneconomic fundamentals. *PS: Political Science & Politics*, 49(4):669–672.

Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. 2023. From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair nlp models. *Preprint*, arXiv:2305.08283.

Lin Gao, Wei Zhang, and Jian Liu. 2022a. Agent-based modeling for election prediction: A network approach. *Computational Social Networks*, 9(1):1–22.

Ming Gao, Zhongyuan Wang, Kai Wang, Chenhui Liu, and Shiping Tang. 2022b. Forecasting elections with agent-based modeling: Two live experiments. *PLOS ONE*, 17(6):1–11.

Andreas Graefe. 2014. Accuracy of vote expectation surveys in forecasting elections. *Public Opinion Quarterly*, 78(1):204–232.

Emaad Haq et al. 2023. Large language models for political science research. *Political Analysis*, pages 1–21.

Thomas M Holbrook. 2016. *Forecasting US presidential elections*. Rowman & Littlefield.

Tiancheng Hu and Nigel Collier. 2024. Quantifying the persona effect in LLM simulations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10289–10307, Bangkok, Thailand. Association for Computational Linguistics.

Emily Johnson and David Lee. 2024. Political sentiment analysis using large language models. In *Proceedings of the International Conference on Computational Social Science*, pages 78–92. ICCSS.

Soo-Jin Kim and Rahul Patel. 2024. Large language models for policy sentiment analysis: A case study. In *Proceedings of the Conference on AI in Public Policy*, pages 201–215. AIPP.

Carlos Lemos, Helder Coelho, and Rui J Lopes. 2019. An agent-based model of voter turnout and preference formation. In *Proceedings of the Social Simulation Conference*, pages 245–256. SSC.

Thomas Lerer et al. 2022. Political sentiment analysis: A comparison of lexicon-based and machine learning approaches. *Social Science Computer Review*, 40(5):1135–1154.

Michael S Lewis-Beck and Tom W Rice. 1992. *Forecasting Elections*. Georgetown University Press.

Zheng Li, Yue Zhao, Nicola Botta, Cezar Ionescu, and Xiyang Hu. 2020a. Copod: copula-based outlier detection. In *2020 IEEE International Conference on Data Mining (ICDM)*, pages 1118–1123. IEEE.

Zheng Li, Yue Zhao, and Jialin Fu. 2020b. Sync: A copula based framework for generating synthetic data from aggregated sources. In *2020 International Conference on Data Mining Workshops (ICDMW)*, pages 571–578. IEEE.

Carmen Lopez and Rajesh Singh. 2023. Bridging the gap: Integrating macro-level language models with micro-level voter behavior. *Journal of Computational Political Science*, 6(4):523–547.

Pavel Merinov, David Massimo, and Francesco Ricci. 2023a. Behaviour-aware tourist profiles data generation. In *IIR*, pages 3–8.

Pavel Merinov, Davide Massimo, and Francesco Ricci. 2023b. Behaviour-aware tourist profiles data generation. In *Proceedings of the 13th Italian Information Retrieval Workshop*.

Pew Research Center. 2014. Political polarization in the american public. https://www.pewresearch.org.

Vamsi K Potluru, Daniel Borrajo, Andrea Coletta, Niccolò Dalmasso, Yousef El-Laham, Elizabeth Fons, Mohsen Ghassemi, Sriram Gopalakrishnan, Vikesh Gosai, Eleonora Kreačić, et al. 2023a. Synthetic data applications in finance. *arXiv preprint arXiv:2401.00081*.

Vijay Kumar Potluru, Diego Borrajo, Andrea Coletta, Niccolò Dalmasso, Sumanta Das, Shashi Gupta, Scott Harmon, Nimish Kakkar, Yue Meng, Prabhakar Natarajan, et al. 2023b. Synthetic data applications in finance. *arXiv preprint arXiv:2301.07827*.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? *arXiv preprint arXiv:2002.08910*.

Jennifer Roberts, Thomas Brown, and Elena Garcia. 2023. Political campus: A benchmark dataset for llm-based election prediction. *Data in Politics*, 4(2):89–112.

Elham Karimi Sichani, Alexandra Smith, Khaled El Emam, and Anna Goldenberg. 2024. Creating high-quality synthetic health data: Framework for model development and validation. *JMIR Formative Research*, 8:e50704.

John Smith and Jane Doe. 2023. Llms in political forecasting: A new frontier. *Journal of Political Technology*, 15(2):123–145.

Phillip D Stevenson, Christopher A Mattson, Eric C Dahlin, and John L Salmon. 2023. Creating predictive social impact models of engineered products using synthetic populations. *Research in Engineering Design*, 34(4):461–476.

American National Election Studies. 2019. Anes 2016 time series study full release. https://www.electionstudies.org. [Dataset and documentation]. September 4, 2019 version.

American National Election Studies. 2022. Anes 2020 time series study full release. https://www.electionstudies.org. [Dataset and documentation]. February 10, 2022 version.

Alice Thompson. 2023. Ethical considerations in ai-driven political analysis. *Journal of AI and Society*, 12(4):567–589.

Jason Wei et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Oliver Williams, Sophia Johnson, and Kevin Lee. 2024. Challenges and opportunities in applying llms to political discourse analysis. In *Proceedings of the Workshop on NLP for Political Science*, pages 56–70. ACL.

Sarah Wilson, James Taylor, and Raj Patel. 2024. Robust political language models: Mitigating bias and misinterpretation. In *Proceedings of the Conference on AI Ethics in Politics*, pages 112–128. AIEP.

Chengxing Xie, Canyu Chen, Feiran Jia, Ziyu Ye, Kai Shu, Adel Bibi, Ziniu Hu, Philip Torr, Bernard Ghanem, and Guohao Li. 2024. Can large language model agents simulate human trust behaviors? *Preprint*, arXiv:2402.04559.

Yilang Xu. 2022. Deep learning in political science: Introducing rdlmatio to estimate the ideological positions of political actors using roll-call data. *Political Analysis*, 30(4):458–467.

Yi-Hsuan Yang et al. 2022. A survey on deep learning for text-to-music generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:1808–1837.

Karan Zhang et al. 2023. Large language models encode clinical knowledge. *Nature*, 618(7965):732–739.

Yun Zhang, Li Wang, and Anil Kumar. 2024. A hybrid framework for election forecasting: Integrating llms and abms. In *Proceedings of the International Conference on Computational Social Science*, pages 301–315. ICCSS.

Xuhui Zhou, Yue Zhang, Leyang Cui, and Dandan Huang. 2020. Evaluating commonsense in pretrained language models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 9733–9740.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2023. Large language models are human-level prompt engineers. In *The Eleventh International Conference on Learning Representations*.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can large language models transform computational social science? *Computational Linguistics*, 50(1):237–291.