Agentmandering: A Game-Theoretic Framework for Fair Redistricting via Large Language Model Agents

Hao Li[1]*, Haotian Chen[2]*, Ruoyuan Gong[1], Juanjuan Wang[3], Hao Jiang[1]†,

[1]Wuhan University, Wuhan, 430072 China
[2]University of California, Los Angeles, Los Angeles, 90095
[3]Zhongnan University of Economics and Law, Wuhan, 430073 China
whulh@whu.edu.cn, barneychen@ucla.edu, GongRuoyuan@whu.edu.cn, Wangjj@zuel.edu.cn, jh@whu.edu.cn

Abstract

Redistricting plays a central role in shaping how votes are translated into political power. While existing computational methods primarily aim to generate large ensembles of legally valid districting plans, they often neglect the strategic dynamics involved in the selection process. This oversight creates opportunities for partisan actors to cherry-pick maps that, while technically compliant, are politically advantageous. Simply satisfying formal constraints does not ensure fairness when the selection process itself can be manipulated. We propose **Agentmandering**, a framework that reimagines redistricting as a turn-based negotiation between two agents representing opposing political interests. Drawing inspiration from game-theoretic ideas, particularly the *Choose-and-Freeze* protocol, our method embeds strategic interaction into the redistricting process via large language model (LLM) agents. Agents alternate between selecting and freezing districts from a small set of candidate maps, gradually partitioning the state through constrained and interpretable choices. Evaluation on post-2020 U.S. Census data across all states shows that Agentmandering significantly reduces partisan bias and unfairness, while achieving 2 to 3 orders of magnitude lower variance than standard baselines. These results demonstrate both fairness and stability, especially in swing-state scenarios.

Code — https://github.com/Lihaogx/AgentMandering

Introduction

In representative democracies, electoral districts determine how citizens are grouped for political representation. In the United States, the winner-takes-all and single member system makes electoral results highly sensitive to district boundaries, which significantly affect the results of congressional and state legislative races (Cox and Katz 2002). Redistricting, the periodic redrawing of district boundaries to reflect population changes, is essential for equitable representation. However, this process is frequently manipulated for political advantage through a practice known as partisan *gerrymandering*, where district lines are intentionally designed to favor the party that controls the drawing (Gelman and King 1994). Common tactics include *packing* voters into a small number of districts to concentrate their influence, or *cracking* them across many districts to dilute their voting power. As illustrated in Figure 1a, a blue-majority population can be divided such that the red-minority secures more districts, demonstrating a classic gerrymandering situation that undermines fair representation.

Modern computational approaches of redistricting focus on generating large ensembles of districting plans that comply with formal legal and demographic constraints. Methods such as Markov Chain Monte Carlo (MCMC) (DeFord, Duchin, and Solomon 2021; Chikina, Frieze, and Pegden 2017; Carter et al. 2019), Sequential Monte Carlo (SMC) (McCartan and Imai 2020), and integer programming (Fravel et al. 2023) are commonly used to produce thousands of plausible alternatives, enabling statistical comparisons to identify instances of extreme partisan bias. However, as shown in Figure 1b, the heatmap reveals that the four key evaluation metrics exhibit low correlations, suggesting that these metrics capture orthogonal dimensions of fairness. This independence creates opportunities for partisan manipulation: as shown in the example, two plans with nearly identical metric scores can lead to starkly different electoral outcomes (Chambers, Miller, and Sobel 2017; Barnes and Solomon 2021). In practice, the abundance of legally compliant maps can be exploited by political actors who select technically valid plans that subtly serve partisan goals. Thus, simply generating maps that meet formal constraints is insufficient. The central challenge is **how to generate plans that are not only compliant, but also robust against strategic selection and capable of achieving fair outcomes under adversarial decision-making**.

Recent work in fair redistricting has proposed negotiation-based protocols that aim to achieve equitable outcomes through structured interaction. One prominent example is the *Choose-and-Freeze* protocol (Pegden, Procaccia, and Yu 2017), which draws from classical ideas in fair division and game theory, particularly the "cake-cutting" paradigm (Brams et al. 2006). In this game-theoretic setting, two opposing parties take turns: one selects a complete districting plan, and the other freezes a single district from it. The process then recurses on the remaining territory. This alternating structure creates a balanced strategic environment in which each side possesses both agency

*These authors contributed equally.
†Corresponding author

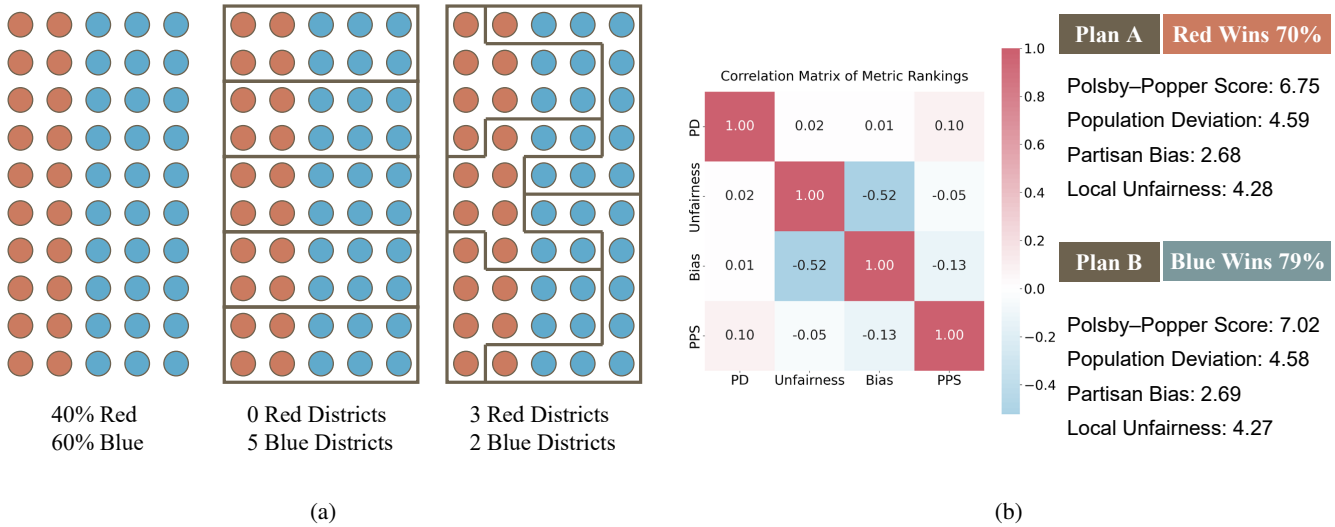

Figure 1: (a) Gerrymandering example showing district manipulation. (b) Correlation analysis of four redistricting metrics and comparison of two districting plans.

and constraint. The protocol has been proven to produce fair outcomes under reasonable assumptions, offering formal guarantees of envy-freeness and symmetry without relying on external arbiters or optimization objectives.

Despite these theoretical advantages, protocols like *Choose-and-Freeze* remain largely disconnected from computational practice. They are designed to guide human negotiation, but cannot be directly implemented using existing algorithmic redistricting pipelines. Most modern methods focus on generating large ensembles of legal plans via sampling or optimization, lacking the interactive structure and strategic balance that these protocols embody.

To bridge the gap between theoretical negotiation protocols and practical redistricting methods, we introduce **Agentmandering**, a framework that implements the game-theoretic *Choose-and-Freeze* protocol (Pegden, Procaccia, and Yu 2017) using large language model (LLM) agents (Li, Gong, and Jiang 2025). At each step, a small set of feasible districting plans is generated over the remaining unpartitioned region. One agent chooses a preferred map, and the opposing agent freezes a single district from it. The process then recurses until the full state is partitioned. By leveraging the LLMs' capacity for strategic reasoning and preference modeling, we simulate partisan decision-making within a structured, bounded-interaction protocol. This design enables the practical realization of theoretically fair procedures via AI agents, yielding districting outcomes that are both interpretable and robust to strategic manipulation.

1. We introduce a new paradigm that leverages LLM agents to implement game-theoretic protocols, bringing abstract fairness principles into practical tools for computational redistricting.
2. We introduce **Agentmandering**, a framework combining the *Choose-and-Freeze* protocol with LLM agents to structure redistricting as a strategic negotiation, constraining partisan manipulation and yielding fairer outcomes.
3. We demonstrate that Agentmandering achieves 2 to 3 orders of magnitude lower variance than existing methods on post-2020 U.S. Census data across all states, while reducing partisan bias and unfairness. This highlights its robustness across all states, particularly in swing-state scenarios.

Related Work

Evaluation and selection of districting plans. Previous research by mathematicians, computer scientists, and legal scholars has pursued two main approaches to combat gerrymandering: 1) developing a metric (such as *efficiency gap* and *compactness*) to evaluate and optimize fairness across large collections of redistricting plans from simulations (Niemi et al. 1990; Stephanopoulos and McGhee 2015; Ko et al. 2022a); and 2) designing map-drawing algorithms to ensure overall partisan fairness. The first approach, however, depends on judicial rulings in partisan gerrymandering cases—after the federal courts' withdrawal in *Rucho v. Common Cause*, this path remains deeply contested (Chen and Rodden 2015; Tam Cho and Liu 2016). Our research aligns with the second approach and contributes to a growing body of work in which scholars propose interactive protocols that partition maps through negotiations between opposing parties (Landau, Reid, and Yershov 2009; Pegden, Procaccia, and Yu 2017; Mixon and Villar 2018; Benade, Procaccia, and Tucker-Foltz 2023; Palmer, Schneer, and DeLuca 2024), analogous to the classic "cake-cutting" problem (Brams et al. 2006). This approach does not rely on independent commissions or special masters to draw the map. Instead, each party acts in its own interest and takes turns making mapping decisions until they reach an unique subgame perfect equilibrium.

Game-theoretic LLM agents in negotiations. With the emergence of LLM agents, a growing body of research has explored whether LLMs can engage in strategic behavior through autonomous reasoning in negotiation settings. Several studies use classic game-theoretic environments to create controlled settings for evaluating LLM agents' human-like strategic interaction (Guo 2023; Mao et al. 2024; Fan et al. 2024; Gemp et al. 2024; Hua et al. 2024). Other work applies LLMs to real-world social deduction games such as Avalon (Light et al. 2025), Werewolf (Xu et al. 2024), and Chameleon (Karabag and Topcu 2025), examining whether agents can navigate complex rule-based behavior similar to humans. A parallel line of research investigates LLM agents in realistic economic contexts, such as trade and auction decision-making (Jiang, Xiong, and Liu 2025; Kwon et al. 2025), to test their ability to reason under market constraints. Building on this literature, our work extends the use of LLM-based agents to the high-stakes domain of political fairness, applying strategic negotiation to the problem of electoral redistricting.

Preliminaries

Redistricting Redistricting is the process of redrawing electoral district boundaries to reflect population changes and maintain fair political representation, typically carried out after each decennial census. In computational terms, it is often formulated as a graph partitioning problem. Let $G = (V, E)$ denote the adjacency graph of population units (e.g., precincts or census blocks), where each node $v \in V$ is assigned a population weight $p(v)$. The task is to partition V into k disjoint subsets $\{V_1, V_2, \ldots, V_N\}$, each representing a district, subject to the following conditions. Each district V_i must induce a connected subgraph of G (contiguity), and the population must be balanced such that

$$\left| \sum_{v \in V_i} p(v) - \frac{1}{N} \sum_{v \in V} p(v) \right| \leq \epsilon,$$

for a given tolerance ϵ. Additional constraints may also apply, including geometric compactness, preservation of communities of interest, or compliance with legal mandates such as the Voting Rights Act.

Metrics of Redistricting Evaluation Given a districting plan $\mathcal{M} = \{V_1, V_2, \ldots, V_N\}$ over a population graph $G = (V, E)$, we evaluate its quality using the following key metrics:

Population Deviation (PD) measures how equally population is distributed across districts (Stephanopoulos and McGhee 2015). It is defined as the average deviation from the ideal district population:

$$\mathrm{PD} = \frac{1}{N} \sum_{i=1}^{N} \left| \sum_{v \in V_i} p(v) - \frac{1}{N} \sum_{v \in V} p(v) \right|.$$

Polsby–Popper Score (PPS) quantifies the geometric compactness of a district (Polsby and Popper 1991). For each district V_i, it is computed as:

$$\mathrm{PPS}(V_i) = \frac{4\pi A_i}{P_i^2},$$

where A_i and P_i are the area and perimeter of district V_i, respectively. We report both the average and minimum PPS over all districts.

Partisan Bias (Bias) captures systemic advantage for one party over another (Grofman and King 2007). It is calculated as the average deviation from parity between vote share and seat share:

$$\mathrm{Bias} = \frac{1}{N} \sum_{i=1}^{N} (2 \cdot \mathrm{pct\_dem}_i - 1),$$

where $\mathrm{pct\_dem}_i$ is the Democratic vote share in district V_i. Values closer to 0 indicate fairer partisan balance. A positive value suggests a bias in favor of the Democratic Party (i.e., districts are drawn to favor Democrats), while a negative value indicates a bias in favor of the Republican Party.

Unfairness measures the proportion of residents whose preferred party did not win in their district (Ko et al. 2022b). For each district V_i with population P_i, we define:

$$\mathrm{unhappy\_votes}_i = \begin{cases} (1 - \mathrm{pct\_dem}_i) \cdot P_i & \text{if Dem win,} \\ \mathrm{pct\_dem}_i \cdot P_i & \text{otherwise.} \end{cases}$$

And the overall unfairness is computed as:

$$\mathrm{Unfairness} = \frac{\sum_i \mathrm{unhappy\_votes}_i}{\sum_i P_i}.$$

Lower *PD* and *Unfairness* indicate better population balance and greater voter satisfaction, higher ***PPS*** implies more compact districts, and smaller absolute ***Bias*** signals reduced partisan skew.

Gerrymandering Gerrymandering refers to the manipulation of electoral district boundaries to favor a specific party or group. The term dates back to 1812, when a Massachusetts district approved by Governor Elbridge Gerry was said to resemble a salamander—thus coining the term 'Gerry-mander.'

Method

In this section, we provide a detailed introduction to the **Agentmandering** framework. Agentmandering models redistricting as a structured interaction between two competing agents over a sequence of map construction rounds. As illustrated in Figure 2, the method consists of four core components: (1) Materials: a set of partisan agents representing competing political interests and corresponding district information, (2) Protocol: a Choose-and-Freeze protocol that alternates these actions until the full state is partitioned, (3) Choose Mechanism: a candidate generator that proposes feasible districting plans over the current unassigned region, and (4) Freeze Mechanism: a freeze mechanism that allows the opposing agent to lock in one district per round.

Materials The Agentmandering framework operates with two core agents: a Republican agent $\mathcal{A}_R$ and a Democratic agent $\mathcal{A}_D$, each representing the strategic interests of one major political party. These agents are powered by LLMs and are prompted to act in alignment with their respective party goals to defend and expand Republican or Democratic

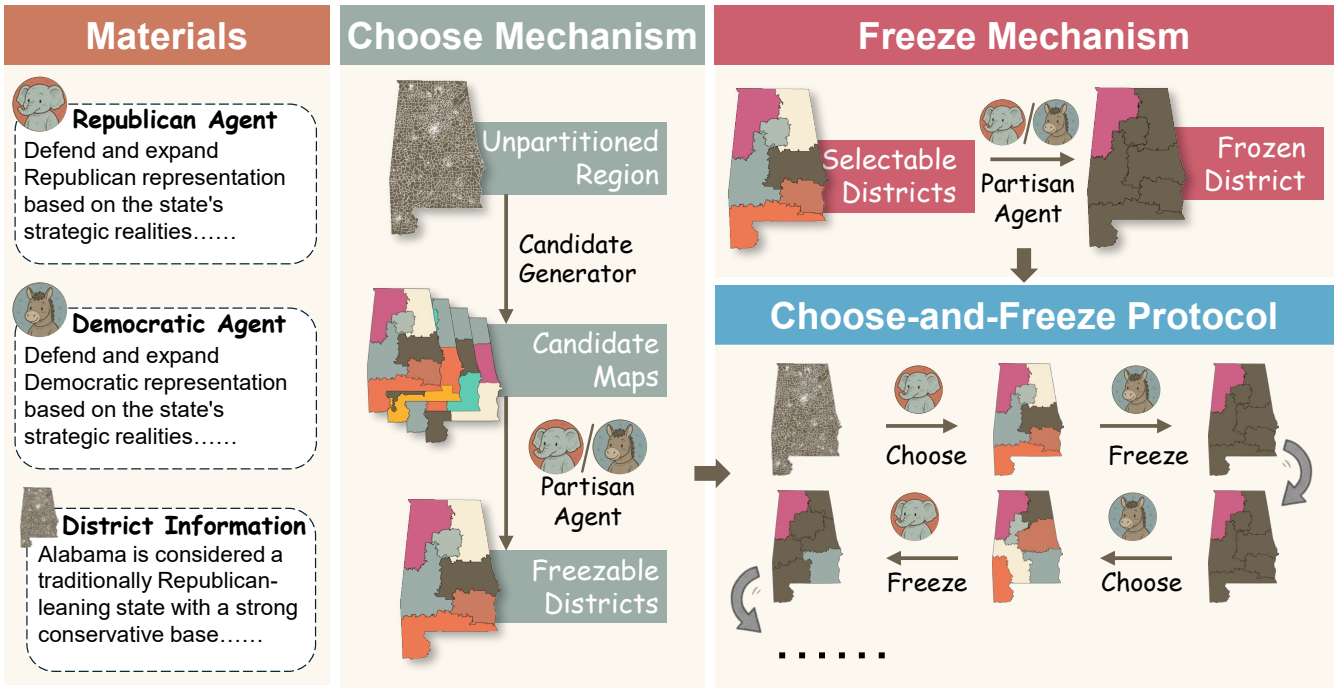

Figure 2: The Agentmandering framework.

representation based on local demographic and political conditions.

Each agent is equipped with a state-specific political profile $\mathcal{P}_{state}$, which includes historical voting trends, demographic composition (sourced from Census data), and partisan geography. This information, combined with racial demographics, provides strategic cues that guide the agent's behavior throughout the redistricting game.

Choose-and-Freeze Protocol The core of the Agentmandering framework is a sequential game played between a Republican agent $\mathcal{A}_R$ and a Democratic agent $\mathcal{A}_D$, who alternate roles over a series of rounds indexed by $n = 1, 2, \ldots, N$. Here N represents the total number of districts in the corresponding state. At each round, the agents jointly construct the redistricting map by interacting over a progressively shrinking unpartitioned region $\mathcal{R}_t$. At initialization, the unpartitioned region $\mathcal{R}_0$ is the entire state.

Each round consists of two key actions:

- **Choose:** One agent, denoted $\mathcal{A}_c \in \{\mathcal{A}_R, \mathcal{A}_D\}$, selects a preferred districting plan M_n^* from a small set of candidate maps $\mathcal{C}_n$ generated over the current region $\mathcal{R}_n$.
- **Freeze:** The opposing agent $\mathcal{A}_f \in \{\mathcal{A}_D, \mathcal{A}_R\}$ where $\mathcal{A}_f \neq \mathcal{A}_c$ selects one district $D_n^* \in M_n^*$ to be permanently fixed. The remaining territory is updated as $\mathcal{R}_{n+1} = \mathcal{R}_n \setminus D_n^*$.

This iterative process continues until the entire territory has been partitioned into districts. In each round, the choose and freeze agents will be switched. The game structure ensures that no single agent can unilaterally control the full outcome; instead, the final map emerges through a series of constrained, adversarial decisions.

Choose Mechanism In each round n, the candidate generator $\mathcal{G}(\cdot)$ produces a set of feasible candidate maps $\mathcal{C}_n$ over the current unpartitioned region $\mathcal{R}_n$. $\mathcal{G}(\cdot)$ is party-agnostic and shared by both sides. The size of $\mathcal{C}_n$ is c, and each plan $M_i \in \mathcal{C}_n$ satisfies population balance, contiguity, and legal constraints. The choice of $\mathcal{G}(\cdot)$ is flexible, in this work, we adopt the ReCom (DeFord, Duchin, and Solomon 2021) algorithm. Then the choosing agent $\mathcal{A}_c \in \{\mathcal{A}_R, \mathcal{A}_D\}$ selects one plan:

$$M_n^* = f_{\text{choose}}(\mathcal{A}_c, \mathcal{C}_n, \mathcal{P}_{state}),$$

where $\mathcal{P}_{state}$ denotes the political profile of the state and f_{choose} is a prompt-driven district selection function. The selected map M_n^* is then passed to the opposing agent for the freeze step.

Freeze Mechanism Upon receiving the selected plan M_n^*, the freezing agent $\mathcal{A}_f$, selects one district $D_n^* \in M_n^*$ to be permanently assigned. The selection is performed via a prompt-driven strategy function:

$$D_n^* = f_{\text{freeze}}(\mathcal{A}_f, M_n^*, \mathcal{P}_{state}),$$

where f_{freeze} evaluates each district in M_n^* based on its partisan composition and strategic implications for the freezing agent. Once frozen, the district D_n^* is removed from the unassigned region:

$$\mathcal{R}_{n+1} = \mathcal{R}_n \setminus D_n^*.$$

This procedure ensures that both agents influence the final map through alternating constrained actions, maintaining strategic balance throughout the game.

Experiments

Datasets, Baselines, and Metrics

We evaluate the effectiveness of the Agentmandering framework using redistricting data from U.S. states based on the post-2020 Census. The data includes population census data, voting history, demographic composition for each state as of 2020, as well as voting data from the 2020 presidential election.[1] The experiments use several baselines including Recom (DeFord, Duchin, and Solomon 2021), MergeSplit (Carter et al. 2019), FlipMCMC (Fifield et al. 2020), and SMCredist (McCartan and Imai 2020).

The evaluation metrics used in our experiments include **Population Deviation (PD)**, **Polsby-Popper Score (PPS)**, **Partisan Bias (Bias)**, and **Local Unfairness (Unfairness)**. *Population Deviation* measures the average population imbalance across districts; lower values indicate better population equality. *Polsby–Popper Score* evaluates the geometric compactness of each district; higher scores are preferred. *Partisan Bias* quantifies the asymmetry in partisan advantage; values closer to zero indicate fairer representation without systematic favor toward either Democrats or Republicans. *Local Unfairness* captures the extent to which voter preferences are respected within neighboring districts; lower values imply better local representational equity.

Performance of Agentmandering

Experiment Setup In this experiment, we evaluate whether Agentmandering produces fairer districting plans compared to traditional ensemble-based sampling methods. Both approaches rely on generating a large number of valid districting plans under legal and demographic constraints, but differ in how these plans are used. We use Gemini 2.5 pro (gemini-2.5-pro-preview-05-06) as our base LLM, and ReCom (DeFord, Duchin, and Solomon 2021) as our Candidate Generator. The temperature was set to 0 for all models and agent steps in every experiment.

In Agentmandering, each round involves selecting one district from a small set of c candidate maps, and a full run covers t districts in total. This results in $l = c \times t$ samples per run. For example, in Alabama, where $t = 7$ and $c = 100$, one complete Agentmandering game uses around 700 samples. If we repeat this process 10 times, the total number of samples is 7000.

To ensure a fair comparison, we allow the baseline method to generate the same total number of plans. However, unlike Agentmandering, which incrementally builds maps through an interactive process over a shrinking unassigned region, the baseline produces complete maps in one step. As a result, Agentmandering generates fewer final maps, but each is shaped through strategic agent interactions and controlled partisan dynamics.

[1]The district geographic information data is sourced from https://data.census.gov, and the voting data is from https://dataverse.harvard.edu/dataverse/electionscience.

Results Table 1 reports results on seven competitive swing states, with bold text indicating the best scores. The row **CD_2020** represents enacted districting plans. Agentmandering shows a strong advantage in *stability*, with standard deviations at least two orders of magnitude smaller than other methods, indicating reduced metric fluctuation and less room for strategic manipulation.

On **PD** (Population Deviation) and **PPS** (Polsby–Popper Score), all computational methods outperform CD_2020, revealing population imbalance and geometric distortion in real-world plans. Agentmandering slightly underperforms on PPS due to its irregular boundaries but achieves the lowest PD, reflecting superior population balance.

For **Bias** and **Unfairness**, which assess partisan neutrality and representational equity, Agentmandering performs best or near-best across most states. Notably, it is the only method to recover the correct partisan direction in Wisconsin, highlighted in italics. Its low Unfairness scores suggest more balanced and satisfying outcomes for voters.

Figure 3 shows that Agentmandering produces fairer and more stable maps than both Recom and CD_2020, despite using Recom for candidate generation. This demonstrates its robustness and capacity to mitigate partisan bias.

Effectiveness of LLM-Based Agent Decisions

Experiment Setup In this section, we evaluate the effectiveness of LLM-based agents' decisions within the Agentmandering framework. As a baseline, we implement rule-based variants for both the *choose* and *freeze* steps. The evaluation metric used is **Unfairness**. Specifically, we compare against the following decision rules:

- **Partisan Bias:** In the *choose* step, the agent selects the map that maximizes partisan advantage for its affiliated party; in the *freeze* step, it freezes the district that offers the greatest partisan gain.
- **Population Deviation:** In both *choose* and *freeze* steps, the agent selects the map or district with the smallest population deviation.
- **Compactness:** In both steps, the agent chooses the most compact option according to the Polsby–Popper score.

Results The results are shown in Figure 4. As observed, **Agentmandering achieves a lower Unfairness score than all rule-based variants**. This indicates that LLM-based agents are more effective in making politically strategic decisions—both in selecting appropriate candidate maps and in freezing reasonable districts. These results provide strong support for the value of integrating LLM reasoning with a game-theoretic mechanism to simulate human-like political behavior. Furthermore, the standard deviation of the rule-based variants is also 2–3 orders of magnitude lower than traditional baselines, suggesting that the overall stability of Agentmandering primarily stems from the *Choose-and-Freeze* protocol.

Effect of LLM Choice

We examine whether the choice of LLM affects Agentmandering's performance, given that redistricting involves polit-

Metric	Flip	Merge-split	SMC	Recom	CD-2020	Agentmandering
			Arizona (AZ)			
PD(10^{-3})	4.59±1.13	4.69±0.97	4.72±0.91	4.72±1.07	62.9	**4.19±0.00324**
PPS(10^{-2})	6.75±2.34	**7.02**±2.36	6.97±2.31	6.77±2.19	1.06	4.77±**0.00317**
Bias(10^{-2})	**2.68**±0.44	**2.68**±0.46	2.69±0.46	**2.68**±0.44	3.05	3.41±**0.00361**
Unfairness(10^{-1})	4.28±0.09	4.28±0.08	4.28±0.09	4.27±0.09	4.24	**4.03±0.00362**
			Georgia (GA)			
PD(10^{-3})	4.58±0.76	4.49±0.74	**4.08**±0.74	4.73±0.78	76.6	6.15±**0.00493**
PPS(10^{-2})	**4.68**±0.68	4.62±0.63	4.55±0.56	4.57±0.69	0.75	4.04±**0.00290**
Bias(10^{-3})	8.03±2.19	8.12±2.33	8.15±1.56	8.13±2.38	6.53	**7.63±0.00508**
Unfairness(10^{-1})	3.68±0.07	3.68±0.07	3.66±0.05	3.67±0.07	3.56	**3.49±0.00333**
			Michigan (MI)			
PD(10^{-3})	4.70±0.76	4.72±0.70	4.72±0.72	4.31±0.83	95.2	**3.70±0.00237**
PPS(10^{-2})	**6.03**±0.81	5.84±0.87	5.92±0.87	6.00±0.84	0.74	5.94±**0.00429**
Bias(10^{-2})	-2.34±0.14	-2.34±0.14	-2.34±0.13	**-2.30**±0.14	-2.06	-2.40±**0.00415**
Unfairness(10^{-1})	4.20±0.08	4.18±0.07	4.18±0.08	4.19±0.08	4.10	**3.96±0.00280**
			North Carolina (NC)			
PD(10^{-3})	4.73±0.73	4.61±0.77	4.64±0.74	4.57±0.75	94.2	**3.72±0.00196**
PPS(10^{-2})	5.98±0.81	6.00±0.88	**6.11**±0.97	5.87±0.91	0.85	5.31±**0.00186**
Bias(10^{-2})	-2.36±0.14	-2.37±0.15	-2.37±0.15	-2.32±0.15	-2.07	**-2.22±0.00134**
Unfairness(10^{-1})	4.18±0.09	4.19±0.09	4.19±0.08	4.19±0.09	4.15	**3.94±0.00209**
			Nevada (NV)			
PD(10^{-3})	4.20±1.44	4.30±1.43	4.06±1.47	4.31±1.44	49.3	**4.03±0.00292**
PPS(10^{-1})	2.04±0.35	2.08±0.39	**2.12**±0.38	2.07±0.37	0.32	2.05±**0.00212**
Bias(10^{-2})	4.91±0.57	4.93±0.54	4.95±0.54	4.87±0.54	3.17	**4.57±0.00419**
Unfairness(10^{-1})	4.39±0.04	4.39±0.03	4.39±0.03	4.39±0.03	4.29	**4.30±0.00468**
			Pennsylvania (PA)			
PD(10^{-3})	4.63±0.73	**4.32**±0.86	4.42±0.75	4.39±0.78	73.5	4.88±**0.00297**
PPS(10^{-2})	**4.84**±0.74	4.72±0.69	4.70±0.63	4.63±0.63	0.76	4.19±**0.00323**
Bias(10^{-3})	8.08±2.43	8.58±2.29	7.83±2.47	8.49±2.32	6.59	**7.86±0.00550**
Unfairness(10^{-1})	3.68±0.07	3.68±0.07	3.69±0.07	3.66±0.07	3.61	**3.48±0.00315**
			Wisconsin (WI)			
PD(10^{-3})	4.13±0.98	4.31±1.01	4.25±0.95	4.25±0.97	30.5	**3.54±0.00264**
PPS(10^{-2})	9.39±1.65	9.39±1.65	9.43±1.74	9.48±1.68	1.22	**9.56±0.00384**
Bias(10^{-3})	-1.20±2.57	-1.39±2.29	-1.11±2.31	-1.32±2.35	8.08	*1.29*±**0.00063**
Unfairness(10^{-1})	4.20±0.11	4.20±0.11	4.21±0.11	4.21±0.11	3.96	**3.93±0.00271**

Table 1: Experimental Results on Key Swing States: Bold indicates the best performance (with mean and standard deviation separated), and italics denote the same bias as the real situation, reflected in the sign of the Bias metric.

State	Gemini	GPT-4om	GPT-o3m	DS-R1	DS-V3	Claude-3.7	Llama3	Mixtral3.1	Qwen3
AZ	$4.03_{3.61}$	$4.10_{2.71}$	$4.07_{3.33}$	$4.04_{2.57}$	$4.05_{2.40}$	$4.04_{3.91}$	$4.06_{3.37}$	$4.08_{4.40}$	$4.10_{4.89}$
GA	$3.49_{3.33}$	$3.48_{2.28}$	$3.48_{2.43}$	$3.48_{1.51}$	$3.48_{3.15}$	$3.49_{4.53}$	$3.50_{2.42}$	$3.48_{4.45}$	$3.48_{5.52}$
MI	$3.96_{2.80}$	$3.95_{2.85}$	$3.95_{2.60}$	$3.96_{2.48}$	$3.95_{2.19}$	$3.97_{5.08}$	$3.97_{1.66}$	$3.97_{5.32}$	$3.95_{4.61}$
NC	$3.94_{2.09}$	$3.98_{2.27}$	$3.96_{2.03}$	$4.04_{2.05}$	$4.05_{2.80}$	$3.97_{3.99}$	$4.06_{2.02}$	$3.94_{4.67}$	$3.98_{3.30}$
NV	$4.30_{4.68}$	$4.29_{4.68}$	$4.38_{3.36}$	$4.34_{3.75}$	$4.26_{4.56}$	$4.37_{5.92}$	$4.36_{4.12}$	$4.30_{3.97}$	$4.29_{4.80}$
PA	$3.48_{3.15}$	$3.48_{2.69}$	$3.47_{3.25}$	$3.47_{2.99}$	$3.48_{3.89}$	$3.46_{2.79}$	$3.48_{3.58}$	$3.49_{3.96}$	$3.49_{3.53}$
WI	$3.93_{2.71}$	$3.97_{2.43}$	$3.95_{2.61}$	$3.94_{2.44}$	$3.96_{2.14}$	$3.96_{3.99}$	$3.94_{2.29}$	$3.96_{4.47}$	$3.97_{5.11}$

Table 2: Performance of Agentmandering across states using different LLMs. The mean values are in scientific notation (10^{-1}), and the standard deviations are in scientific notation (10^{-4}).

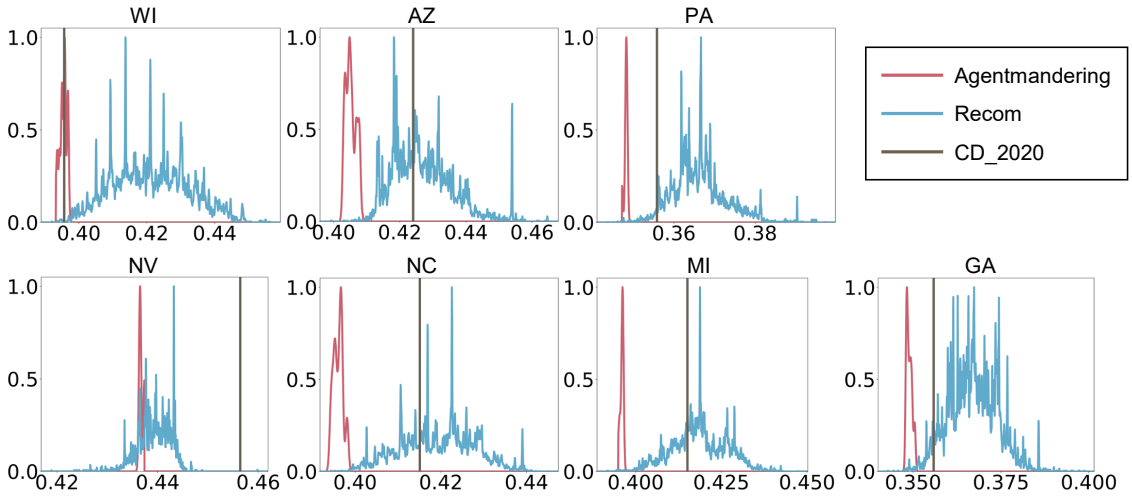

Figure 3: Normalized distribution of Unfairness across seven states for Agentmandering and Recom.

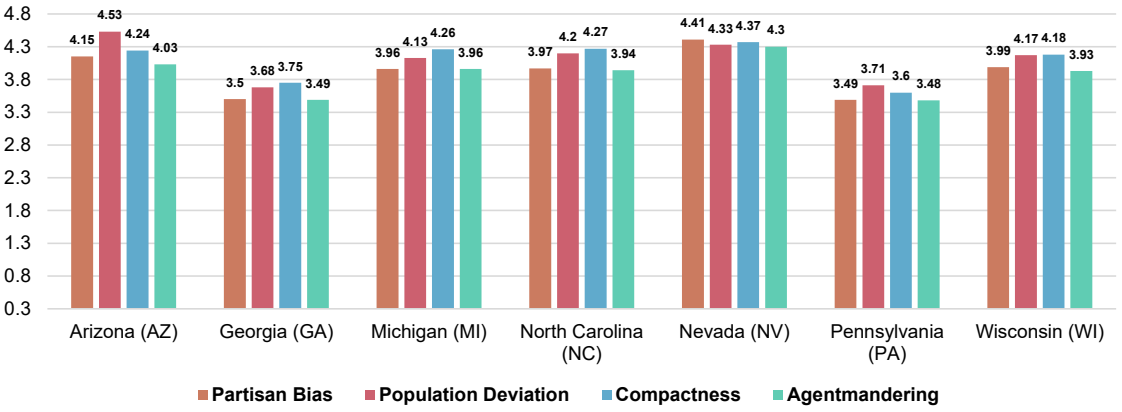

Figure 4: Comparison of LLM-based and rule-based agent decisions in terms of Unfairness.

ically sensitive reasoning and LLMs may differ in bias (Motoki, Pinho Neto, and Rodrigues 2024; Rozado 2024). We evaluate performance using **Unfairness** across a range of models, including Gemini 2.5 Pro, GPT-4o-mini, GPT-o3-mini, Deepseek-R1, Deepseek-V3, Claude-3.7, LLaMA-3-70B, Mixtral-3.1, and Qwen. The proprietary models are accessed via API; open source models run on a 4×A6000 Linux server.

Results. Table 2 shows that all models, including those developed in the United States (such as OpenAI and Anthropic), Europe (such as Mistral), and China (such as Deepseek and Qwen), achieve similar Unfairness scores across states and configurations. This consistency across national and institutional contexts suggests that Agentmandering is robust to differences in model origin, training data, or political orientation. The *Choose-and-Freeze* strategy provides sufficient structural guidance to ensure fairness, even when underlying LLMs vary, enabling institutions to substitute or upgrade models without degrading performance.

Conclusion

We introduce **Agentmandering**, a novel redistricting framework that harnesses large language model (LLM) agents to implement game-theoretic negotiation in practice. By simulating the *Choose-and-Freeze* protocol through interactive LLM agents, our approach transforms an abstract fairness mechanism into a scalable solution for real-world redistricting challenges. The resulting plans are procedurally transparent, strategically robust, and empirically fair across multiple metrics. This work makes two key contributions. First, it provides a new computational lens for political science by demonstrating how LLMs can model strategic partisan behavior in institutional settings. Second, it shows how LLM agents can bridge game-theoretic fairness and applied algorithmic decision-making.

Looking forward, future work will extend to investigate the challenges of applying this framework to multi-party systems, where fairness and strategy must be redefined to accommodate diverse party dynamics, coalition effects, and proportionality requirements.

References

Barnes, R.; and Solomon, J. 2021. Gerrymandering and Compactness: Implementation Flexibility and Abuse. *Political Analysis*, 29(4): 448–466.

Benade, G.; Procaccia, A. D.; and Tucker-Foltz, J. 2023. You can have your cake and redistrict it too. *arXiv preprint arXiv:2305.12079*.

Brams, S. J.; Jones, M. A.; Klamler, C.; et al. 2006. Better ways to cut a cake. *Notices of the AMS*, 53(11): 1314–1321.

Carter, D.; Herschlag, G.; Hunter, Z.; and Mattingly, J. 2019. A merge-split proposal for reversible Monte Carlo Markov chain sampling of redistricting plans. *arXiv [cs.DS]*.

Chambers, C. P.; Miller, A. D.; and Sobel, J. 2017. Flaws in the efficiency gap. *JL & Pol.*, 33: 1.

Chen, J.; and Rodden, J. 2015. Cutting Through the Thicket: Redistricting Simulations and the Detection of Partisan Gerrymanders. *Election Law Journal*, 14(4): 331–345.

Chikina, M.; Frieze, A.; and Pegden, W. 2017. Assessing significance in a Markov chain without mixing. *Proceedings of the National Academy of Sciences of the United States of America*, 114(11): 2860–2864.

Cox, G. W.; and Katz, J. N. 2002. *Elbridge Gerry's salamander: The electoral consequences of the reapportionment revolution*. Cambridge University Press.

DeFord, D.; Duchin, M.; and Solomon, J. 2021. Recombination: A family of Markov chains for redistricting. *Harvard Data Science Review*, 3(1): 3.

Fan, C.; Chen, J.; Jin, Y.; and He, H. 2024. Can large language models serve as rational players in game theory? a systematic analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 17960–17967.

Fifield, B.; Higgins, M.; Imai, K.; and Tarr, A. 2020. Automated redistricting simulation using Markov chain Monte Carlo. *Journal of Computational and Graphical Statistics*, 29(4): 715–728.

Fravel, J.; Hildebrand, R.; Goedert, N.; Travis, L.; and Pierson, M. 2023. Optimizing representation in redistricting: Dual bounds for partitioning problems with non-convex objectives. *arXiv preprint arXiv:2305.17298*.

Gelman, A.; and King, G. 1994. Enhancing Democracy Through Legislative Redistricting. *The American Political Science Review*, 88(3): 541–559.

Gemp, I.; Patel, R.; Bachrach, Y.; Lanctot, M.; Dasagi, V.; Marris, L.; Piliouras, G.; Liu, S.; and Tuyls, K. 2024. Steering Language Models with Game-Theoretic Solvers. arXiv:2402.01704.

Grofman, B.; and King, G. 2007. The future of partisan symmetry as a judicial test for partisan gerrymandering after LULAC v. Perry. *Election Law Journal*, 6(1): 2–35.

Guo, F. 2023. GPT in Game Theory Experiments. arXiv:2305.05516.

Hua, W.; Liu, O.; Li, L.; Amayuelas, A.; Chen, J.; Jiang, L.; Jin, M.; Fan, L.; Sun, F.; Wang, W.; Wang, X.; and Zhang, Y. 2024. Game-theoretic LLM: Agent Workflow for Negotiation Games. arXiv:2411.05990.

Jiang, K.; Xiong, L.; and Liu, F. 2025. HARBOR: Exploring Persona Dynamics in Multi-Agent Competition. arXiv:2502.12149.

Karabag, M. O.; and Topcu, U. 2025. Do LLMs Strategically Reveal, Conceal, and Infer Information? A Theoretical and Empirical Analysis in The Chameleon Game. arXiv:2501.19398.

Ko, S.-H.; Taylor, E.; Agarwal, P.; and Munagala, K. 2022a. All Politics is Local: Redistricting via Local Fairness. *Neural Information Processing Systems*, abs/2210.11643: 17443–17455.

Ko, S.-H.; Taylor, E.; Agarwal, P.; and Munagala, K. 2022b. All Politics is Local: Redistricting via Local Fairness. In Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; and Oh, A., eds., *Advances in Neural Information Processing Systems*, volume 35, 17443–17455. Curran Associates, Inc.

Kwon, D.; Hae, J.; Clift, E.; Shamsoddini, D.; Gratch, J.; and Lucas, G. M. 2025. ASTRA: A Negotiation Agent with Adaptive and Strategic Reasoning through Action in Dynamic Offer Optimization. arXiv:2503.07129.

Landau, Z.; Reid, O.; and Yershov, I. 2009. A fair division solution to the problem of redistricting. *Social Choice and Welfare*, 32(3): 479–492.

Li, H.; Gong, R.; and Jiang, H. 2025. Political actor agent: Simulating legislative system for roll call votes prediction with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 388–396.

Light, J.; Cai, M.; Chen, W.; Wang, G.; Chen, X.; Cheng, W.; Yue, Y.; and Hu, Z. 2025. Strategist: Self-improvement of LLM Decision Making via Bi-Level Tree Search. In *The Thirteenth International Conference on Learning Representations*.

Mao, S.; Cai, Y.; Xia, Y.; Wu, W.; Wang, X.; Wang, F.; Ge, T.; and Wei, F. 2024. ALYMPICS: LLM Agents Meet Game Theory – Exploring Strategic Decision-Making with AI Agents. arXiv:2311.03220.

McCartan, C.; and Imai, K. 2020. Sequential Monte Carlo for sampling balanced and compact redistricting plans. *arXiv [stat.AP]*.

Mixon, D. G.; and Villar, S. 2018. Utility Ghost: Gamified redistricting with partisan symmetry. *arXiv preprint arXiv:1812.07377*.

Motoki, F.; Pinho Neto, V.; and Rodrigues, V. 2024. More human than human: measuring ChatGPT political bias. *Public Choice*, 198(1): 3–23.

Niemi, R. G.; Grofman, B.; Carlucci, C.; and Hofeller, T. 1990. Measuring Compactness and the Role of a Compactness Standard in a Test for Partisan and Racial Gerrymandering. *The Journal of Politics*, 52(4): 1155–1181.

Palmer, M.; Schneer, B.; and DeLuca, K. 2024. A partisan solution to partisan gerrymandering: The define–combine procedure. *Political Analysis*, 32(3): 295–310.

Pegden, W.; Procaccia, A. D.; and Yu, D. 2017. A partisan districting protocol with provably nonpartisan outcomes. *arXiv [cs.GT]*.

Polsby, D. D.; and Popper, R. D. 1991. The third criterion: Compactness as a procedural safeguard against partisan gerrymandering. *Yale L. & Pol'y Rev.*, 9: 301.

Rozado, D. 2024. The political preferences of LLMs. *PloS one*, 19(7): e0306621.

Stephanopoulos, N. O.; and McGhee, E. M. 2015. Partisan gerrymandering and the efficiency gap. *U. Chi. L. Rev.*, 82: 831.

Tam Cho, W. K.; and Liu, Y. Y. 2016. Toward a Talismanic Redistricting Tool: A Computational Method for Identifying Extreme Redistricting Plans. *Election Law Journal*, 15(4): 351–366.

Xu, Y.; Wang, S.; Li, P.; Luo, F.; Wang, X.; Liu, W.; and Liu, Y. 2024. Exploring Large Language Models for Communication Games: An Empirical Study on Werewolf. arXiv:2309.04658.