LLM Voting: Human Choices and AI Collective Decision-Making

Joshua C. Yang, Damian Dailisan, Marcin Korecki, Carina I. Hausladen, Dirk Helbing

Computational Social Science, ETH Zurich, Switzerland
{joyang, ddailisan, mkorecki, carinah, dhelbing}@ethz.ch

Abstract

This paper investigates the voting behaviors of Large Language Models (LLMs), specifically GPT-4 and LLaMA-2, their biases, and how they align with human voting patterns. Our methodology involved using a dataset from a human voting experiment to establish a baseline for human preferences and conducting a corresponding experiment with LLM agents. We observed that the choice of voting methods and the presentation order influenced LLM voting outcomes. We found that varying the persona can reduce some of these biases and enhance alignment with human choices. While the Chain-of-Thought approach did not improve prediction accuracy, it has potential for AI explainability in the voting process. We also identified a trade-off between preference diversity and alignment accuracy in LLMs, influenced by different temperature settings. Our findings indicate that LLMs may lead to less diverse collective outcomes and biased assumptions when used in voting scenarios, emphasizing the need for cautious integration of LLMs into democratic processes.

Introduction

Recent breakthroughs in generative models have marked a significant achievement in Artificial Intelligence (AI): the creation of machines capable of fluently processing human language. Large Language Models (LLMs) are increasingly being integrated into a variety of services. As with many discoveries before, a great deal of enthusiasm has been associated with LLMs and their potential applications. As exciting and potentially useful LLMs may be, however, they are also limited in various ways, and their naïve integration may bring many unforeseen consequences.

LLMs have been used to reproduce classic economic, psycho-linguistic, and social psychology experiments with humans (Aher, Arriaga, and Kalai 2023; Argyle et al. 2023b; Horton 2023), as well as behavior on social media platforms (Törnberg et al. 2023). Some propose using LLMs to improve online democratic conversations (Argyle et al. 2023a). Another proposal is to use LLMs to perform human-related scientific research (Boiko, MacKnight, and Gomes 2023; Bran et al. 2023; Sourati and Evans 2023). Despite the impressive development of LLMs, however, it is clear that LLMs are currently still limited and prone to errors (Sobieszek and Price 2022; Floridi 2023). The "hallucinations" (Yao et al. 2023), proclivity for lies (Azaria and Mitchell 2023), effects such as a strong primacy effect (Wang et al. 2023b), and clear political biases (Feng et al. 2023) all question that LLMs can be safely used in social settings.

In the area of digital democracy, an application of AI that has sparked considerable debate is the concept of assisted real-time voting. The proposal to use AI "digital twins" to replace politicians has been raised (Hidalgo 2018). These AI systems, including customized LLM agents, envisioned to vote and mirror individual voter preferences (Gersbach and Martinelli 2023). Proponents believe that such agents could enable more nuanced and granular voting under voter supervision (Tang, Weyl, and the Plurality Community 2023). Yet, ethical concerns about automation, democratic integrity, and agent bias necessitate careful consideration. Allen and Weyl (2024) argue that Generative Foundation Models (GFMs), such as GPT-4, present unprecedented challenges for democratic institutions. The deployment of such technology in democratic processes would require a cautious approach, ensuring that the core values of democracy are upheld (Helbing et al. 2022).

In our study, we explore the limitations and potentials of integrating LLMs into collective decision-making processes. We focus on examining LLM behavior in voting scenarios by contrasting the voting patterns of human participants with those of LLaMA-2 and GPT-4 agents. This analysis uses data from a Participatory Budgeting (PB) voting experimental study conducted by Yang et al. (2024). Our research aims to address two research questions: (1) What biases exist in LLM voting behaviors compared to human voters, and how can these biases be mitigated to more closely align LLM decision-making with human processes? (2) Can the diversity of voting choices by LLMs be enhanced to more closely resemble the variability seen in human voters, and what are the potential trade-offs involved in achieving this level of diversity? The overall setting of the study is illustrated in Fig 1. By investigating the behavior of LLMs in this context, we seek to contribute to the broader discourse on LLM decision-making, value alignment, AI explainability, and ethical use of AI in society.

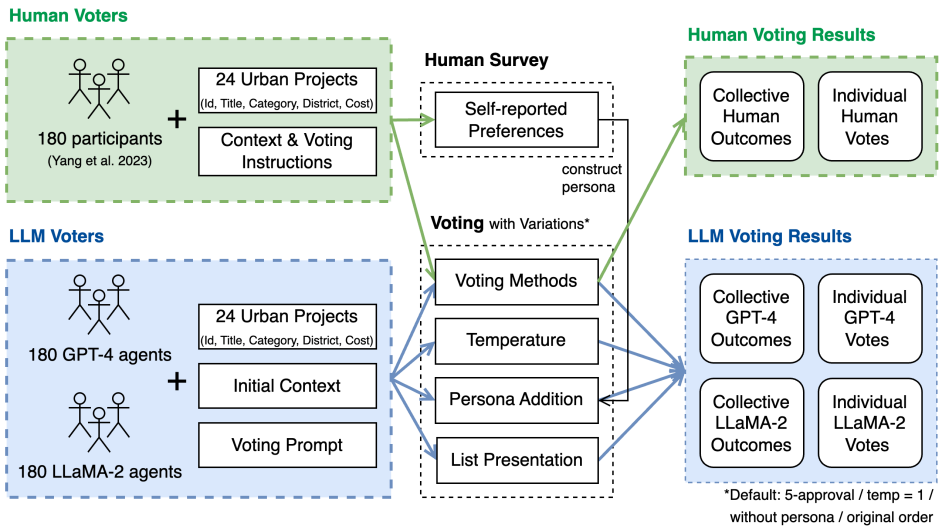

Figure 1: Overview of the LLM voting experimental setup

Related Literature

LLM Autonomous Agents

The popularity of LLMs sparked a trend wherein researchers used LLMs to model "agents" as entities with several submodules. In their landmark paper, Park et al. (2023) show LLM agents interacting with each other in a sandbox community exhibits human-like behavior. Boiko, MacKnight, and Gomes (2023) suggests combining multiple LLMs for the autonomous design, planning, and execution of scientific experiments. For a literature survey on creating autonomous LLM agents, we refer the reader to Wang et al. (2023a).

Research into multi-agent LLM systems, drawing inspiration from social and biological cooperative systems, explores the dynamics of cooperative artificial agents (Axelrod and Hamilton 1981; Shoham and Leyton-Brown 2009; Bonabeau, Dorigo, and Theraulaz 1999). Studies have investigated various LLM agents in both software and embodied forms for complex tasks, such as text evaluation (Chan et al. 2023), image captioning, and text-to-image synthesis (Zhuge et al. 2023; Talebirad and Nadiri 2023; Zhang et al. 2023b). Innovations include proactivity and adaptability enhancement (Zhang et al. 2023a), standardized prompting mechanisms (Hong et al. 2023), roleplaying frameworks for cooperation study (Li et al. 2023), and debating methods to improve answer quality (Wei et al. 2022; Hao et al. 2023; Du et al. 2023). Furthermore, researchers have explored the use of different personas within a single agent to enhance complex task performance (Wang et al. 2023c). While most studies focus on surpassing benchmarks in complex tasks, some also focus on the intrinsic characteristics of multi-agent LLM systems (Serapio-García et al. 2023; Liang et al. 2023).

LLM Societies

One particular example of a multi-agent LLM system is the creation of LLM societies. Such systems create environments for studying debates and societal simulations. Liu et al. (2023) introduces an open-source platform for simulating artificial human societies, while Hao et al. (2023) proposes a ChatLLM network system for more objective and comprehensive decision-making. Zhuge et al. (2023) explores the concept of natural language-based societies of mind, questioning their optimal governance structure. In Wang et al. (2023c), the idea of Solo Performance Prompting (SPP) is proposed, transforming a single LLM into a "cognitive synergist" through multi-turn self-collaboration with multiple personas. Du et al. (2023) presents a method where multiple language model instances propose, debate, and refine their responses to reach consensus.

Liang et al. (2023) addresses the Degeneration-of-Thought (DoT) problem with a Multi-Agent Debate framework, where agents engage in "tit for tat" arguments and a judge oversees the debate to derive a final solution. Talebirad and Nadiri (2023) introduces a collaborative environment, where intelligent agents with distinct attributes and roles communicate. Wu et al. (2023) discusses "Autogen", an open-source library facilitating conversation among multiple LLM agents. Furthermore, Ferland (2015) and Rank and Mushtare (2019) explore the use of personas in civic communication and the potential of digital twins to understand voter behavior and address biases in democratic processes.

Social Choice and Rationality

Social welfare functions are crucial for aggregating individual preferences into a collective social ranking, denoted by R, derived from individual rankings $\{R_i\}$. They help establish a coherent social order by balancing fairness and rational decision-making (Sen 1995; Arrow 1951).

Rationality in social choice theory impacts voting behaviors and preference aggregation. It encompasses: **Internal Consistency of Choice**: Fundamental in decision theory and

economics, this aspect requires individual choices to be coherent and logically consistent (Arrow 1951). It is tested in our research by varying voting methods, which, as prior studies indicate, impact voter preferences and behaviors significantly (Yang et al. 2024; Hausladen et al. 2023). **Self-Interest Maximization**: Central to classical economic theory, this assumes individuals maximize their self-interest, often modeled as a utility function (Becker 1976; Walsh 1994), guiding choices to optimize personal utility. **Maximization in General**: Extending rationality, this principle involves maximizing a defined objective function (Sen 1995), often aiming to maximize collective utility or welfare. This aligns with the concept of other-regarding preferences, where individual utility considers the well-being of others (Blanco, Engelmann, and Normann 2011).

Our study explores these three dimensions of rational choice, analyzing their influence on individual and collective voting behaviors.

Voting in Multi-Winner Systems

Modern democracies are increasingly characterized by individualized, issue-driven actions, leading to what has been described as "chaotic pluralism" that challenges traditional democratic processes (Margetts et al. 2016). In response, multi-winner systems, such as Participatory Budgeting, allow voters to select multiple candidates from a broad field, potentially resulting in multiple winners. These systems are designed to reflect pluralistic societal values and accommodate diverse demands (Skowron, Faliszewski, and Slinko 2019; Elkind et al. 2017), aiming to maximize collective utility or fairness (Aziz and Huang 2017).

Multi-winner systems translates individual voter preferences into a collective decision, often seeking to maximize collective utility or fairness (Aziz and Huang 2017). Approval voting is one common method, where the top candidates receiving the most votes are elected (Brams and Fishburn 1983; Laslier and Van Der Straeten 2016). Given that most off-the-shelf LLMs are programmed to maintain neutrality in political decisions, such as voting between candidates, they are not used to directly vote on politicians in this study. Instead, we focus on Participatory Budgeting to examine LLM behavior in a multi-winner election format, concentrating on urban issues.

Human Biases in Voting

Human voting behavior is often influenced by various cognitive biases that can distort decision-making processes. The *primacy effect*, where voters tend to prefer candidates listed first on the ballot, is a well-documented bias (Miller and Krosnick 1998; van Erkel and Thijssen 2016). Similarly, the *recency effect* can cause voters to favor the last options they assess (Koppell and Steen 2004). These biases in human decision-making raise important considerations for the design and presentation of voting systems. In this study, we investigate whether similar biases are exhibited by LLMs and how LLMs might mimic or diverge from these complex human behaviors in voting decisions.

Opinion Sampling and Persona Creation

Our methodology, which compares actual population opinions with LLM-generated responses, is in line with the work of Durmus et al. (2023). They developed the GlobalOpinionQA dataset and a metric for comparing LLM and human responses, which is crucial to assess how well LLMs reflect human opinions. Similarly, Törnberg et al. (2023) simulated social media environments using digital personas based on demographic data. Their findings suggest that this approach leads to healthier, less divisive online discussions.

LLMs and Human Behaviour

Exploring the intricacies of political biases and personality traits in LLMs is crucial for their ethical application in human interactions. Serapio-García et al. (2023) investigate the role of personality traits in LLMs, developing methods to assess and customize them in advanced models. Strachan et al. (2024) demostrate that LLMs exhibit behavior that is consistent with the outputs of mentalistic inference in humans. Feng et al. (2023) delve into political biases of LLMs, examining their origins from diverse pretraining data and effects on tasks like hate speech detection. Meanwhile, Argyle et al. (2023b) explore how LLMs can positively influence online political discourse, with AI chat assistants improving conversation quality and receptiveness to different views. In addition, Wang et al. (2023b) examine the primacy effect in ChatGPT, finding that: i) ChatGPT's decisions are sensitive to the order of labels in prompts, and ii) there is a higher tendency for ChatGPT to select labels in earlier positions as the answer. (Abdurahman et al. 2024) also finds that AI models resemble a single WEIRD (Western, Educated, Industrialized, Rich, Democratic) individual rather than simulating a diverse pool of participants. In summary, these studies underscore the significance of recognizing personality traits in LLMs and managing biases to ensure their ethical and effective use. The study by Wei et al. (2022) demonstrates that "Chain-of-Thought"(CoT) prompting, which involves providing intermediate reasoning steps, enhances the problem-solving abilities of LLMs. In their experiments, CoT significantly improved performance on arithmetic, commonsense, and symbolic reasoning tasks. Step-Back prompting (Zheng et al. 2023), extends this CoT concept by explicitly asking LLMs to abstract high-level concepts and first principles before tackling the task. Coda-Forno et al. (2024) finds that Step-Back prompting fosters model-based behaviors in cognitive psychology experiments involving LLMs.

LLMs and Collective Intelligence

As language models (LLMs) are increasingly used to mimic human behavior, recent research has also begun exploring the collective behavior of LLMs. The work of Jarrett et al. (2023) formalizes the problem of digital representation as the simulation of an agent's behavior to yield equivalent outcomes from the collective decision-making mechanism. Chuang et al. (2024) examines whether groups of LLMs can mimic human behavior when role-playing as partisan personas (like Democrats or Republicans). The authors prompted the LLMs to role-play as different personas created with varying levels of background detail. The research

finds that incorporating CoT reasoning or a lack of detailed persona tends to diminish the wisdom of partisan crowds effect, making the group less likely to converge to more accurate beliefs.

Fish et al. (2023) merges social choice theory with the text generation abilities of LLMs. This framework is designed to facilitate complex decision-making, such as selecting textual statements that represent collective preferences. A recent work by Gudiño-Rosero, Grandi, and Hidalgo (2024) suggests that LLMs using augmented data can more accurately predict the preferences of an entire participant population compared to probabilistic samples, which may not be representative. They indicates that LLMs have significant potential for constructing systems of augmented democracy.

Methods

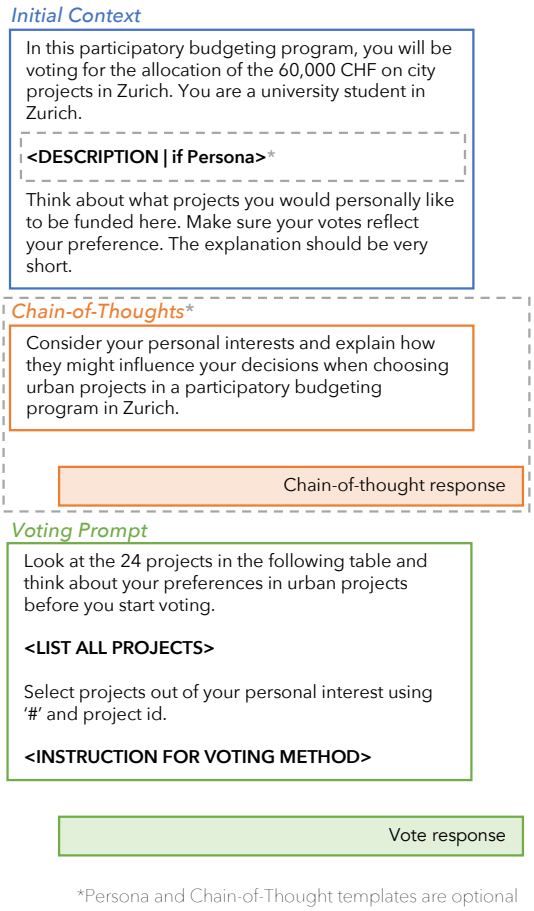

Figure 2: Overview of the LLM prompt template.

In this paper, we compare the behavior of human voters in a PB setting to that of LLM voters. For human voters, we used data from Yang et al. (2024) who simulated a Participatory Budgeting (PB) scenario in Zurich in an online experiment. The dataset[1] contains 180 university student participants' votes on urban projects, reflecting real urban interests for allocating a CHF 60,000 budget across 24 projects. We simulated the voting of human participants using LLaMA-2 and GPT-4 Turbo models. These models take in an ***Initial Context*** with instructions and background that the human voters also received in their experiements, followed by a ***Voting Prompt*** for casting votes on the projects in a format that can be processed (see Fig. 2).

Both the human participants and LLM agents were presented the same list with 24 projects (Table 3 in the appendix) in the voting process, detailed by ID, Name, District (Nord, Süd, Ost, West), Urban Category (Nature, Culture, Transportation), and Cost (CHF 5,000 or CHF 10,000). There is exactly one project for each combination of these characteristics.

Testing Different LLMs: Our study uses two types of agents: LLaMA-2 70B (Q8_0) and GPT-4 Turbo (GPT-4-1106-preview). We chose the open-source LLaMA-2 to ensure other researchers can replicate and verify our experiments, thereby enhancing the reliability of our research. In contrast, GPT-4 Turbo was selected for its widespread use and relevance in current applications, making our findings more applicable to real-world scenarios.

Aggregation: In each experimental setting, we replicated the voting process with 180 LLM voters, matching the sample size of the human experiment. LLM votes, formatted with a '#' and project ID in the textual responese, were parsed using `regex`. The Borda Count method was applied to Ranked votes to convert ranks into points, while the 10-point Cumulative votes were normalized to ensure equal weighting. Outcomes were then aggregated, ordering projects by total votes or points for a direct comparison with human voter results.

Voting Methods

We explore the four most representative voting input methods in multi-winner settings, as identified by Yang et al. (2024) in their voting experiment from which the data originate. Among these, 5-Approval voting is used as our baseline due to its common usage in PB settings and its simplicity, which facilitates the comparison of outcomes when each voter is constrained to select an equal number of projects. For each voting method, we use the same textual voting instructions that were given to human voters for the LLM agents. The voting instructions are as follows:

- **Approval**: *"Select any number of projects. Here, in this vote, you can select all the projects you approve of."*
- **5-Approval**: *"Select exactly 5 projects."*
- **Cumulative**: *"Distribute 10 points among the projects you like. List the projects and the points you allocate, separated by a colon."*
- **Ranked**: *"Select 5 projects and rank them from the most preferred to the 5th most preferred."*

Note: The conversion from ranks to points in **Ranked** voting follows the conventional Borda's Method. (1st rank: 5 points, 2nd rank: 4 points, and so forth)

[1] https://github.com/joshuay1/zurich-pb-voting

Evaluation of Voting Outcomes

In our study, we use three indices to compare voting outcomes from LLM agents and human participants:

- **Aggregated Preferences**: We employ Kendall's τ to assess the similarity between LLMs and humans: $\tau = \frac{2}{n(n-1)} \sum_{i<j} \text{sgn}(x_i - x_j)\text{sgn}(y_i - y_j)$, where n is the number of observations, and x_i, x_j, y_i, y_j represent ranked preferences.
- **Individual Vote Comparison**: We utilize the Jaccard Similarity index, defined as $J_i(A, B) = \frac{|A \cap B|}{|A \cup B|}$, where A and B are the vote sets from the i$^{\text{th}}$ LLM and human agents, respectively.
- **Preference Diversity**: The complement of Jaccard Similarity (i.e. Jaccard Distance), $D_{J_i}(A, B) = 1 - J_i(A, B)$, quantifies the diversity of preferences of voters. A higher value indicates greater heterogeneity in voting patterns.

Experiments

Variation of the Temperature Parameter: The temperature parameter t modulates the probability of choosing the next token following $p(\text{token}_i) = \frac{\exp \text{logit}_i/t}{\sum_j \exp \text{logit}_j/t}$. We systematically vary the temperature parameter within the range $t \in [0, 2]$ with 0.5 increments, noting that the standard temperature setting in ChatGPT is 1. This helps assess the impact of different randomness levels on the model's outputs.

Variation of List Presentation: Our study focuses on two list presentation variations in the context of multi-candidate elections:

1. **Primacy Effect:** This investigates how the positioning of items at the beginning of a list affects their choices.
2. **Numerical Labeling Effect:** This investigates the impact of numerical IDs on the LLM's decision-making process.

By examining both the primacy effect and the numerical labeling effect, the study aims to disentangle the influences of item ordering and numerical labeling on LLM decision-making.

Addition of Persona We construct persona descriptions from participant responses in the survey data conducted in the same study of human votes. From the survey data, we have the self-reported rating on urban category preferences (Culture, Nature, Transport), district (Nord, Süd, Ost, West), and importance of decision-making factors (district, urban category, cost). These details are incorporated into the *Initial Context* to introduce a variation of persona. By relying on self-reported preferences rather than personal demographic data, we effectively simulate diverse voter profiles for the assisted-voting scenario, thus avoiding the ethical complexities and privacy concerns associated with AI stereotyping and the use of sensitive personal information.

Each entry consists of the participant's ID, urban category preferences, district information, connection to the city, and importance ratings. The process of persona generation involves processing survey data and using the numerical scores to generate descriptive human language labels that reflect the participants' preference intensity. The function further combines this with participants' district preferences and their prioritization of decision factors like district importance, urban category significance, and cost, creating a comprehensive and personalized narrative for each participant. Here is the persona constructed based on the example in Appendix C:

> You are a university student from Nord district in Zurich. In urban topics, you have a strong preference for transport. When deciding on projects, you find the district to be moderately important, the urban category very important, and the cost of the project not important at all.

Addition of Chain-of-Thought (CoT): In the attempt to improve performance without further complicating the comparison between the results of human voters and LLM agents, we applied CoT reasoning technique to the best-performing model. In this experiment, the model used was focused on GPT-4. The agents would be prompted an additional time before the voting prompt with the project list (see Fig. 2).

Similar to the Step-Back promoting technique (Zheng et al. 2023), we specifically prompted agents to reflect on their preferences before presenting the PB projects. This additional prompting, referred to as CoT in its broad definition within this paper, aimed to mitigate the issue of "unfaithful" CoT explanations, as described by Turpin et al. (2024). By withholding project information during the initial CoT phase, we prevented the LLM agents from using pre-existing choices to rationalize their decisions.

Data and Code Availability For reproducibility, the code and data used in our experiments are available in this GitHub repository: github.com/ethz-coss/LLM_voting. For technical appendices and the **appendix** referenced in this paper, see the extended version: arxiv.org/abs/2402.01766.

Results

Number of Selected Projects

The three voter types show distinct behaviors with respect to approval and cumulative voting. As shown in figure 3, human participants display a broad spectrum of *approval* patterns, ranging from a minimum of 2 to a maximum of all 24 projects. In contrast, LLaMA's approval patterns form bell-shaped distribution, peaking around 7 projects. GPT4 demonstrates an even narrower distribution, with a median of 5 projects. In *cumulative voting*, LLaMA's point allocations frequently surpassed the prescribed maximum of 10 points, reflecting a deficiency in the model's numerical reasoning capabilities. In contrast, GPT4 adhered strictly to the voting instruction of assigning 10 points.

Consistency Across Voting Methods

Figure 4A shows the voting consistency of humans and LLMs across voting methods measured by Kendall's τ. This metric ranges from -1 for complete disagreement, 0 indicating no discernible pattern in voting, to 1 for perfect agreement across the various voting methods.

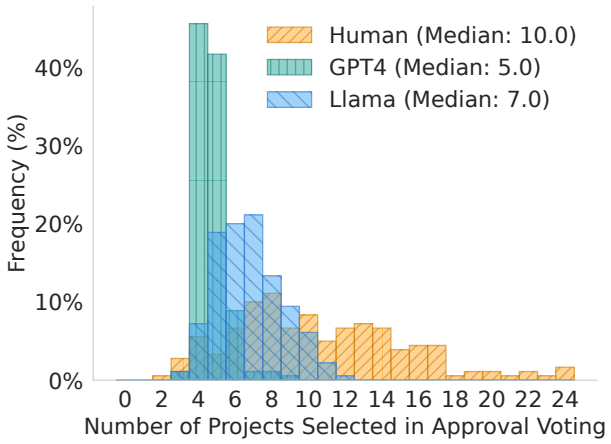

Figure 3: Histograms showing the frequency percentage of a certain number of selected projects out of 24 projects in Approval Voting between human and LLM voters.

Human voters displayed high average consistency (τ: 0.81) across different voting input methods, that is, human voters are clear with their preference and vote consistently when different voting input methods are presented. Particularly, ranked voting (rank_h) exhibited the highest consistency values for human voters, suggesting that the outcome of ranked voting is more representative of the collective human preferences.

GPT-4 also showed considerable consistency (τ: 0.71) across voting input methods, with its preferences most distinctly aligned in cumulative voting (cumu_g). LLaMA-2 had a lower average consistency score as compared to humans and GPT-4 agents (τ: 0.45). The LLaMA-2 τ value in ranked voting (rank_l) was notably lower, indicating the least alignment with other voting methods in their selection of projects

Consistency Across List Order Variations

To investigate how sensitive the LLMs are to the order of projects, we presented the 24 projects in varying sequences when prompting the LLM: original order, reversed order, and original order with reversed project IDs. Upon examining the responses individually, each agent provided a convincing rationale behind their votes, as denoted in example responses in the appendix. However, the aggregated votes reveal a real shift in the collective outcome that is not obvious in individual votes.

LLaMA-2 agents demonstrated considerable volatility in preferences due to changing order; the reversed order caused a substantial reshuffling ($\tau = -0.2$). When the IDs were reversed, LLaMA-2 agents with reversed IDs displayed more stability by keeping two of the top three projects (dark green) at the top (Fig. 4B). GPT-4 agents showed a stronger consistency when the order is reversed, but the reversed ID scenario resulted in a notable disturbance, elevating the lowest-ranked projects (dark blue). Kendall's τ coefficients at the bottom of Fig. 4B show that both LLMs experienced a significant drop with reversed orders and IDs.

These results highlight the sensitivity of LLMs to list presentation, aligning with findings where ChatGPT showed a similar sensitivity to label order (Wang et al. 2023b). However, our study did not observe the expected *Primacy Effect*, where top-listed items are typically preferred. As Table 1 in the appendix shows, there was no significant correlation between the rankings of projects and their IDs, which represented their list order, for both humans and LLMs.

Incorporating Persona into LLM Voting

The personas of the LLM agents are constructed based on human survey responses regarding the importance of urban categories and residential districts as well as the prioritization of district characteristics, urban categories, and decision-related costs.

When personas are applied via a static prompt to the LLM models, individual votes become more aligned to human votes for both, LLaMA-2 (from $J = 0.14$ to 0.21) and GPT-4 ($J = 0.18$ to 0.30).

The incorporation of personas not only influences individual sets of votes but also increases the similarity in collective project rankings. Notably, GPT-4 with persona achieves the highest τ of 0.54. Moreover, GPT-4 voters without persona adjustments ($\tau = 0.39$) still outperform LLaMA-2 voters with persona variations ($\tau = 0.12$). LLaMA-2 without persona creates an outcome ($\tau = -0.04$) that is akin to a random outcome. Qualitatively, GPT-4 without persona exhibits a pronounced bias towards transportation projects. With persona adjustments, however, GPT-4's voting distribution across various project types becomes more diverse and therefore more aligned with human outcomes (Fig. 5B).

Incorporating Chain-of-Thought into LLM Voting

We analyze whether the inclusion of CoT reasoning enhances collective decision-making in LLMs. Our findings indicate that CoT has minimal impact on both individual and collective decision-making levels. As illustrated in Figure 5A, GPT-4 and LLaMA-2 agents that underwent the CoT process demonstrate similar collective preferences to those without CoT prompting. Figure 5B reveals that for GPT-4, Jaccard Similarity values remain consistent with or without CoT. For LLaMA-2, there's even a slight decrease in similarity (0.21 to 0.19) when CoT is used. The statistical analysis reinforces these observations, showing no significant differences for either mode (GPT-4: T-statistic $= 0.0119$, P-value $= 0.9906$; LLaMA-2: T-statistic $= 1.6506$, P-value $= 0.0997$).

However, in analyzing the responses qualitatively, we find that the addition of CoT has great potential to improve the explanability of AI decision-making processes.

GPT-4 agents in 5-approval voting scenarios return simple selections of projects. This output fails to clarify the rationale behind the decisions. By contrast, incorporating CoT provides a more comprehensive thought response, with detailed reasoning of what projects the voter prefer based on their persona (Full response in extended appendix). This additional thought response offers valuable insights into the

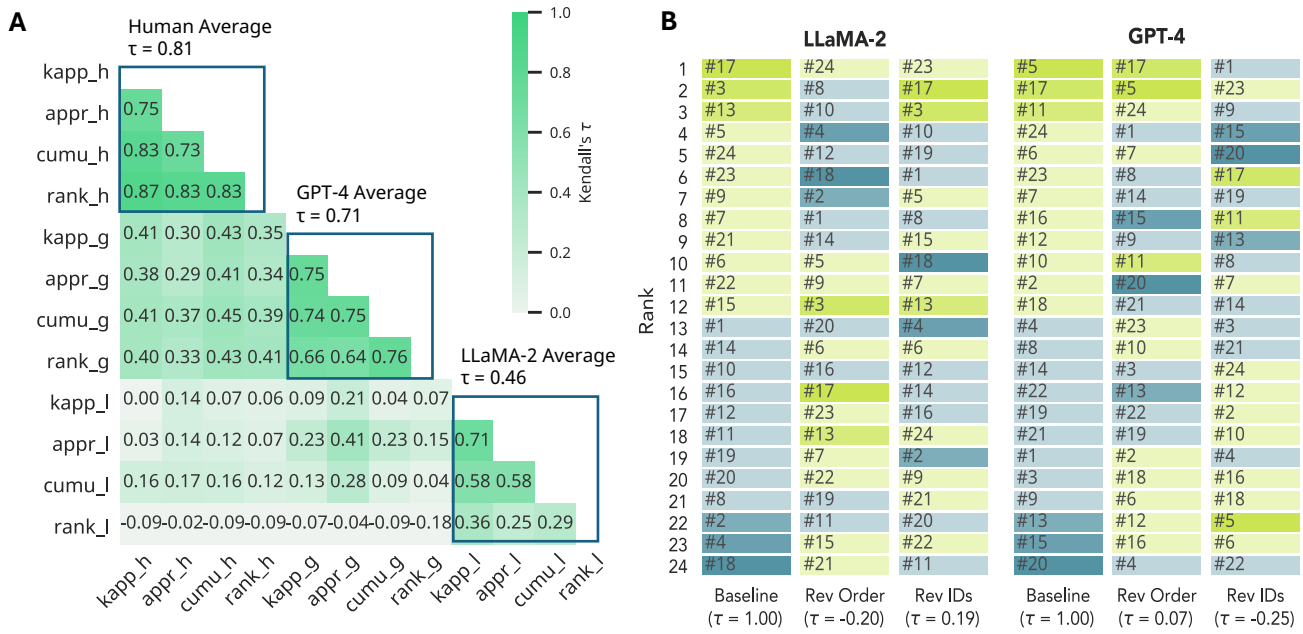

Figure 4: **A:** Heatmap of Kendall's τ across different voting methods with votes cast by human ("_h"), GPT-4 ("_g"), and LLaMA-2 ("_l") voters, showing the similarity between ranking orders of 24 projects, across different groups and methods. Voting methods include *5-Approval* (kapp), *Approval* (appr), *Cumulative (10 points)* (cumu), and *Ranked* (rank). Higher values of τ (green) show greater agreement between the compared voting methods. **B:** Comparative analysis of the primacy effect on LLM ranking behaviors in 5-Approval voting. The outcomes of human votes (baseline) are juxtaposed against results with reversed presentation order or ID sequence, showing how the presentation sequence affects ranking. Project rankings are color-coded to reflect their relative positioning, with green representing projects in the top half, and blue in the bottom half. In addition, the top and bottom three projects are shown with darker hues. Disparities in τ values underscore the variable susceptibility of each LLM to ordering effects in vote aggregation.

reasoning behind the decisions for AI explainability. The implications of this effect are further discussed in the subsequent discussion section.

Qualitative Vote Comparison to Human Voters

There are some notable differences in the proportion of votes allocated to certain types of projects by the LLaMA-2, GPT-4, and human voters (Fig. 5A). LLaMA-2 agents demonstrated a tendency to favor more budget-conscious options, shown in the cost bar plot in Fig. 5A, often selecting projects costing around CHF 5,000. Human voters are the least cost-conscious voter in this study, with the highest votes favoring the more expensive CHF 10,000 projects.

GPT-4 with no persona variation, in contrast, exhibits a strong preference in urban category for transportation-related projects. The intensity of this preference towards transportation is not shared with human voters. This discrepancy may stem from an over-reliance on the default "university student" demographic profile, coupled with a stereotypical assumption about the preferences of this group. Incorporating persona variations into GPT-4 agents helps mitigate this bias, leading to voting outcomes that more closely aligned with those of human voters.

For a more detailed comparison, see Figure 8 in the appendix, which lists project-specific differences. The voting patterns of GPT-4 agents align slightly more with human preferences than those of LLaMA-2 agents, with average deviations of 101.9% and 141.9%, respectively. LLaMA-2 agents exhibit a distinct preference for kid-oriented projects (#3 and #9), diverging from human choices. Meanwhile, GPT-4 agents without persona variation heavily favor projects enhancing bicycle infrastructure (#5, #7, #11).

Alignment Between Stated Preference and Votes

We analyzed how votes align with surveyed preferences of human voters. As shown in Figure 6, human voters typically allocate nearly half of their votes to projects within their self-reported district and preferred urban category, indicating openness to supporting projects outside their primary interests. For both LLaMA-2 and GPT-4 agents, integrating personas leads to a shift toward choices more aligned with stated preferences and human votes. However, when personas are incorporated, GPT-4 agents tend to over-align with self-reported districts and urban interests, surpassing the human tendency for self-prioritization. This overfitting suggests an excessive adherence to the characteristics of the provided human preferences.

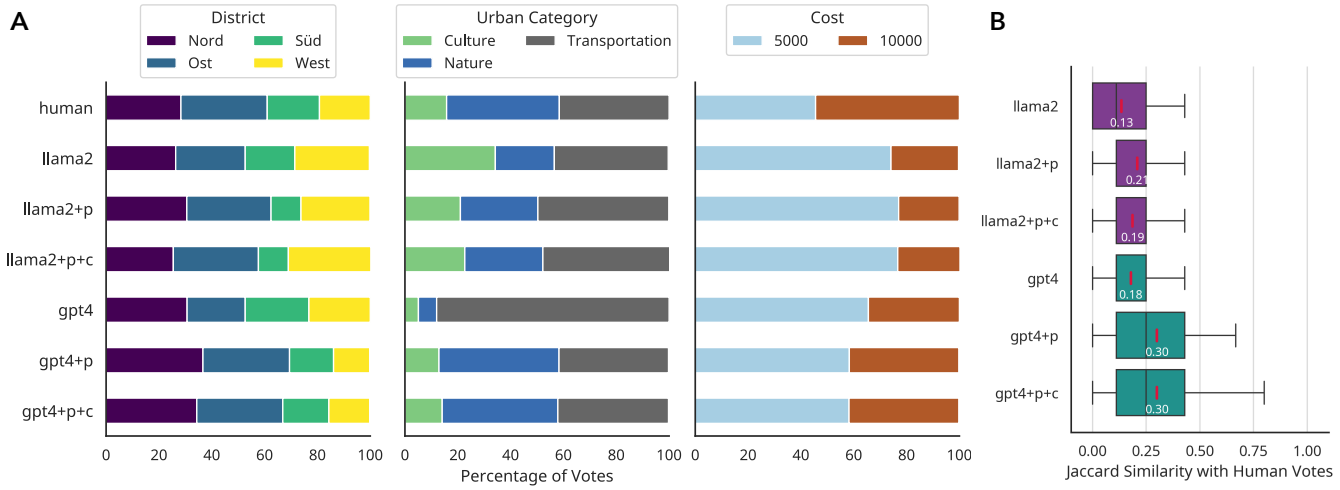

Figure 5: **A:** Stacked bar plots displaying the distribution of 5-approval votes among different districts (left) and urban project categories (right) for various voter types, including human participants, LLaMA-2, GPT-4, and their respective enhancements (persona, CoT). The addition of persona and CoT is denoted as "+p" and "+c" respectively. Each bar's segment denotes the proportion of votes contributed by each voter type to the district, category, or cost. **B:** Box plot of Jaccard Similarity of *Individual* human votes against the agent votes. The addition of persona and CoT is denoted as "+p" and "+c" respectively. The short red lines and annotated numbers denote mean values. The closer agents' votes are to human votes, the closer the value is to 1.0.

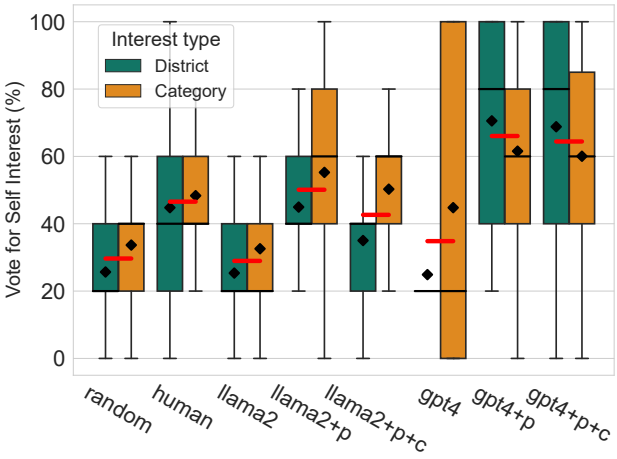

Figure 6: Box plots illustrating the distribution of vote percentages towards the voters' self-identified district and category interests. The red lines represent the median values, while the diamond symbols denote the mean values for each group. A random voting situation is added for a clearer comparison with voting outcomes that consist of a particular preference.

Diversity in Different Temperature Settings

Analysis of LLM agents' voting behaviors across different temperature settings reveals distinct trends in how these settings influence collective decision-making processes. At lower temperatures, both GPT-4 and LLaMA-2 agents demonstrate highly concentrated preferences, consistently selecting similar project sets. This behavior aligns with the deterministic nature of lower temperatures, where the LLMs favor the most probable outcomes based on their training, leading to limited diversity in choices.

As temperatures increase, there is a notable shift towards broader and more diverse preferences. This is particularly evident at a temperature of 2, where LLM agents begin to explore a wider array of options, reflecting a range of human interests (Fig. 7B). The increased randomness allows the models to simulate a more human-like variance in preferences, although this comes at the cost of reduced predictability in outcomes.

For our analysis, we simplified our description by equating the limit as temperature approaches zero ($t \to +0$) with a temperature of 0. Mathematically, as $t \to +0$, the softmax function converges to argmax, selecting the token in a perfectly deterministic manner. This behavior was observed in our experiments with LLAMA-2 but notably not with GPT-4 in Figure 7B, suggesting a possible divergence in their softmax implementation.

The alignment of LLaMA-2 and GPT-4 agents across different temperature settings with human voting patterns is illustrated in Figure 7A. Among the LLaMA-2 and GPT-4 voters themselves, outcomes generated with the temperature setting of 1.0 are found to be the most stable, as they consistently exhibit high Kendall's τ values above 0.8 compared to settings of 0.5 and 1.5.

When we compare LLM outcomes with human outcome, however, the Kendall's τ values are relatively low across all temperature settings for LLaMA-2, indicating a consistent lack of alignment with human voting outcomes. GPT-4 agents achieve a better alignment with human outcome at a temperature setting of 1.0, as evidenced by the highest τ values, 0.39, in the first column of Figure 7A.

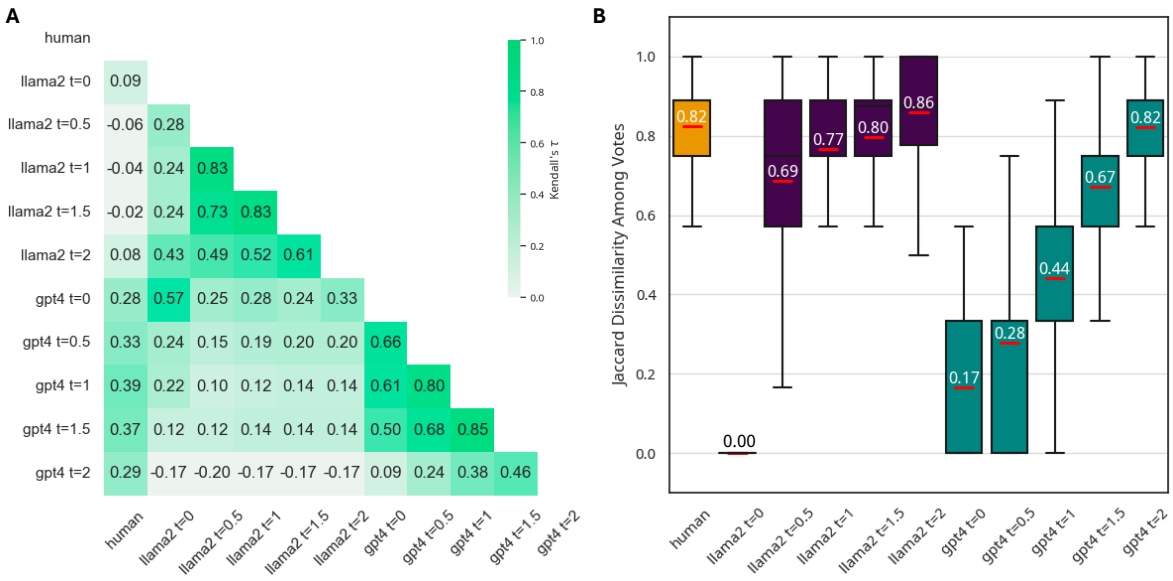

Figure 7: **A:** Heatmap displaying Kendall's τ coefficient across various temperature settings of LLaMA-2 and GPT-4 models, compared directly with human vote outcomes. Each cell represents the similarity between the *collective* ranked results of voting sets, with temperature denoted by t. Values closer to 1 indicate a higher similarity in the overall ordering of outcomes, whereas values closer to 0 indicate randomness. **B:** Box plot illustrating the Jaccard Dissimilarity between all *individual* votes from human voters and various temperature settings of LLaMA-2 and GPT-4 models. The mean for each group is indicated by a short red line, with the numerical value displayed. Dissimilarity values closer to 1 suggest greater average deviation between vote sets, indicating less similarity in voting patterns.

As expected, increasing the temperature setting in GPT-4 introduces greater diversity in the voting outcomes but also leads to more pronounced deviations from human-like decision-making. For instance, when the temperature is raised from 1.0 to 1.5, there is only a minor reduction in Kendall's τ from 0.39 to 0.37, which suggests a slight decrease in alignment with human outcomes. At the same time, this increase in temperature causes a significant rise in the Jaccard Dissimilarity of the votes, from 0.44 to 0.67, as shown in Figure 7B. In the subsequent discussion, we will assess whether the modest reduction in alignment is an acceptable compromise for the substantial gain in response diversity.

The alignment of LLaMA-2 and GPT-4 agents across different temperature settings with human voting patterns is illustrated in Figure 7A. Among the LLaMA-2 and GPT-4 voters, outcomes generated with the temperature setting of 1.0 seems to be the most stable, as they are relatively consistent with that of 0.5 or 1.5 with τ all above 0.8.

For LLaMA-2 agents, the Kendall's τ values are close to 0 across all temperature settings, indicating a consistent lack of alignment with human voting outcomes. In contrast, GPT-4 agents achieve optimal alignment at a temperature setting of 1, as evidenced by the highest τ values. Overall, when a higher temperature introduces more diversity into the voting outcomes, it also results in a greater deviation from human-like decision-making patterns. However, increasing the temperature to 1.5 results in only a slight decrease in τ from 0.39 to 0.37, yet affects the Jaccard Dissimilarity or diversity of the votes, as shown in Figure 7B, jumps from 0.44 to 0.67. The temperature of 1.5 seems to balance the need for diversity and the desire to maintain coherence with human-like decision-making.

Discussion

We compared the voting behaviors of LLMs and humans using voting data from real-world PB experimental study.

Our results reveal distinct differences in human and LLM voting patterns and collective outcomes. Humans display significant variation in the number of approved projects, while LLM agents tend to approve fewer and exhibit more uniform behavior, even with the addition of varying personas. This insight is consistent with the finding of (Abdurahman et al. 2024) that synthetic AI-simulated sampling has a WEIRD (Western, Educated, Industrialized, Rich, Democratic) bias and often fail to show meaningful variance (or diversity) in their judgments.

The results also show that the choice of voting method influence LLM votes more so than it affects humans votes. Humans demonstrated proficiency in ranking and relative comparisons. For the LLM agents, GPT-4 excelled in quantifying and distributing preferences using cumulative voting, whereas LLaMA-2 notably showed inconsistent voting, with ranked voting outcome being the least consistent outcome. This suggests that for LLM agents to make collective decisions, cumulative voting, where agents can assign precise

points, might be a better choice than ranked voting.

Additionally, list ordering also influenced the LLM voting. Although these agents often provide convincing justifications for their votes in their responses (Appendix E), a quantifiable change in collective votes reveals an inconsistency in their preferences. This highlights the importance of looking into aggregated votes on top of individual LLM responses in understanding LLM behavior in the democratic context.

Without adding a persona, LLMs display distinct preferences not shared by humans. For example, LLaMa-2 prefers kids-related projects, while GPT-4 shows a strong inclination towards transportation projects. The addition of personas partly offsets these innate preferences but introduces new challenges, leading LLMs to base their votes predominantly on persona-specific reported districts and urban interests. In contrast, human votes rely less on these factors, suggesting that their choices reflect not only self-interest but are also driven by other-regarding preferences, such as considering the community's interest.

Several strategies proposed in the literature to improve LLM behavior did not help in aligning them more closely with humans: Chain-of-thought (CoT) reasoning neither aligns LLM individual preferences nor collective outcomes more with humans, consistent with previously reported results (Chuang et al. 2024). Furthermore, varying the temperature introduces a trade-off between diversity and accuracy in human alignment. At a temperature of 1, where GPT-4 shows the highest similarity to human voting, the preference range is overly concentrated. At higher temperatures, LLM agents match the human preference diversity range but produce more random and less accurate votes.

Some aspects of LLMs could support human voting patterns in PB settings. Specifically, unlike human voters who are often less cost-conscious in their selection of projects, LLMs show a higher preference for low-budget projects. LLMs are also relatively effective in quantifying their preferences, as shown in the consistency of cumulative voting compared to the outcomes with other voting methods. This suggests that LLMs could potentially assist humans in decision-making processes involving cost considerations or more accurately present human preferences in a quantifiable manner.

Limitations: (1) Context Specificity: Our findings concerning participatory budgeting and multi-winner elections may not be so generalizable, as it does not directly apply to the more common single-winner election formats (such as presidential elections). (2) Order Effects: We find that LLM votes are susceptible to order effects. While humans also exhibit primacy and recency effects (Van Schaik, Kusev, and Juliusson 2011), our PB voting data did not include a treatment for measuring these order effects in humans. As a result, we could not quantify the difference between humans and LLMs. Future research is needed to address this gap. (3) Persona Construction: the personas were developed solely from self-reported preferences gathered through questionnaires and were not varied.

Future Directions: (1) Altruistic Factor: Our results show that personas align LLM preferences more closely with human preferences in terms of topic preference. However, they also cause a divergence from human preferences regarding the balance between self-interest and other-regarding preferences. Future research should focus on creating persona descriptions that balance both aspects to better mirror human voting behavior. (2) Performance and Diversity: Varying temperature revealed a trade-off between model performance and diversity in LLM voting. Future research should investigate the extent to which improved alignment with human votes justifies the lack of diversity among decision-making agents. (3) Exploring Various Election Types: We tested a multi-winner election format in a PB setting. Testing additional election formats in different contexts would broaden the understanding of LLM applicability in democratic processes. (4) Persona Construction: Future research could investigate scenarios where participants voluntarily provide socio-demographic details, deeming them relevant to their political decisions and accepting potential stereotyping. This approach could enhance the personalization and relevance of LLM responses in political contexts.

Conclusion

This study analyzed and compared the voting behaviors of LLM agents and human voters using data from real-world Participatory Budgeting experimental study. Our findings reveal significant differences between LLM and human voting behaviors in several key areas: the number of projects approved, consistency across different voting methods, qualitative preferences, and the degree of self-interest exhibited in voting patterns. We explored various strategies to align LLM behaviors more closely with human behaviors, but each strategy presented notable trade-offs. Incorporating personas increased similarity in preferences for certain topics but also made LLMs more self-interested compared to humans. Furthermore, no temperature setting effectively maintained both the diversity and the human-like nature of preferences. A notable advantage of LLMs in the context of PB is their cost awareness and their ability to rationalize decisions. While humans often lack cost-consciousness in participatory budgeting settings, LLMs demonstrate a greater consideration of costs. Additionally, LLMs using the Chain-of-Thought approach can provide reasoning before voting decisions, making them potentially useful for explaining collective decision-making outcomes, which can be difficult for voters with varying opinions to understand.

Overall, we empirically quantified some the biases and limitations of using LLMs in a democratic context. While LLM agents with personas can be useful in predicting voting patterns, our results indicate that the biases and lack of diversity in LLM decisions require critical evaluation before their deployment in democratic decision-making processes. Instead, we propose that LLMs be more effectively utilized within a human-in-the-loop framework. In this context, LLMs can function as support tools, assisting humans in addressing specific deficiencies in their decision-making processes, such as enhancing cost-consciousness, overcoming common human cognitive biases, summarization, explanation, or providing nuances and additional context for public policy and public sentiment.

Impact Statement

This study critically examines the use of LLMs in democratic systems, particularly in the voting process. It reveals that the current capabilities of LLMs are inadequate for capturing the full spectrum of human perspectives, which can threaten the integrity of democratic outcomes. The findings highlight the need for strict ethical guidelines to manage the integration of AI into democratic contexts. A human-centered approach is essential to ensure that AI deployment supports rather than compromises the collective intelligence derived from diverse human preferences in society.

Acknowledgments

JY would like to express gratitude to the Swiss National Science Foundation (SNSF) for the financial support provided for the previously conducted PB voting experiment in Yang et al. (2024), which was part of the National Research Programme NRP77 on Digital Transformation (project no. 187249). JY also thanks Dominik Peters, Regula Hänggli Fricker, and Evangelos Pournaras for their intellectual contributions. MK, CIH, and DH acknowledge the support from the "CoCi: Co-Evolving City Life" project, funded by the European Research Council (ERC) under the EU's Horizon 2020 research and innovation program (grant no. 833168). DD acknowledges the support from the Distributed Intelligence and Technology for Traffic and Mobility Management (DIT4TraM) project, funded by the EU's Horizon 2020 Research and Innovation Programme (grant no. 953783).

References

Abdurahman, S.; Atari, M.; Karimi-Malekabadi, F.; Xue, M. J.; Trager, J.; Park, P. S.; Golazizian, P.; Omrani, A.; and Dehghani, M. 2024. Perils and opportunities in using large language models in psychological research. *PNAS Nexus*, 3.

Aher, G.; Arriaga, R. I.; and Kalai, A. T. 2023. Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies. arXiv:2208.10264.

Allen, D.; and Weyl, E. 2024. The Real Dangers of Generative AI. *Journal of Democracy*, 35(1): 147–162.

Argyle, L. P.; Bail, C. A.; Busby, E. C.; Gubler, J. R.; Howe, T.; Rytting, C.; Sorensen, T.; and Wingate, D. 2023a. Leveraging AI for democratic discourse: Chat interventions can improve online political conversations at scale. *Proceedings of the National Academy of Sciences*, 120(41): e2311627120. Publisher: Proceedings of the National Academy of Sciences.

Argyle, L. P.; Busby, E. C.; Fulda, N.; Gubler, J. R.; Rytting, C.; and Wingate, D. 2023b. Out of One, Many: Using Language Models to Simulate Human Samples. *Political Analysis*, 31(3): 337–351.

Arrow, K. J. 1951. *Social Choice and Individual Values*. Yale University Press.

Axelrod, R.; and Hamilton, W. D. 1981. *The Evolution of Cooperation*. Basic Books.

Azaria, A.; and Mitchell, T. 2023. The Internal State of an LLM Knows When It's Lying. arXiv:2304.13734.

Aziz, H.; and Huang, S. 2017. A Polynomial-time Algorithm to Achieve Extended Justified Representation. *ar5iv*.

Becker, G. S. 1976. *The Economic Approach to Human Behavior*. University of Chicago Press.

Blanco, M.; Engelmann, D.; and Normann, H. T. 2011. A within-subject analysis of other-regarding preferences. *Games and Economic Behavior*, 72(2): 321–338.

Boiko, D. A.; MacKnight, R.; and Gomes, G. 2023. Emergent autonomous scientific research capabilities of large language models. ArXiv:2304.05332 [physics].

Bonabeau, E.; Dorigo, M.; and Theraulaz, G. 1999. *Swarm Intelligence: From Natural to Artificial Systems*. Oxford University Press.

Brams, S. J.; and Fishburn, P. C. 1983. *Approval Voting*. Boston: Birkhäuser. ISBN 9783764331085.

Bran, A. M.; Cox, S.; Schilter, O.; Baldassari, C.; White, A. D.; and Schwaller, P. 2023. ChemCrow: Augmenting large-language models with chemistry tools. arXiv:2304.05376.

Chan, C.-M.; Chen, W.; Su, Y.; Yu, J.; Xue, W.; Zhang, S.; Fu, J.; and Liu, Z. 2023. ChatEval: Towards Better LLM-based Evaluators through Multi-Agent Debate. arXiv:2308.07201.

Chuang, Y.-S.; Harlalka, N.; Suresh, S.; Goyal, A.; Hawkins, R.; Yang, S.; Shah, D.; Hu, J.; and Rogers, T. T. 2024. The Wisdom of Partisan Crowds: Comparing Collective Intelligence in Humans and LLM-based Agents. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 46.

Coda-Forno, J.; Binz, M.; Wang, J. X.; and Schulz, E. 2024. CogBench: a large language model walks into a psychology lab. *arXiv preprint arXiv:2402.18225*.

Du, Y.; Li, S.; Torralba, A.; Tenenbaum, J. B.; and Mordatch, I. 2023. Improving Factuality and Reasoning in Language Models through Multiagent Debate. ArXiv:2305.14325 [cs].

Durmus, E.; Nyugen, K.; Liao, T. I.; Schiefer, N.; Askell, A.; Bakhtin, A.; Chen, C.; Hatfield-Dodds, Z.; Hernandez, D.; Joseph, N.; Lovitt, L.; McCandlish, S.; Sikder, O.; Tamkin, A.; Thamkul, J.; Kaplan, J.; Clark, J.; and Ganguli, D. 2023. Towards Measuring the Representation of Subjective Global Opinions in Language Models. ArXiv:2306.16388 [cs].

Elkind, E.; Faliszewski, P.; Skowron, P.; and et al. 2017. Properties of Multiwinner Voting Rules. *Social Choice and Welfare*, 48(3): 599–632.

Feng, S.; Park, C. Y.; Liu, Y.; and Tsvetkov, Y. 2023. From Pretraining Data to Language Models to Downstream Tasks: Tracking the Trails of Political Biases Leading to Unfair NLP Models. ArXiv:2305.08283 [cs].

Ferland, B. 2015. A rational or a virtuous citizenry? – The asymmetric impact of biases in votes-seats translation on citizens' satisfaction with democracy. *Electoral Studies*, 40: 394–408.

Fish, S.; Gölz, P.; Parkes, D. C.; Procaccia, A. D.; Rusak, G.; Shapira, I.; and Wüthrich, M. 2023. Generative social choice. *arXiv preprint arXiv:2309.01291*.

Floridi, L. 2023. AI as agency without intelligence: on ChatGPT, large language models, and other generative models. *Philosophy & Technology*, 36(1): 15.

Gersbach, H.; and Martinelli, C. 2023. Supported democracy: reinventing direct democracy, AI and voting twice. Accessed: 2024-01-24.

Gudiño-Rosero, J.; Grandi, U.; and Hidalgo, C. A. 2024. Large Language Models (LLMs) as Agents for Augmented Democracy. *arXiv preprint arXiv:2405.03452*.

Hao, R.; Hu, L.; Qi, W.; Wu, Q.; Zhang, Y.; and Nie, L. 2023. ChatLLM Network: More brains, More intelligence. ArXiv:2304.12998 [cs].

Hausladen, C. I.; Hänggli, R.; Helbing, D.; Kunz, R.; Wang, J.; and Pournaras, E. 2023. On the Legitimacy of Voting Methods. *Available at SSRN 4372245*.

Helbing, D.; Mahajan, S.; Hänggli, R.; Musso, A.; Hausladen, C. I.; Carissimo, C.; Carpentras, D.; Stockinger, E.; Argota Sánchez-Vaquerizo, J.; Yang, J.; et al. 2022. Democracy by Design: Perspectives for Digitally Assisted, Participatory Upgrades of Society. *Participatory Upgrades of Society (November 2, 2022)*, 71.

Hidalgo, C. 2018. Augmented Democracy. Accessed: 2024-01-12.

Hong, S.; Zhuge, M.; Chen, J.; Zheng, X.; Cheng, Y.; Zhang, C.; Wang, J.; Wang, Z.; Yau, S. K. S.; Lin, Z.; Zhou, L.; Ran, C.; Xiao, L.; Wu, C.; and Schmidhuber, J. 2023. MetaGPT: Meta Programming for A Multi-Agent Collaborative Framework. arXiv:2308.00352.

Horton, J. J. 2023. Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus? ArXiv:2301.07543 [econ, q-fin].

Jarrett, D.; Pislar, M.; Bakker, M. A.; Tessler, M. H.; Koster, R.; Balaguer, J.; Elie, R.; Summerfield, C.; and Tacchetti, A. 2023. Language agents as digital representatives in collective decision-making. In *NeurIPS 2023 Foundation Models for Decision Making Workshop*.

Koppell, J. G.; and Steen, J. A. 2004. The Effects of Ballot Position on Election Outcomes. Publications from President Jonathan G.S. Koppell. 10.

Laslier, J.-F.; and Van Der Straeten, K. 2016. Strategic voting in multi-winner elections with approval balloting: a theory for large electorates. *Social Choice and Welfare*, 47(3): 559–587.

Li, G.; Hammoud, H. A. A. K.; Itani, H.; Khizbullin, D.; and Ghanem, B. 2023. CAMEL: Communicative Agents for "Mind" Exploration of Large Language Model Society. arXiv:2303.17760.

Liang, T.; He, Z.; Jiao, W.; Wang, X.; Wang, Y.; Wang, R.; Yang, Y.; Tu, Z.; and Shi, S. 2023. Encouraging Divergent Thinking in Large Language Models through Multi-Agent Debate.

Liu, R.; Yang, R.; Jia, C.; Zhang, G.; Zhou, D.; Dai, A. M.; Yang, D.; and Vosoughi, S. 2023. Training Socially Aligned Language Models on Simulated Social Interactions. ArXiv:2305.16960 [cs].

Margetts, H.; John, P.; Hale, S.; and Yasseri, T. 2016. *Political Turbulence: How Social Media Shape Collective Action*. Princeton University Press. ISBN 9780691159225.

Miller, J. M.; and Krosnick, J. A. 1998. The Impact of Candidate Name Order on Election Outcomes. *The Public Opinion Quarterly*, 62(3): 291–330.

Park, J. S.; O'Brien, J. C.; Cai, C. J.; Morris, M. R.; Liang, P.; and Bernstein, M. S. 2023. Generative Agents: Interactive Simulacra of Human Behavior. ArXiv:2304.03442 [cs].

Rank, A. D.; and Mushtare, R. 2019. Using Personas as a Tool for Teaching Civic Communication. *The Journal of General Education*, 68(3-4): 252–262.

Sen, A. 1995. Rationality and social choice. *The American economic review*, 85(1): 1.

Serapio-García, G.; Safdari, M.; Crepy, C.; Sun, L.; Fitz, S.; Romero, P.; Abdulhai, M.; Faust, A.; and Matarić, M. 2023. Personality Traits in Large Language Models. ArXiv:2307.00184 [cs].

Shoham, Y.; and Leyton-Brown, K. 2009. *Multiagent Systems: Algorithmic, Game-Theoretic, and Logical Foundations*. Cambridge University Press.

Skowron, P.; Faliszewski, P.; and Slinko, A. 2019. Axiomatic characterization of committee scoring rules. *Journal of Economic Theory*, 180: 244–273.

Sobieszek, A.; and Price, T. 2022. Playing Games with Ais: The Limits of GPT-3 and Similar Large Language Models. *Minds and Machines*, 32(2): 341–364.

Sourati, J.; and Evans, J. A. 2023. Accelerating science with human-aware artificial intelligence. *Nature Human Behaviour*, 7(10): 1682–1696.

Strachan, J. W. A.; Albergo, D.; Borghini, G.; Pansardi, O.; Scaliti, E.; Gupta, S.; Saxena, K.; Rufo, A.; Panzeri, S.; Manzi, G.; Graziano, M. S. A.; and Becchio, C. 2024. Testing theory of mind in large language models and humans. *Nature Human Behaviour*, 8: 1285 – 1295.

Talebirad, Y.; and Nadiri, A. 2023. Multi-Agent Collaboration: Harnessing the Power of Intelligent LLM Agents. ArXiv:2306.03314 [cs].

Tang, A.; Weyl, G.; and the Plurality Community. 2023. Plurality: The Future of Collaborative Technology and Democracy.

Turpin, M.; Michael, J.; Perez, E.; and Bowman, S. 2024. Language models don't always say what they think: unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36.

Törnberg, P.; Valeeva, D.; Uitermark, J.; and Bail, C. 2023. Simulating Social Media Using Large Language Models to Evaluate Alternative News Feed Algorithms. ArXiv:2310.05984 [cs].

van Erkel, P. F.; and Thijssen, P. 2016. The first one wins: Distilling the primacy effect. *Electoral Studies*, 44: 245–254.

Van Schaik, P.; Kusev, P.; and Juliusson, A. 2011. Human preferences and risky choices.

Walsh, V. 1994. Rationality as Self-Interest versus Rationality as Present Aims. *The American Economic Review*, 84(2): 401–405.

Wang, L.; Ma, C.; Feng, X.; Zhang, Z.; Yang, H.; Zhang, J.; Chen, Z.; Tang, J.; Chen, X.; Lin, Y.; Zhao, W. X.; Wei, Z.; and Wen, J.-R. 2023a. A Survey on Large Language Model based Autonomous Agents. ArXiv:2308.11432 [cs].

Wang, Y.; Cai, Y.; Chen, M.; Liang, Y.; and Hooi, B. 2023b. Primacy Effect of ChatGPT. ArXiv:2310.13206 [cs].

Wang, Z.; Mao, S.; Wu, W.; Ge, T.; Wei, F.; and Ji, H. 2023c. Unleashing Cognitive Synergy in Large Language Models: A Task-Solving Agent through Multi-Persona Self-Collaboration. ArXiv:2307.05300 [cs].

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35: 24824–24837.

Wu, Q.; Bansal, G.; Zhang, J.; Wu, Y.; Li, B.; Zhu, E.; Jiang, L.; Zhang, X.; Zhang, S.; Liu, J.; Awadallah, A. H.; White, R. W.; Burger, D.; and Wang, C. 2023. AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation. ArXiv:2308.08155 [cs].

Yang, J. C.; Hausladen, C. I.; Peters, D.; Pournaras, E.; Fricker, R. H.; and Helbing, D. 2024. Designing Digital Voting Systems for Citizens: Achieving Fairness and Legitimacy in Digital Participatory Budgeting Voting Mechanism. *ACM Digital Government: Research and Practice*.

Yao, J.-Y.; Ning, K.-P.; Liu, Z.-H.; Ning, M.-N.; and Yuan, L. 2023. LLM Lies: Hallucinations are not Bugs, but Features as Adversarial Examples. arXiv:2310.01469.

Zhang, C.; Yang, K.; Hu, S.; Wang, Z.; Li, G.; Sun, Y.; Zhang, C.; Zhang, Z.; Liu, A.; Zhu, S.-C.; Chang, X.; Zhang, J.; Yin, F.; Liang, Y.; and Yang, Y. 2023a. ProAgent: Building Proactive Cooperative AI with Large Language Models. arXiv:2308.11339.

Zhang, H.; Du, W.; Shan, J.; Zhou, Q.; Du, Y.; Tenenbaum, J. B.; Shu, T.; and Gan, C. 2023b. Building Cooperative Embodied Agents Modularly with Large Language Models. arXiv:2307.02485.

Zheng, H. S.; Mishra, S.; Chen, X.; Cheng, H.-T.; hsin Chi, E. H.; Le, Q. V.; and Zhou, D. 2023. Take a Step Back: Evoking Reasoning via Abstraction in Large Language Models. *ArXiv*, abs/2310.06117.

Zhuge, M.; Liu, H.; Faccio, F.; Ashley, D. R.; Csordás, R.; Gopalakrishnan, A.; Hamdi, A.; Hammoud, H. A. A. K.; Herrmann, V.; Irie, K.; Kirsch, L.; Li, B.; Li, G.; Liu, S.; Mai, J.; Piękos, P.; Ramesh, A.; Schlag, I.; Shi, W.; Stanić, A.; Wang, W.; Wang, Y.; Xu, M.; Fan, D.-P.; Ghanem, B.; and Schmidhuber, J. 2023. Mindstorms in Natural Language-Based Societies of Mind. ArXiv:2305.17066 [cs].