Addressing the alignment problem in transportation policy making: an LLM approach

Xiaoyu Yan
Northwestern University
xiaoyuyan2028@u.northwestern.edu

Tianxing Dai
Northwestern University
tianxing.dai@northwestern.edu

Yu (Marco) Nie
Northwestern University
y-nie@northwestern.edu

Abstract

A key challenge in transportation planning is that the collective preferences of heterogeneous travelers often diverge from the policies produced by model-driven decision tools. This misalignment frequently results in implementation delays or failures. Here, we investigate whether large language models (LLMs)—noted for their capabilities in reasoning and simulating human decision-making—can help inform and address this alignment problem. We develop a multi-agent simulation in which LLMs, acting as agents representing residents from different communities in a city, participate in a referendum on a set of transit policy proposals. Using chain-of-thought reasoning, LLM agents provide Ranked-Choice or approval-based preferences, which are aggregated using instant-runoff voting (IRV) to model democratic consensus. We implement this simulation framework with both GPT-4o and Claude-3.5-Sonnet, and apply it for Chicago and Houston. Our findings suggest that LLM agents are capable of approximating plausible collective preferences and responding to local context, while also displaying model-specific behavioral biases and modest divergences from optimization-based benchmarks. These capabilities underscore both promise and limitations of LLMs as tools for solving the alignment problem in transportation decision-making.

1 Introduction

Urban transportation policy plays a central role in shaping regional development. Designing effective policy requires access to multidimensional data and a deep understanding of individual preferences across heterogeneous communities. Conventional approaches typically rely on structured mathematical models that identify an optimal policy under specified objectives and constraints. However, these models often rest on rigid assumptions and oversimplified behavioral representations. As a result, they may produce solutions that are analytically tractable yet poorly aligned with public sentiment or the complex realities of policy implementation. This misalignment frequently contributes to delays—or even failures—in policy approval and execution.

Recent advances in large language models (LLMs) offer a promising opportunity to address this alignment problem. Trained on vast corpora of text encompassing news, facts, and human discourse, LLMs possess a rich contextual understanding that could help policymakers infer public preferences and explore trade-offs before implementation. Their ability to interpret unstructured information, reason about competing objectives in natural language, and adapt to specific contexts suggests a new form of decision support—one that complements the traditional paradigm.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: .

In this study, we implement a multi-agent voting framework to examine the potential of LLMs in supporting transportation policy design. We simulate collective decision-making by deploying autonomous LLM agents as representatives of heterogeneous communities within a large city. These agents deliberate over transit policy proposals involving three levers: a dedicated sales tax for transit services, fare policies, and driver fees (e.g., congestion charges). This design enables us to study how collective preferences form, how trade-offs are negotiated across constituencies, and how democratic mechanisms can be modeled within an AI-driven environment. Crucially, the framework distributes reasoning tasks across agents, enabling scalable deliberation.

To ground the experiment in established transportation planning practices, we incorporate a standard utility-based travel demand model. Travelers, characterized by their daily disposable income, choose between driving and transit, both affected by congestion externalities. The model estimates how each policy scenario impacts travel experience and utility. These outputs are provided to LLM agents to guide deliberation. Additionally, the model yields utility distributions across the income spectrum, allowing us to rank policy alternatives based on normative objectives. The rankings serve as benchmarks to evaluate LLM agents' choices. We implement both ranked-choice and approval voting to examine how different aggregation rules shape collective outcomes. Finally, we systematically vary the information available to agents to assess how domain knowledge affects decision quality and alignment with model-based benchmarks.

Our research addresses several key questions: (1) To what extent can LLM agents generate coherent, collective policy preferences in a simulated voting environment? (2) How do different voting mechanisms and prompt designs influence stability, diversity, and bias in decision outcomes? (3) To what extent do the policies emerged from LLM-based voting deviate from those recommended by the conventional model and how do we explain these discrepancies? (4) How do results generalize across different urban contexts and language models?

We find LLM-based referendums often align with model-based benchmarks, selecting similar priorities despite detail differences. LLM agents show stronger tax aversion than model optima. GPT-4o yields more consistent patterns than Claude-3.5-Sonnet, though both converge on average preferences. Top policies remain stable across voting methods, indicating procedural robustness. Sentiment analysis shows GPT-4o maintains uniformly positive tones, while Claude-3.5-Sonnet is more varied, affecting vote dispersion. Context sensitivity emerges: in Houston, GPT-4o favors lower taxes and higher driver fees than in Chicago, reflecting awareness of local sociopolitical conditions.

2 Related studies

Transportation planning Travel forecasting models have guided urban transportation planning for more than half a century [3, 18, 33, 21, 4]. The traditional four-step model, developed in the postwar era, forecasts long-term demand based on assumptions about population, land use, and infrastructure. Since the 1970's, these models have been grounded in random utility theory, assuming travelers maximize latent utility based on cost, time, and socio-demographic attributes [20]. While statistically convenient and influential in shaping investments, this paradigm faces enduring criticism for behavioral rigidity, inability to capture bounded rationality and social influence [2, 16, 30], and poor adaptation to structural shifts and exogenous shocks [28, 14]. More fundamentally, the optimization focus of these models—typically welfare or efficiency—often clashes with public values. Model-derived "optimal" policies can misalign with community preferences, generating friction during implementation, a phenomenon termed the *technocratic disconnect* [12, 31, 32]. The process, portrayed as value-neutral, overlooks the inherently political nature of planning and its distributive consequences [19, 9]. Accuracy issues further compound these challenges. Studies report demand forecast errors exceeding 20% for most rail projects and ±30–50% for major road investments [34, 11, 14]. These persistent deficiencies underscore the need for complementary tools that integrate public sentiment and deliberation. Motivated by this gap, we explore whether LLMs can serve as a bridge between technical analysis and democratic alignment in transportation planning.

Applications of LLMs in simulating complex human decision-making processes Recent work has leveraged LLMs to simulate complex human decision-making without relying on explicit structural assumptions. Han [13] showed that LLMs adaptively infer relevant factors in classical decision tasks, while Ross et al. [29] mapped behavioral biases in LLMs across economic games, finding decision patterns distinct from both rational-agent and human baselines. LLMs have also been used to model social dynamics. Park et al. [26] demonstrated that persona-driven agents can reproduce opinion formation and collective behaviors, and Park et al. [27] showed strong alignment between LLM-generated survey responses and human data, supporting their use as proxies in behavioral experiments. Beyond experimental contexts, LLM agents have been applied to practical domains, including auction design [7], legal reasoning [6], and macroeconomic forecasting [17], underscoring their potential as flexible tools for decision analysis in complex, high-stakes environments.

LLMs in urban and transportation planning Recent studies highlight the potential of LLMs in urban planning. Zhou et al. [36] showed that LLMs can match or surpass traditional reinforcement learning in complex city planning tasks. Building on this, Zhou et al. [37] introduced a multi-agent framework where LLM agents, simulating diverse community residents, used a fishbowl discussion mechanism to allocate land under competing priorities, demonstrating LLMs' ability to support participatory planning. Similarly, Ni et al. [23] developed a closed-loop system where resident agents provided real-time evaluations of planning proposals, enabling adaptive, stakeholder-informed strategies and modeling human-centered planning dynamics.

3 Methodology

We consider a stylized city composed of $\mathbb{I} = \{1, \cdots, I\}$ communities. Its transportation agency is evaluating a policy change to its current transit system, consisting of three key components: a flat fare paid by each rider (denote by r), a dedicated sales tax levied on all residents (t), and a per-trip fee paid by drivers (τ). Accordingly, use a tuple $\boldsymbol{p}_k = \{r_k, t_k, \tau_k\}$ to denote the values of a policy $k \in \mathbb{K} = \{1, \cdots, K\}$. To streamline the decision process, the agency defines three levels—low (l), medium (m), and high(h)—for each of the three policy levers. In other words, r_k, t_k, and τ_k must take one of three values contained in set $\{l, m, h\}$. This results in a policy set $\mathbb{K}$ consisting of $K = 27$ distinct policy proposals.

The analysis proceeds in two stages. First, the agency evaluates the K proposals using a conventional transit policy design model. This model provides estimates of performance metrics such as travel times, trip costs, and congestion levels. Second, the same set of proposals is submitted to a multi-agent simulation framework powered by a LLM. This simulator emulates a city-wide referendum, where agents representing different communities deliberate and vote, informed in part by the output of the conventional model.

3.1 Transit design model

We take the transit design model from Dai et al. [9], which begins with a simplified representation of a city and its transportation system. The city is modeled as a square with a grid street network and evenly distributed residents and travel demand. Individuals differ by income level, which shapes their travel choices and how they experience transportation costs and benefits. A bus network is overlaid on the city's street grid. Its service quality is governed by three operational parameters: headway (how frequently buses arrive), stop spacing, and route spacing. These design elements collectively determine the accessibility of the bus system. Moreover, the system is financed through a combination of rider fare r, a local sales tax on residents t, a fee paid by drivers τ (e.g., a congestion charge or fuel tax), and an exogenous government subsidy. The transit agency must determine both the service configuration and the financing strategy, balancing operational quality with fiscal feasibility.

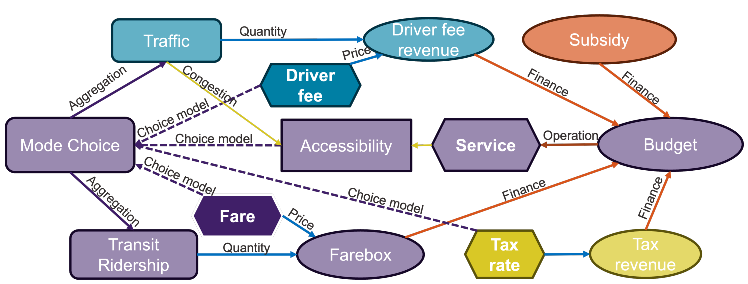


Figure 1: Joint design of public transit service and policy.

Figure 1 illustrates the structure of this integrated design model. Key decision variables include both service parameters (headway, stop spacing, and route coverage) and policy instruments (r, t, and τ). At the center of the system is individual mode choice between bus or car, which depends on the utility of each mode. The utility is a function of income, cost, and accessibility. At the same time, individual choices aggregate into system-level travel patterns that shape congestion, influence accessibility and affect overall revenue for transit through a complex feedback loop. The agency then adjusts its service and financial decisions, subject to a budget constraint requiring that operating costs be covered by fares, taxes, driver fees, and subsidies.

The above model is employed to evaluate system objectives, transit ridership, and distributive effects after a transit policy $k \in \mathbb{K}$ is implemented. For each policy $k \in \mathbb{K}$, the corresponding transit mode share is denoted as γ_k. We produce two normative objectives: the total utility of all travelers U_k (a utilitarian objective) and the utility of the most disadvantaged traveler u_k(an egalitarian objective). Roughly speaking, U_k measures efficiency whereas u_k gauges distributive effects. Another metric, a more direct indicator of distribution effects, is the Gini index, denoted as G_k. These metrics serve as a benchmark for evaluating the performance of our LLM-based simulation framework.

3.2 LLM-based multi-agent simulation

We implement a multi-agent simulation framework in which LLM agents serve as representatives of communities within a city, as illustrated in Figure 2. Each agent is tasked with evaluating and voting on the K transit policies, as described at the beginning of Section 3. Recall that each policy $p \in \mathbb{K}$ is formed by choosing a transit fare r_k, sales tax rate t_k, and driver fee τ_k from one of three values $\{l, m, h\}$. The simulation is designed to explore whether LLMs can approximate community-based deliberation and consensus formation across diverse constituencies.Two state-of-the-art LLMs are used in our simulation: GPT-4o from OpenAI [25] and Claude-3.5-Sonnet from Anthropic [1]. Both models are used with temperature set to zero to ensure reproducibility and eliminate stochastic variation in the outputs. See Appendix G for the sample prompts.

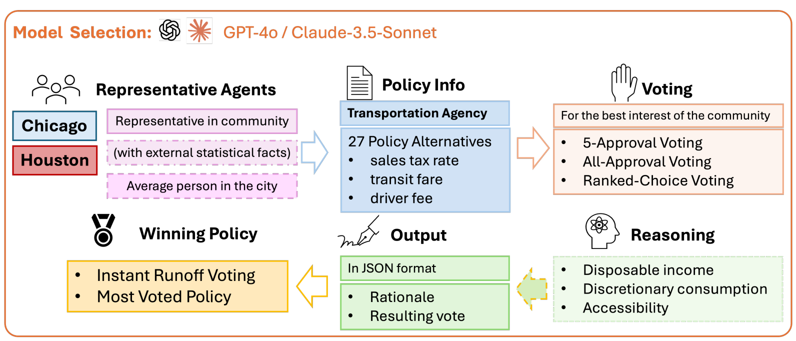


Figure 2: Framework of the multi-agent LLM simulation.

We construct three types of agents to examine the role of context and knowledge in shaping LLM decision-making: (1) **Community-based agents (CHI-com)** represent individual communities and acted based solely with pre-tained knowledge; (2) **Knowledge-augmented agents (CHI-know)** build on the first configuration, but are additionally prompted using localized demographic and economic data—such as average household expenditures, transit reliance, and income levels—to test whether broader contextual grounding improves judgment; (3) **City-average agents (CHI-avg)** simulate a generic "average" resident of the city, without community-specific difference.

Each agent $i \in \mathcal{I}$ is instructed to evaluate the K policy alternatives and cast votes based on what best serves its community's interests. Let $\mathbf{v}_i = \{v_{i,0}, \ldots, v_{i,K-1}\}$ denote the vote cast by agent i. Since there is no universally accepted voting rule for simulated multi-agent environments, we experiment with three collective decision rules commonly studied in the literature [35]: (1) **5-Approval voting**: Each agent approves exactly five proposals. Formally, $v_{i,k} \in \{0, 1\}$ for all k, with $\sum_{k=0}^{K-1} v_{i,k} = 5$ for each $i \in \mathcal{I}$; (2) **All-approval voting**: Agents may approve any number of proposals they deem acceptable, i.e., $v_{i,k} \in \{0, 1\}$, without any constraint on the number of approvals; (3) **Ranked-choice voting**: Each agent assigns a unique rank to up to five proposals. Formally, $v_{i,k} \in \{1, 2, 3, 4, 5, \text{null}\}$, where lower values indicate higher preference, and each agent assigns at most one value from

$\{1, 2, 3, 4, 5\}$ to any proposal. That is, for all $i \in \mathcal{I}$, $|\{k : v_{i,k} \in \{1, 2, 3, 4, 5\}\}| \leq 5$ and $v_{i,k} \neq v_{i,k'}$ for $k \neq k'$ whenever both are ranked.

To encourage deeper reasoning over pattern matching, we use structured chain-of-thought prompts, directing each agent to consider three key factors before voting: (1) **Disposable income**: How the proposed policy affects residents' income after taxes, fares, and fees; (2) **Discretionary consumption**: How much income is left for non-essential goods and services; (3) **Accessibility**: The ability to reach daily destinations—such as work, shopping, or healthcare—under the transit and congestion conditions implied by each policy.

The prompts aim to elicit civic-style reasoning, allowing LLMs to weigh trade-offs using contextual knowledge. Each agent outputs its community ID, a written rationale, and a ranked vote. To ensure consistent and structured reasoning, we designed standardized system and user prompts (Appendix G). The system prompt casts the agent as a representative of a specific community, directing it to reason step by step and rank policy proposals based on three key factors. The user prompt outlines the referendum context, describes the 27 policy options, and presents performance metrics (e.g., cost, travel time), emphasizing their effects on travel behavior and household budgets.

4 Results

In this section, we first benchmark the conventional model as a normative reference. Then we report Chicago simulation outcomes and compare different voting methods, contextual information and LLM models. We also conduct a regression analysis to better understand LLM agents' preference and voting behavior. Lastly, we test generalizability via a Chicago–Houston comparison. Experiments are performed in Python 3.11 with Apple M2 chip, 10 GB memory.

4.1 Model-based results

The transit design model by Dai et al. [9] employs utility-based mode choice, with utility influenced by income, congestion, service levels, fares, and policy levers (taxes and driver fees). The model is calibrated to match four key statistics: (i) mode share, (ii) ridership, (iii) farebox recovery, and (iv) budget. We adopt the same parameters for Chicago; Houston is re-calibrated to local conditions. See Table 6 in Appendix C for calibration outcomes.

In all experiments, the policy space has three levers $\{l, m, h\}$ with three levels each: sales tax t : $\{0.5\%, 1\%, 1.5\%\}$; fare r: $\{\$0.75, \$1.25, \$1.75\}$/trip; driver fee τ: $\{\$0, \$0.50, \$1\}$/trip. We evaluate 27 policy combinations with performance metrics computed by the calibrated model for both the case of Chicago and Houston (see Appendix B). LLM agents receive only mode-specific travel times and trip costs to guide voting, while other outputs—average utility (U_k), minimum utility (u_k), Gini index (G_k), and transit mode share (γ_k)—are reserved for evaluation. U_k measures system efficiency, and u_k, G_k, and γ_k capture distributional equity.

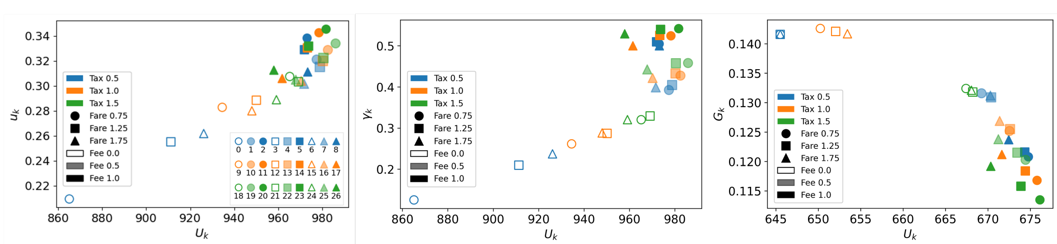


Figure 3: Comparison of u_k, γ_k, and G_k against U_k for different policy configurations in Chicago.

The status quo (*Policy 12*) does not lie on the Pareto frontier, which is defined by the upper-right envelope in the left and middle plots and the lower-right envelope in the right plot of Figure 3. By increasing the sales tax and driver fee while keeping the transit fare constant or reduced, one can simultaneously increase U_k and γ_k, and decrease G_k, thereby achieving both more utilitarian and more egalitarian outcomes. The underlying mechanism is intuitive: greater transit subsidy via taxes and driver fees enhances the level of service, induces a shift from driving to transit, and ultimately reduces congestion—improving the overall efficiency of the transportation system.

In terms of model-optimal outcomes, Figure 3 shows that the *Utilitarian* solution corresponds to *Policy 19*, which maximizes the total utility U_k. However, the minimum utility u_{19} under this policy is not the highest among all u_k, indicating that maximizing U_k doesn't necessarily benefit the most disadvantaged travelers. By contrast, the *Egalitarian* solution is *Policy 20* , which achieves the highest minimum utility, the greatest transit mode share, and the lowest Gini index among all policy options. Generally, lowering the transit fare tends to move outcomes closer to the Pareto frontier. More importantly, we observe a tight cluster of policies located near the Pareto frontiers, meaning that many policies produce outcomes similar to the utility-based Pareto-optimal policies. If one were to adopt an alternative social welfare function, any of these clustered policies could emerge as optimal. This diversity sets the stage for the LLM-based voting experiment that follows.

4.2 LLM referendum for Chicago

Chicago has 77 community areas. Thus, the simulated referendum is participated by 77 communities members represented by LLM agents. We begin by comparing the outcomes produced by the three voting methods described earlier. We then explore how contextual information and knowledge embedded in the prompts influence agents' voting patterns. All results reported in this section are generated using the GPT-4o model. The comparison with Claude-3.5-Sonnet will be discussed later.

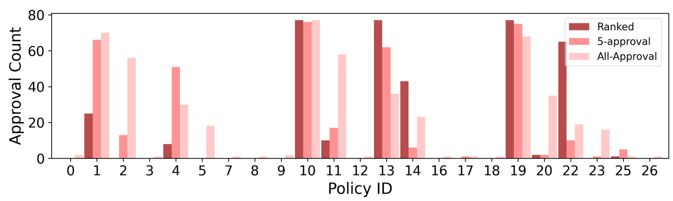


Figure 4: Voting counts of different voting types (single round).

Results from different voting methods The readers are referred to Appendix H for examples of responses from the three voting methods. Figure 4 shows the distribution of votes from a single round under each of the three voting methods. Overall, the results reveal a broadly consistent preference for a small set of top-ranked policies, particularly *Policies 10, 13, and 19*. Under the Instant Runoff Voting rule, *Policy 10* emerges as the winner in the ranked-choice voting scheme (see Appendix A.1 for the mathematical definition of the winning policy under different voting methods). It also receives the highest number of votes in both the 5-approval and all-approval settings, making it the winning policy across all three methods. While *Policy 10* is neither the *Utilitarian* nor the *Egalitarian* optimum identified by the model, it lies close to the Pareto frontier (see Figure 3), suggesting that the simulated referendum is capable of selecting policies that align well with model-based recommendations.

That said, the approval-based methods (5-approval and all-approval) produced more dispersed voting patterns, showing a greater tendency to favor policies with lower tax rates beyond the commonly agreed-upon options. The all-approval method, in particular, admits several policies (e.g., Policies 5, 7, 8, 9, 16, and 26) that are not selected under either the ranked-choice or 5-approval methods. These outlier policies are neither efficient nor equitable based on the model-based evaluation in Figure 3.

Table 1: Winning policies, mean policy values and entropy (single round).

	Winning Policy				Mean policy values			Entropy
	ID	t	r	τ	t	r	τ	E
Ranked-Choice	10	1.000	0.750	0.500	0.833	0.917	0.750	2.739
5-Approval	10	1.000	0.750	0.500	1.077	1.096	0.731	2.928
All-Approval	10	1.000	0.750	0.500	0.978	1.185	0.587	3.565

Table 1 shows that *Policy 10* wins across all voting methods. Compared to mean policy values, it features similar tax rates but lower fares and driver fees—indicating a democratic compromise that favors broader support over stronger user-based funding preferences. Among all options, *Policy 19* consistently ranks as a favored policy in all three voting methods. Notably, this policy coincides with the *Utilitarian* policy identified by the model. In contrast, the *Egalitarian* policy, *Policy 20*, receives substantially fewer votes—failing to make the top five even under the all-approval method. This result suggests that the aggregate preferences captured by GPT-4o agents in the Chicago simulation tend to align more closely with utilitarian values than with egalitarian ones.

Overall, we find strong consistency across the three voting mechanisms, with ranked-choice and 5-approval producing especially similar outcomes. All-approval voting, on the other hand, yields more scattered results and diverges notably from the other two. For this reason, we focus on the ranked-choice method in the remainder of our analysis.

Impact of contextual information We introduce three referendum scenarios for the three types of agents previously introduced: (1) *CHI-com* - **community-based agents**, (2) *CHI-know* - **knowledge-augmented agents**, and (3) *CHI-avg* - **City-average agents**. For each scenario, we run the simulation for 10 rounds, identify the winner for each round and analyze the distribution of votes across policies and the three policy levers. As shown in Table 2, GPT-4o under the CHI-com setting exhibits highly stable preferences. Across 10 simulated referendum rounds, *Policy 10* consistently emerges as the winning option. The distribution of votes across all policies shows only modest variation, reflected in a moderate entropy value of $E = 2.739$ (the larger the entropy, the more variations in the votes, see Appendix A.2 for details). The mean policy values indicate a clear and consistent preference for low fares, medium tax rates, and medium driver fees.

Table 2: Winning policy, mean entropy values and mean policy values across ten-round referendums.

Model	Scenario	Winner (counts)	Entropy E	$\bar{t}_1$ ($\bar{e}_{t\|1}$)	$\bar{r}_1$ ($\bar{e}_{r\|1}$)	$\bar{\tau}_1$ ($\bar{e}_{\tau\|1}$)
GPT-4o	CHI-com	P10 (10)	2.739	0.983 (0.21)	0.782 (0.34)	0.511 (0.15)
	CHI-know	P10 (8), P11 (2)	2.804	0.999 (0.02)	0.772 (0.25)	0.721 (0.98)
	CHI-avg	P10 (8), P11 (2)	2.611	1.000 (0.00)	0.750 (0.00)	0.600 (0.72)
Claude-3.5	CHI-com	P1 (6), P10 (4)	4.022	0.802 (1.29)	0.951 (1.08)	0.546 (1.52)
	CHI-know	P1 (8), P4 (2)	3.902	0.782 (1.34)	0.986 (1.04)	0.579 (1.28)
GPT-4o	HOU-com	P2 (10)	3.583	0.635 (0.92)	0.785 (0.29)	0.895 (0.74)

Note: $E \in [0, 4.75]$ measures the concentration of votes across all policies; $\bar{e}_{x|s} \in [0, 1.58]$ measures the concentration of the rank s votes on the three policy levers (x) across the three levels.

When agents are augmented with external contextual knowledge (CHI-know), the preference landscape shifts slightly. The winning set expands to include higher driver fee options—*Policy 11* wins in 20% of the rounds. This change is also evident in the rank-1 mean policy values, where the average driver fee increases by 50% compared to CHI-com. The entropy value rises slightly, suggesting a more dispersed distribution of preferences when additional knowledge is available. In the CHI-avg setting, where a single agent represents the average Chicagoan, the set and frequency of winning policies mirror those under CHI-know. However, the entropy value for CHI-avg is noticeably lower, indicating more concentrated preferences. Most strikingly, all rank-1 votes in this scenario support policies with a medium tax rate and the lowest fare—resulting in zero entropy for those policy levers.

In sum, while all three scenarios broadly agree on the most preferred policies, community-specific simulations introduce greater diversity in preference expression — more so when additional contextual information is provided, while the averaged-agent setting yields more concentrated but potentially less nuanced outcomes.

4.3 GPT-4o vs. Calude-3.5-Sonnet

Table 2 shows that, for CHI-com, the winning policy set expands from *Policy 10* under GPT-4o to include both *Policy 10* and *Policy 1* under Claude-3.5-Sonnet and the winning probability of *Policy 10* falls from 100% to just 40% when switching to Claude-3.5-Sonnet. The key distinction between these two policies lies in the sales tax rate: *Policy 10* preserves the status quo at 1%, while *Policy 1* lowers it to 0.5%. For CHI-know, the winning policies diverge entirely: Claude-3.5-Sonnet selects *Policy 1* and *Policy 4*, while GPT-4o favors *Policy 10* and *Policy 11*. Their mean policy values also drift much further apart, underscoring that the two LLMs respond to local contexts in markedly different ways.

Compared to GPT-4o, Claude-3.5-Sonnet exhibits significantly more dispersed voting behavior, as reflected in a roughly 50% increase in the overall entropy measure (also see Appendix D for vote distribution visualizations). The entropy values for individual policy levers are even higher—for instance, the entropy of rank-1 votes for the driver fee is 1.52, indicating a near-uniform distribution and aggregate indifference. To verify this difference between the two models, we conducted sentiment analyses of the LLM-generated rationale texts from both models and found that GPT-4o exhibits consistent positive sentiments across all communities while Claude-3.5-Sonnet produces a wider dispersion of sentiments. See the definition and results in Appendix A.3 and E, respectively.

4.4 Preference of LLM agents via regression

To delve deeper into the decision making rationale of LLM agents and unveil the impact of contextual information, we specify a set of regression models to reveal the relation between the agents' preferences for each of the three policy levers and their sociodemographic attributes. To quantify each agent's preference, we apply the Borda Count method, which assigns scores based on the rank order of each policy. Formally, let $\bar{s}_{ix}$ denote the average Borda score assigned according to the policy lever $x \in \mathcal{L}$ for community area i over 10 simulation rounds, i.e

$$\bar{s}_{ix} = \frac{1}{J} \sum_{j=1}^{J} \frac{\sum_{s=1}^{5} (6-s) q(x|k(i,j,s))}{\sum_{s=1}^{5} s}. \tag{1}$$

where j indexes the referendum round in the total set of J rounds, and $s \in \{1, 2, 3, 4, 5\}$ denotes the rank position (with $s = 1$ corresponding to the top rank). The term $\frac{6-s}{\sum_{s=1}^{5} s}$ specifies the normalized weight for the s-th rank, while $q(x|k(i, j, s))$ returns the level (i.e., l, m or h) of policy lever x in policy k that is ranked by community i at s in round j. Thus, $\bar{s}_{ix}$ is the average normalized Borda score assigned to policy lever x by community i across J rounds under the ranked voting process.

In addition to standard sociodemographic attributes, we introduce interaction terms between each attribute and a binary treatment indicator D_i, where $D_i = 1$ if the observation originates from the CHI-know condition (with factual context) and $D_i = 0$ if from the CHI-com condition (without additional facts). We specify an Ordinary Least Squares (OLS) regression model as follows:

$$\bar{s}_{ix} = \beta_{x0} + \boldsymbol{\beta}_x^\top \mathcal{X}_i + \gamma_{x0} D_i + \boldsymbol{\gamma}_x^\top \mathcal{X}_i D_i + \varepsilon_{ix}, \tag{2}$$

where β_{x0} is the intersect associated with policy lever x, $\mathcal{X}_i$ denotes the vector of sociodemographic covariates for agent i, γ_{x0} is the coefficient for the treatment indicator D_i, $\boldsymbol{\beta}_x$ is the vector of coefficients for the covariates and $\boldsymbol{\gamma}_x$ is the vector of coefficients of the interaction terms.

Table 3: Regression results for tax rate, transit fare, and driver fee under GPT-4o and Claude-3.5.

	Tax rate		**Transit fare**		**Driver fee**	
	GPT-4o	Claude-3.5	GPT-4o	Claude-3.5	GPT-4o	Claude-3.5
constant	1.123***	0.868***	0.968***	0.963***	0.535***	0.522***
none-white %	0.004	-0.031	0.005	0.010	-0.003	-0.085*
number of cars	-0.009	0.073**	0.008	0.018	0.024**	0.032
transit commuter %	0.010†	0.025	-0.002	-0.059***	-0.008	0.036
income < 25k %	-0.033***	-0.056*	-0.013†	0.001	-0.008	0.010
income > 150k %	-0.013	0.076***	0.021**	0.157***	0.024**	0.111*
info_dummy	-0.025***	-0.018	-0.047***	0.037**	0.150***	0.024
none-white % × info	0.007	0.017	0.001	-0.005	0.008	0.086†
number of cars × info	0.000	-0.017	0.003	-0.048†	0.001	0.011
transit commuter % × info	-0.009	0.004	-0.030***	0.026	0.025**	0.032
income < 25k % × info	0.035**	0.020	-0.012	-0.015	0.014	-0.037
income > 150k % × info	-0.006	0.045	0.015	-0.067*	0.018	0.035
R^2	0.45	0.724	0.729	0.782	0.875	0.593
Adj. R^2	0.407	0.702	0.708	0.765	0.865	0.561
F-statistics	10.54 ***	33.81***	34.79***	46.18***	89.99***	18.79***

Note. Significance: $^{\dagger}p < 0.1$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Table 3 reports the regression results (refer to Appendix F for the details on model selection). The effect of the information dummy is pronounced in the GPT-4o models. Across all three policy domains, the variable is highly significant, indicating that the provision of additional information systematically shapes preferences and improves explanatory power. Notably, the negative coefficients for tax rate and transit fare suggest that, when contextual signals are incorporated, GPT-4o registers stronger aversion to financial burdens imposed by higher taxes or fares. By contrast, the positive and relatively large coefficient for driver fee implies that information makes the model more inclined to endorse this revenue instrument, reflecting a redistribution logic that shifts costs toward car users rather than vulnerable transit riders. Claude-3.5-Sonnet, in comparison, shows weaker and less consistent responses to the information variable, suggesting that it internalizes contextual cues less directly. With local context, Claude-3.5-Sonnet agents become more open to raising transit fare, while indifferent to the other two instruments.

Overall, the regression model explains more variance in tax rate and transit fare outcomes in the data generated by Claude-3.5-Sonnet, while showing stronger explanatory power for driver fee outcomes in GPT-4o. Substantively, GPT-4o agents focus narrowly on income and car ownership as key determinants, whereas Claude-3.5-Sonnet agents draw on a broader set of sociodemographic factors, including car ownership, transit share, and minority status. Reassuringly, the signs of the significant correlations align with intuitive expectations: low-income communities tend to resist higher taxes and fares, while car-less or transit-reliant communities support higher fees on drivers.

4.5 Chicago vs Houston

To assess urban context effects, we replicate the simulation in Houston, a similarly sized but more car-oriented city with lower transit use and less public investment than Chicago. These contrasts provide a testbed for evaluating the LLM-based voting framework's adaptability and generalization.

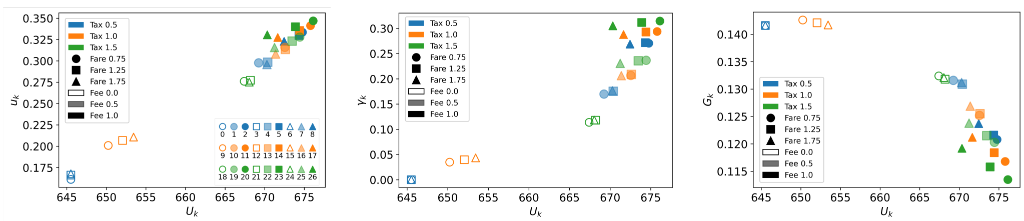


Figure 5: Comparison of u_k, γ_k, and G_k against U_k for different policy configurations in Houston.

Figure 5 presents the model-based evaluation of the 27 policy proposals for Houston. Notably, *Policy 20* emerges as the clear Pareto-optimal choice—it lies at the frontier in all three metrics considered: total utility (U_k), minimum utility (u_k), and the Gini index (G_k). Compared to the common winner in Chicago (*Policy 10*), *Policy 20* maintains the same transit fare but sets both the sales tax and driver fee at their highest levels (1.5% and $1, respectively). This result is consistent with the model's logic. Given Houston's low transit share and poor service levels, there is significant potential for improvement. Raising additional revenue through broader taxation and congestion-based fees—while keeping fares affordable—is the most effective way to boost transit access and system efficiency.

In Houston, GPT-4o agents consistently select Policy 2 as the winning option in all ten rounds. The only divergence from the model's optimal Policy 20 is the tax rate: while the model favors the maximum (1%), GPT-4o agents prefer the minimum (0.5%). Comparing *Policy 2* (Houston) with *Policy 10* (Chicago) suggests Houston agents are less supportive of taxation and more accepting of driver fees. Entropy values in Table 3 further show higher dispersion in Houston, indicating more diverse preferences among agents.

While preliminary, these findings offer intriguing clues about the LLM's capacity to reflect contextual variation. The divergence in voting patterns may signal an implicit sensitivity to the political and cultural ethos of different urban environments—an area that warrants further investigation.

5 Discussion

This study investigates how large language models (LLMs), operating as autonomous agents in a simulated citywide referendum, compare to conventional analytical models in shaping urban transportation policy. By embedding agents within a realistic decision framework and prompting them with localized context and performance metrics, we evaluate their collective behavior across voting methods, model types, and urban environments.

Insights include: First, LLM agents selected policies that, while often differing in detail, reflect broadly similar priorities as the model-based Pareto-optimal solutions. Second, GPT-4o produced more consistent and concentrated voting patterns than Claude-3.5-Sonnet, and generally reason with positive sentiments. GPT-4o results are stable across different voting methods. Lastly, while sensitive to contexts and information, GPT-4o agents are generally aware of the cultural variation across urban settings, proving generalizability.

The observed variation in LLM behavior across urban contexts and model types bring up key questions for future research. First, divergence in GPT-4o and Claude-3.5-Sonnet model outcomes deserves systematic scrutiny on model alignment and interpretability. Second, verification of the reasoning output to survey or ethnographic data is crucial to assess how faithfully models simulate actual community preferences. Lastly, prompt structure and persona design could be explored to enhance validity of the results.

At their core, LLMs are not just decision aids or behavioral simulators—they are also repositories of embedded social priors. As such models increasingly participate in policy deliberation, understanding—and where necessary, correcting—these priors will be essential to safeguarding democratic legitimacy, procedural fairness, and value alignment in AI-assisted public planning.

References

[1] Anthropic (2024). Introducing Claude 3.5 Sonnet. https://www.anthropic.com/news/claude-3-5-sonnet. Accessed on July 23, 2025.

[2] Arrow, K. J. (1966). Exposition of the theory of choice under uncertainty. *Synthese*, 16:253–269.

[3] Beckmann, M., McGuire, C. B., and Winsten, C. B. (1956). *Studies in the Economics of Transportation*. Yale University Press.

[4] Boyce, D. E. and Williams, H. C. (2015). *Forecasting urban travel: Past, present and future*. Edward Elgar Publishing.

[5] CensusReporter (2023). Houston, tx.

[6] Chen, G., Fan, L., Gong, Z., Xie, N., Li, Z., Liu, Z., Li, C., Qu, Q., Ni, S., and Yang, M. (2024). Agentcourt: Simulating court with adversarial evolvable lawyer agents. *arXiv preprint arXiv:2408.08089*.

[7] Chen, J., Yuan, S., Ye, R., Majumder, B. P., and Richardson, K. (2023). Put your money where your mouth is: Evaluating strategic planning and execution of llm agents in an auction arena. *arXiv preprint arXiv:2310.05746*.

[8] Comeaux, D. (2021). A pre-pandemic snapshot of travel in northeastern illinois key findings. *Chicago Metropolitan Agency for Planning*.

[9] Dai, T., Zheng, H., and Nie, M. (2024). Is fare free transit just? quantifying the impact of moral principles on transit design and finance. *Quantifying the Impact of Moral Principles on Transit Design and Finance (May 24, 2024)*.

[10] FHWA (2019). Urban congestion report.

[11] Flyvbjerg, B., Bruzelius, N., and Rothengatter, W. (2003). *Megaprojects and Risk: An Anatomy of Ambition*. Cambridge University Press.

[12] Forester, J. (1989). *Planning in the Face of Power*. University of California Press.

[13] Han, S. (2024). Mining causality: Ai-assisted search for instrumental variables. *arXiv preprint arXiv:2409.14202*.

[14] Hartgen, D. T. (2013). Hubris or humility? accuracy issues for the next 50 years of travel demand modeling. 40(6):1133–1157.

[15] Hutto, C. and Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 216–225.

[16] Kahneman, D. (1979). Prospect theory: An analysis of decisions under risk. *Econometrica*, 47:278.

[17] Li, N., Gao, C., Li, M., Li, Y., and Liao, Q. (2023). Econagent: large language model-empowered agents for simulating macroeconomic activities. *arXiv preprint arXiv:2310.10436*.

[18] Manheim, M. L. (1979). *Fundamentals of Transportation systems analysis; Volume 1: Basic concepts*.

[19] Martens, K. (2017). *Transport justice: Designing fair transportation systems*. Routledge.

[20] McFadden, D. (1973). Conditional logit analysis of qualitative choice behavior.

[21] Meyer, M. D. and Miller, E. J. (2001). *Urban transportation planning: A decision-oriented approach*.

[22] MITEDSL (2023). *Instant Runoff Voting*. MIT Election Data and Science Lab. Online explainer, last updated April 25 2023.

[23] Ni, H., Wang, Y., and Liu, H. (2024). Planning, living and judging: A multi-agent llm-based framework for cyclical urban planning. *arXiv preprint arXiv:2412.20505*.

[24] NTD (2020). Transit profiles: 2019 top 50 reporters.

[25] OpenAI (2025). Chatgpt response to user query. https://chat.openai.com/. Accessed: 2025-06-29.

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

[27] Park, J. S., Zou, C. Q., Shaw, A., Hill, B. M., Cai, C., Morris, M. R., Willer, R., Liang, P., and Bernstein, M. S. (2024). Generative agent simulations of 1,000 people. *arXiv preprint arXiv:2411.10109*.

[28] Polak, J. (1987). A comment on supernak's critique of transport modelling. 14(1).

[29] Ross, J., Kim, Y., and Lo, A. W. (2024). Llm economicus? mapping the behavioral biases of llms via utility theory. *arXiv preprint arXiv:2408.02784*.

[30] Simon, H. A. (1990). Bounded rationality. In Eatwell, J., Milgate, M., and Newman, P., editors, *Utility and Probability*, pages 15–18. Palgrave Macmilla.

[31] Wachs, M. (1982). Ethical dilemmas in forecasting for public policy. *Public Administration Review*, 42(6):562–567.

[32] Wachs, M. (1989). When planners lie with numbers. *Journal of the American Planning Association*, 55(4):476–479.

[33] Weiner, E. (1997). Urban transportation planning in the united states: an historical overview. *Transportation Research Board*.

[34] Williams, H. and Ortuzar, J. D. D. (1982). Travel demand and response analysis—some integrating themes. 16(5):345–362.

[35] Yang, J. C., Dalisan, D., Korecki, M., Hausladen, C. I., and Helbing, D. (2024). Llm voting: Human choices and ai collective decision-making. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 1696–1708.

[36] Zhou, Y., Muresanu, A. I., Han, Z., Paster, K., Pitis, S., Chan, H., and Ba, J. (2022). Large language models are human-level prompt engineers. In *The Eleventh International Conference on Learning Representations*.

[37] Zhou, Z., Lin, Y., Jin, D., and Li, Y. (2024). Large language model for participatory urban planning. *arXiv preprint arXiv:2402.17161*.

A Evaluation metrics

A.1 Winner of referendum and policy means of selected policies

In the simulation, voting is carried out in multiple rounds to ensure the results are stable. Accordingly, we let $\boldsymbol{v}_i^j = \{v_{i,k}^j\}$ be the vote of agent $i \in \mathbb{I}$ in round j. In each round, we can determine the winner and compute the mean values of the polices voted for in a voting method. The policy means can then be averaged over the rounds.

1. **5-Approval voting**: The winner in each round is the policy that receives the most votes, i.e., $k^* = \operatorname{argmax}_k \sum_{i \in \mathbb{I}} v_{i,k}$. We can compute the mean values of all polices that receives at least one vote as

$$\bar{\boldsymbol{p}}_A^j = \frac{\sum_{k \in \mathbb{K}} \sum_{i \in \mathbb{I}} \boldsymbol{p}_k v_{i,k}^j}{I}.$$

2. **All-approval voting**: Both the determination of the winner and the calculation of the mean policy values are identical to Top-5 approval voting.

3. **Ranked-choice voting**: In ranked-choice voting, the winner is determined using the instant-runoff process [22], where voters rank 5 policies in order of preference, and the policy with the fewest rank 1 votes is eliminated in successive rounds until one policy has a majority.

   Assuming all agents cast complete ranked ballots of five proposals in referendum round j, let $\boldsymbol{p}^j_{i,s}$ denote the value vector associated with the policy ranked at position $s \in \{1, 2, 3, 4, 5\}$ by agent $i \in \mathcal{I}$. The mean policy value at rank s in round j is then defined as:

   $$\bar{\boldsymbol{p}}^j_s = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \boldsymbol{p}^j_{i,s}.$$

   This yields a ranked list of average policy vectors $\{\bar{\boldsymbol{p}}^j_1, \bar{\boldsymbol{p}}^j_2, \ldots, \bar{\boldsymbol{p}}^j_5\}$, which summarizes aggregate preferences across agents at each rank level.

A.2 Entropy

We measure cross-agent variation in round j via Shannon entropy. Let P^j_k be the probability of policy k being selected. For 5- or All-Approval Voting,

$$P^j_k = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} v^j_{i,k},$$

and for Ranked-Choice Voting,

$$P^j_k = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \mathbb{I}(\text{policy } k \in \{k^j_{i,1}, \ldots, k^j_{i,5}\}).$$

Policy entropy is

$$E^j = -\sum_k P^j_k \log_2 P^j_k.$$

For levers, let $P^j_{x,y}$ be the probability of choosing lever $x \in \{r, t, \tau\}$ at level $y \in \{l, m, h\}$. Lever entropy is

$$e^j_x = -\sum_y P^j_{x,y} \log_2 P^j_{x,y},$$

and for ranked votes, conditional on rank s,

$$e^j_{x|s} = -\sum_y P^j_{x,y|s} \log_2 P^j_{x,y|s}.$$

A.3 VADER sentimental analysis

We apply VADER (Valence Aware Dictionary and sEntiment Reasoner) to measure the sentiment of LLM-generated rationale texts [15]. VADER assigns a compound score $S_i \in [-1, 1]$ for each agent i, computed as

$$S_i = \frac{o_i - n_i}{\sqrt{(o_i - n_i)^2 + \alpha}} = \text{VADER}(T_i), \alpha = 15,$$

where o_i and n_i are the aggregated positive and negative polarity scores assigned by the VADER lexicon. α is a normalized constant. We analyze S_i across agents to examine sentiment trends and the strength of preferences in policy deliberations.

B Available policy package sets

Table 4: Policy and their model-based performance metrics for Chicago.

ID	tax rate	transit fare	driver fee	drive_time	bus_time	drive_cost	bus_cost	Transit%	Utotal	Umin	Gini
0	0.5	0.75	0	25.5	68.72	5.43	0.75	12.51	865.0972	0.2093	0.1544
1	0.5	0.75	0.5	19.11	57.43	5.93	0.75	39.28	977.3733	0.3212	0.1086
2	0.5	0.75	1	18.02	54.64	6.43	0.75	50.50	973.0646	0.3386	0.0957
3	0.5	1.25	0	23.16	64.12	5.43	1.25	20.98	911.2723	0.2552	0.1365
4	0.5	1.25	0.5	19.01	54.51	5.93	1.25	40.49	978.7941	0.3153	0.1096
5	0.5	1.25	1	17.97	51.99	6.43	1.25	50.98	971.9173	0.3287	0.0991
6	0.5	1.75	0	22.40	60.33	5.43	1.75	23.71	926.1360	0.2617	0.1326
7	0.5	1.75	0.5	19.06	52.48	5.93	1.75	39.89	971.5802	0.3019	0.1149
8	0.5	1.75	1	18.08	50.56	6.43	1.75	50.00	973.1771	0.3111	0.1071
9	1.0	0.75	0	21.72	62.79	5.43	0.75	26.14	934.5490	0.2827	0.1257
10	1.0	0.75	0.5	18.80	55.79	5.93	0.75	42.79	982.4324	0.3287	0.1027
11	1.0	0.75	1	17.80	53.32	6.43	0.75	52.44	978.2246	0.3425	0.0917
12	**1.0**	**1.25**	**0**	**21.00**	**58.80**	**5.43**	**1.25**	**28.72**	**950.1400**	**0.2884**	**0.1219**
13	1.0	1.25	0.5	18.76	53.21	5.93	1.25	43.31	980.2506	0.3196	0.1056
14	1.0	1.25	1	17.79	51.00	6.43	1.25	52.56	973.3634	0.3304	0.0964
15	1.0	1.75	0	20.98	56.44	5.43	1.75	28.80	948.0086	0.2799	0.1244
16	1.0	1.75	0.5	18.86	51.47	5.93	1.75	42.21	970.1432	0.3039	0.1123
17	1.0	1.75	1	18.08	51.07	6.43	1.75	50.00	961.5373	0.3061	0.1089
18	1.5	0.75	0	20.07	58.51	5.43	0.75	32.02	965.1986	0.3074	0.1142
19	1.5	0.75	0.5	18.52	54.37	5.93	0.75	45.82	985.7996	0.3340	0.0981
20	1.5	0.75	1	17.58	52.16	6.43	0.75	54.21	981.6979	0.3454	0.0883
21	1.5	1.25	0	19.81	55.50	5.43	1.25	32.94	969.0859	0.3030	0.1150
22	1.5	1.25	0.5	18.52	52.10	5.93	1.25	45.79	980.4515	0.3224	0.1025
23	1.5	1.25	1	17.61	50.14	6.43	1.25	54.02	973.6848	0.3315	0.0942
24	1.5	1.75	0	20.07	53.91	5.43	1.75	32.03	959.1524	0.2888	0.1201
25	1.5	1.75	0.5	18.66	50.61	5.93	1.75	44.28	967.8360	0.3050	0.1102
26	1.5	1.75	1	17.74	48.82	6.43	1.75	52.92	957.9162	0.3126	0.1029

Table 5: Policy and their model-based performance metrics for Houston.

ID	tax rate	transit fare	driver fee	drive_time	transit time	drive cost	transit cost	Transit %	Utotal	Umin	Gini
0	0.5	0.75	0	25.98	61.55	7.19	0.75	0.00	645.5176	0.1611	0.1416
1	0.5	0.75	0.5	23.47	58.58	7.69	0.75	16.99	669.2392	0.2977	0.1316
2	0.5	0.75	1	22.62	56.06	8.19	0.75	27.06	674.7613	0.3336	0.1208
3	0.5	1.25	0	25.98	61.25	7.19	1.25	0.00	645.5176	0.1662	0.1416
4	0.5	1.25	0.5	23.42	57.54	7.69	1.25	17.53	670.3823	0.2985	0.1309
5	0.5	1.25	1	22.61	55.05	8.19	1.25	27.17	674.3346	0.3297	0.1216
6	0.5	1.75	0	25.98	61.01	7.19	1.75	0.00	645.5176	0.1677	0.1416
7	0.5	1.75	0.5	23.41	56.62	7.69	1.75	17.65	670.2816	0.2958	0.1312
8	0.5	1.75	1	22.64	54.23	8.19	1.75	26.88	672.4483	0.3229	0.1237
9	1.0	0.75	0	25.12	61.23	7.19	0.75	3.48	650.2375	0.2009	0.1426
10	1.0	0.75	0.5	23.17	57.20	7.69	0.75	20.67	672.5462	0.3159	0.1252
11	1.0	0.75	1	22.40	55.07	8.19	0.75	29.39	675.7829	0.3413	0.1168
12	**1.0**	**1.25**	**0**	**25.00**	**60.77**	**7.19**	**1.25**	**3.96**	**652.0485**	**0.2068**	**0.1421**
13	1.0	1.25	0.5	23.16	56.22	7.69	1.25	20.87	672.6134	0.3135	0.1255
14	1.0	1.25	1	22.41	54.17	8.19	1.25	29.27	674.4099	0.3356	0.1184
15	1.0	1.75	0	24.91	60.25	7.19	1.75	4.33	653.4359	0.2107	0.1417
16	1.0	1.75	0.5	23.18	55.39	7.69	1.75	20.64	671.3721	0.3080	0.1269
17	1.0	1.75	1	22.46	53.45	8.19	1.75	28.78	671.6410	0.3274	0.1212
18	1.5	0.75	0	23.88	58.93	7.19	0.75	11.37	667.4096	0.2761	0.1324
19	1.5	0.75	0.5	22.92	56.03	7.69	0.75	23.68	674.4277	0.3280	0.1203
20	1.5	0.75	1	22.19	54.20	8.19	0.75	31.47	676.1273	0.3470	0.1135
21	1.5	1.25	0	23.85	58.10	7.19	1.25	11.80	668.2043	0.2773	0.1318
22	1.5	1.25	0.5	22.93	55.14	7.69	1.25	23.58	673.4684	0.3232	0.1215
23	1.5	1.25	1	22.23	53.40	8.19	1.25	31.16	673.8875	0.3400	0.1158
24	1.5	1.75	0	23.84	57.33	7.19	1.75	11.83	668.0523	0.2751	0.1321
25	1.5	1.75	0.5	22.97	54.41	7.69	1.75	23.08	671.2299	0.3157	0.1238
26	1.5	1.75	1	22.29	52.77	8.19	1.75	30.50	670.3263	0.3306	0.1192

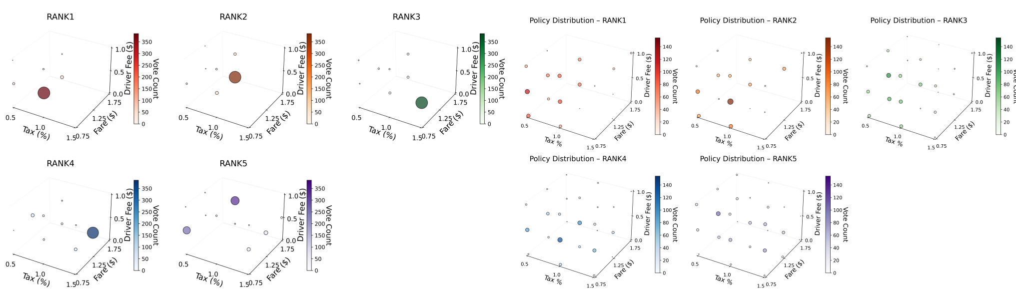
C Transit design model calibration

Table 6: Key statistics produced by the calibrated model vs. empirical data.

	Chicago		Houston	
	Data	Model	Data	Model
Transit mode share	30% [8]	29%	4% [5]	4%
Daily ridership	1.47 million [24]	1.46 million	0.29 million [24]	0.30 million
Farebox	43% [24]	44%	12% [24]	11%
Peak hour budget	$441/km^2 [24]	$441/km^2	$125/km^2 [24]	$125/km^2
Congestion index	0.77 [10]	0.74	0.79 [10]	0.77

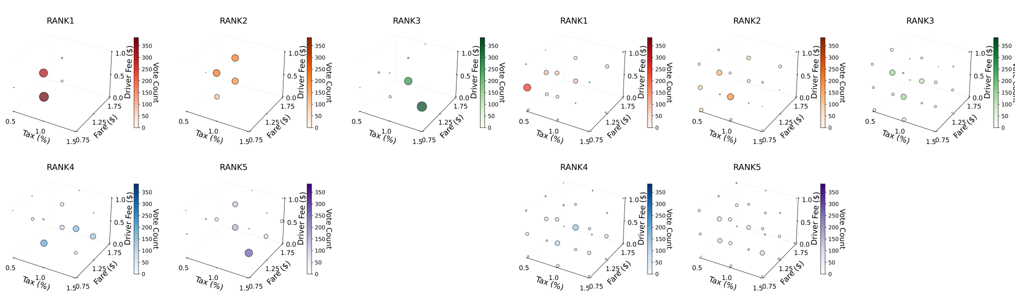
D Voting distribution

Figure 6 compares the full Ranked-Choice voting distributions generated in the referendums using GPT-4o and Claude-3.5-Sonnet. Each plot contains 27 lattice points representing the policy alternatives, with the size of each point proportional to the number of votes the policy received at a given rank across ten simulation rounds.



(a) GPT-4o. (b) Claude-3.5-Sonnet.

Figure 6: Ranked-Choice voting outcomes generated by GPT-4o and Claude-3.5-Sonnet. Scenario CHI-com, ten rounds.



(a) CHI-know GPT-4o. (b) CHI-know Claude-3.5-Sonnet.

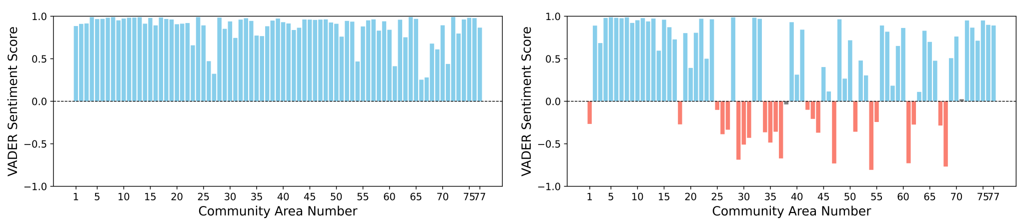
Figure 7: Ranked-Choice voting outcomes generated by GPT-4o and Claude-3.5-Sonnet. Scenario *CHI-know*, ten rounds.

For CHI-com, the plots confirm that GPT-4o generates a much more concentrated voting distribution than Claude-3.5-Sonnet, as evidenced by the larger and more clearly defined clusters in Figure 6(a). Despite the greater dispersion in Claude-3.5-Sonnet's rankings, the average policy values associated with rank-1 votes remain similar across the two models, though Claude-3.5-Sonnet shows a slight tilt toward policies with higher transit fares (see Table 2).

As shown in Figure 7, providing additional contextual information disperses GPT-4o's voting patterns, leading to consistently higher entropy values across the board (see Table 2). Interestingly, the effect is reversed for Claude-3.5-Sonnet. While less pronounced, the evidence—from the shift in winning policies (from a 4:6 split to a 2:8 split) and from the entropy values in Table 2—indicates that Claude-3.5-Sonnet's voting patterns become more concentrated once contextual information is introduced.

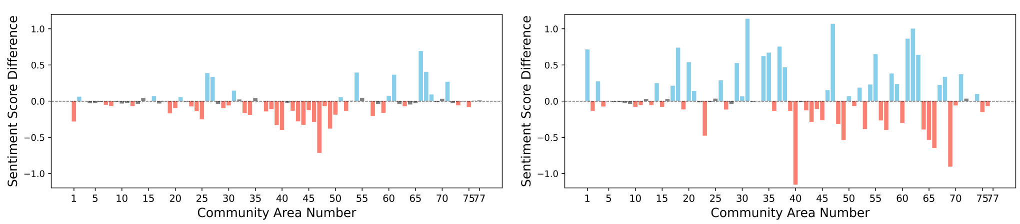
E Sentiment analysis

To better understand the divergent voting behaviors exhibited by the two LLMs—despite their broadly similar aggregate preferences—we turn to the sentiments expressed in the agents' rationale texts. These rationales, generated as part of the structured response for each agent, provide insight into the tone and emotional framing of their decision-making. We process these texts using the VADER sentiment analyzer, which produces a compound score ranging from -1 (highly negative) to 1 (highly positive) for each community.



(a) GPT-4o. (b) Claude-3.5-Sonnet.

Figure 8: Sentiment scores of rationale texts for *CHI-com*. GPT-4o vs. Claude-3.5-Sonnet across ten rounds. The scores range from -1 to 1. Each column corresponds to a distinct community area in Chicago, ordered by their official community area index used by the City of Chicago.



(a) GPT-4o. (b) Claude-3.5-Sonnet.

Figure 9: Differences in the sentiment scores of rationale texts between *CHI-know* (with contextual information) and *CHI-com* (without). GPT-4o vs. Claude-3.5-Sonnet agents across ten rounds.

As shown in Figure 8, GPT-4o agents exhibit consistently positive sentiment across all communities. In contrast, Claude-3.5-Sonnet produces a much wider dispersion: while some agents are equally upbeat, a notable share—nearly one-third—express neutral or negative sentiment in their reasoning. This discrepancy is striking and may offer a partial explanation for the higher entropy and broader vote distribution observed with Claude-3.5-Sonnet.

In Figure 9, we compare the differences in sentiment scores between the CHI-know and CHI-com scenarios. The results highlight distinct ways in which the two models respond to contextual information. For GPT-4o, the information tends to moderate overly optimistic tones, pushing sentiment scores downward. While corrections are frequent, the overall variation remains within a relatively narrow range. GPT-4o thus appears to internalize contextual information in a cautious manner, limiting drastic changes across community areas. In contrast, Claude-3.5-Sonnet exhibits much stronger sensitivity to contextual inputs. The sentiment differences are both larger in magnitude and more polarized, with sharp positive spikes in some community areas and pronounced negative shifts in others.

While we cannot assert a direct causal link between sentiment and vote outcome, it is reasonable to hypothesize that sentiment affects how agents weigh trade-offs and interpret community needs. For example, a more pessimistic framing might lead agents to avoid policies with perceived downside risks, whereas a more optimistic tone could favor ambitious, redistributive strategies.

The more uniformly positive tone of GPT-4o raises further questions. Might this reflect a bias introduced during pretraining—perhaps an overexposure to institutional or promotional language that emphasizes uplift and resolution? Or could it indicate an under-representation of narratives from marginalized communities, which may affect how the model perceives hardship that some of the Chicago communities must have been experiencing? These questions, though beyond the scope of the present study, point to a broader concern: sentiment in LLM-generated reasoning may not merely reflect mood but encode deeper assumptions about the world. As such, it deserves closer scrutiny in future research on model alignment and fairness.

F Regression model selection and summary statistics

We begin with a full set of available sociodemographic variables (e.g. income and race) and travel attributes (e.g., car ownership and access to transit). Using the voting results simulated by GPT-4o, we iteratively test the explanatory power of each variable. A variable is retained if it (i) exhibits statistical significance according to standard t-tests and F-tests and (ii) is not strongly correlated with one another. We also attempt to balance across different categories so that the selection includes at least one variable related to either race, car ownership, travel mode, or income. As the final step, we compute the Variance Inflation Factors (VIFs) of the included variables to rule out the concern for multicollinearity. Table 7 reports descriptive statistics for the selected variables. Since all variables are measured in percentages, they are all represented by a real number between 0 and 1 in regression to avoid the bias from scale differences.

Table 7: Summary statistics of sociodemographic and travel variables.

Variable	**Description**	**Min**	**Max**	**Mean**	**Std**
none-white	Percentage of residents of non-White race	16.00%	99.80%	72.77%	25.80%
number of cars	Percentage of households without cars	3.70%	58.00%	24.12%	12.92%
transit commuter %	Percentage of commuters traveling by public transit	4.20%	36.90%	19.50%	7.53%
income $<$ 25k	Percentage of households with annual income below $25k	6.10%	70.00%	23.19%	12.26%
income $>$ 150k	Percentage of households with annual income above $150k	0.60%	49.50%	17.21%	12.26%

G Sample prompt

System Prompt

Initial Context:

You are a representative from **<community area>,** one of the seventy-seven communities in the City of Chicago. On behalf your community, you will be participating in a referendum in which representatives from all communities vote on a set of transportation policy proposals. **<Instruction of Voting method:** e.g. You are allowed to choose up to five proposals and submit your vote as a ranked list of proposals.**>** Remember you must act in the best interests of your community.

Output Format:

Think step by step and submit the top five policy proposals in a descending order of preference. Return a JSON object:
{"community_area": "<name>",
"thinking": {
"disposable_income": "<summary of the disposable income of the community area>",
"discretionary_consumption": "<summary of the discretionary consumption of the community area>",
"accessibility": "<summary of the accessibility to resources and services by different modes of transportation in the community area>",
"decision_rationale": "<rationale of the ranked voting decision, showing factores that influence the tradeoffs and ranking rationale, think in step by step>"
"vote": [<rank1>, <rank2>, <rank3>,<rank4>, <rank5>]}
* The vote list must contain 5 distinct integers from 0 to 26. * No additional keys, no Markdown, no code fences.
}

User Prompt

Referendum Setting:

In the referendum, residents from all seventy-seven communities will vote on 27 transportation policy proposals.
A policy proposal consists of three policies: (i) transit fare policy, which may set a per-trip fare for riding transit to either $0.75, $1.25, or $1.75,
(ii) tax policy, which may set a dedicated sales tax rate to either 0.5%, 1%, or 1.5%;
and (iii) a driver-fee policy, which may set a per-trip fee for driving to either $0.00, $0.50, or $1.00.
Since there are three options for each policy, you have 27 proposals to vote on.
You must pick the top five proposals and rank them from 1 to 5 according to how they serve your interest. Your interest should be defined by weighing the cost and benefits of each package.

The Chicago Metropolitan Agency for Planning (CMAP) has estimated the travel times per trip by transit and driving corresponding to each policy, along with the corresponding fare and fees.
So, you should use this information to gauge your interest. You should also be aware that paying either fare for transit or a fee for driving will take a portion of your income away from other consumption (such as food, housing, cloth, vacation, tuition etc.).
Hence, you should carefully assess how such a loss of income might affect your life. In doing so, please bear in mind that you are acting as a representative resident in your community – so you should try to use as much data (census, employment, transit availability, car ownership, etc.) from your community to make your decision.

Here is some other information that might help your deliberations:
1. The dedicated sales tax is meant to support transit services. The higher the tax, the more money the transit authority will have to provide the transit service. However, remember paying the sales tax means you will have less money to consume other goods and services. If you consume $20000 per year, for example, a 1% sales tax will cost you $200 a year.
2. If you ride in transit, you will pay fare; if you drive, you will pay the driver fee. All revenues from fare and driver fee will be used to support transit, which means higher fare and driver fee generally mean better transit services. However, higher fare and driver fees also mean you have less money to spend on other goods and services. Also, remember the fare and fee are paid on the per trip basis.
3. Driver fees and better transit might persuade some drivers to switch from driving to transit. This could help reduce congestion, lowering driving time and reducing emissions.

Policy Options & Transportation Information

Current Transportation Policy and the travel time and cost:
the transit fare is $1.25/ride, the sales tax is 1%, and the driving fee is $0.00/trip.
The travel time by car is 21 minutes, and the travel time by transit is 58.8 minutes, the cost per trip by car is $5.43, and the cost per trip by transit is $1.25.

Candidate Policy Options and the associated travel time and cost estimated by Chicago Metropolitan Agency for Planning:
{policy_options}

Chain-of-Thought

Consider the following factors based on the unique characteristics of the community area:
1. The disposable income of the community area
2. The discretionary consumption of the community area
3. The accessibility to resources and services by different modes of transportation in the community area
Then, consider the tradeoffs and ranking rationale of the ranked voting decision.

Voting Prompt

Evaluate these 27 policy combinations for the {community} community area.
Return top 5 ranked policies (0-26) and reasoning as a JSON object.
"""

Figure 10: Example Prompt.

H Sample LLM response in CHI-com scenario

H.1 Sample response of All-Approval by GPT-4o

Community Area: Loop
Chain-of-Thought:

1. Disposable Income: “The Loop is a central business district with a mix of high-income professionals and lower-income service workers. Disposable income varies significantly, but there is a substantial portion of residents with higher disposable income."
2. Discretionary Consumption: “Residents in the Loop have a high level of discretionary consumption due to the presence of numerous dining, entertainment, and shopping options. However, the cost of living is also high, which can impact discretionary spending."
3. Accessibility: “The Loop has excellent accessibility to resources and services, with extensive public transit options including buses and trains. Car ownership is less common due to the availability of public transit and the high cost of parking."
4. Decision Rationale: “Given the Loop’s strong public transit infrastructure and the presence of high-income residents, policies that enhance transit services while maintaining reasonable costs are preferred. A lower transit fare and moderate driver fees can encourage transit use and reduce congestion. A moderate sales tax is acceptable if it significantly improves transit services, as many residents can absorb the cost. Therefore, policies with a $0.75 fare and a $0.5 driver fee, combined with a 1% or 1.5% sales tax, are prioritized to balance cost and service improvement."

Vote Decision: [10, 11, 19, 20]

H.2 Sample response of 5-Approval by GPT-4o

Community Area: Loop
Chain-of-Thought:

1. Disposable Income: “The Loop is a central business district with a mix of high-income professionals and lower-income service workers. Disposable income varies significantly, but there is a substantial portion of residents with higher disposable income."
2. Discretionary Consumption: “Residents in the Loop have a higher level of discretionary consumption due to the presence of numerous dining, entertainment, and shopping options. However, the cost of living is also high, which affects overall discretionary spending."
3. Accessibility: “The Loop has excellent accessibility to resources and services, with extensive public transit options including buses and trains. Car ownership is lower compared to other areas due to the availability of public transit and the high cost of parking."
4. Decision Rationale: “Given the Loop’s strong public transit infrastructure and the mixed income levels, policies that enhance transit services without significantly increasing costs for lower-income residents are preferred. A lower transit fare is beneficial for encouraging public transit use, while a moderate driver fee can help reduce congestion and emissions. A moderate sales tax is acceptable if it significantly improves transit services. Therefore, policies with lower fares and moderate driver fees are prioritized, while ensuring the sales tax does not overly burden residents."

Vote Decision: [1, 4, 10, 13, 19].

H.3 Sample response of ranked-choice voting by GPT-4o

Community Area: Loop
Chain-of-Thought:

1. Disposable Income: “The Loop is a central business district with a mix of high-income professionals and lower-income service workers. Disposable income varies significantly, but overall, the area has a relatively high average income."

2. Discretionary Consumption: “Residents in the Loop have a high level of discretionary consumption due to the presence of numerous dining, shopping, and entertainment options. However, the cost of living is also high, which affects disposable income."
3. Accessibility: “The Loop has excellent accessibility to resources and services, with extensive public transit options including buses and trains. Walking and biking are also viable options due to the dense urban environment."
4. Decision Rationale: “Given the high accessibility to public transit and the mixed income levels, the Loop community would benefit from policies that enhance transit services while keeping costs manageable. Lower transit fares and moderate driver fees can encourage more transit use, reducing congestion and emissions. A moderate sales tax is acceptable if it significantly improves transit services. The focus is on balancing cost with improved transit efficiency and reduced travel times."

Voting Decision: [11, 14, 13, 10, 19]

I Limitations

Our evaluation is limited by the absence of direct human ground truth. We compare LLM-simulated community preferences against outcomes from a calibrated transportation model, rather than against empirically observed public opinion, stakeholder input, or stated-preference data from real residents. This gap reflects a practical challenge: for hypothetical policy scenarios (e.g., introducing a new driver fee or restructuring fare policy), it is difficult to obtain high-resolution, representative human responses at city scale, especially across dozens of neighborhoods. As a result, the best we can do is alignment with a stylized model of welfare.

Another limitation of the current study is setting GPT-4o and Claude-3.5-Sonnet as static “agents,” sampled deterministically. This preserves tractability and interpretability of the model outcomes but it suppresses several dynamics that matter in real decision-making. In particular: (1) different models have different alignment tuning and normative priors, which directly shapes their stated community preferences; (2) there is no iterative feedback loop, so models do not update their positions in response to criticism, persuasion, or evidence over time.

J Broader impacts

This work explores a new paradigm for transportation policy design. Traditional approaches rely on optimization models that encode behavioral assumptions and attempt to identify a single “best” policy under specified objectives and constraints. Our LLM-based multi-agent framework follows a similar decision-making logic, but it has the capacity to reason in natural language about trade-offs, distributional impacts, and local context. As such, it can reveal policy tensions that are difficult to capture with parametric utility-based formulations.

We view this approach as complementary rather than substitutive. In principle, model-driven performance metrics and LLM-simulated community preferences could be used together to iteratively calibrate policy choices. With appropriate integration of real public feedback, such as stated-preference surveys and stakeholder input, the system could support more equitable and more representative decision-making in practice. LLM-based simulations should be used as a supplementary tool, alongside transparent reporting, validation against empirical data, and direct engagement with the people affected.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We outline the scope and contribution of our work. Specifically, we list the contribution on the last paragraph.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: It's discussed in Section I.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: There is no theory assumptions nor proofs in this paper.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: It includes the parameter setting and the selected models with the prompts and example responses in the Appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The code will be made publicly available upon acceptance.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: It's stated in the Methodology and Results Section.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: It's stated in Table 3.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: It's stated in Section 3.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We read and followed the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The broader impact is in Section J.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to

generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: We do not release high risk models.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [NA]

Justification: We do not use existing assets.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: We do not release new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: We do not conduct crowdsourcing experiments and research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: This paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: We deploy multiple autonomous agents instantiated with GPT-4o and Claude-3.5-Sonnet to represent heterogeneous communities, generate policy rationales, and cast votes over transportation policy options. These LLM-generated preferences and justifications form the primary data analyzed in the paper. We describe the prompting framework, agent design, voting protocol, and evaluation setup in detail in Section 3.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.