The Wisdom of Partisan Crowds: Comparing Collective Intelligence in Humans and LLM-based Agents

Anonymous ACL submission

Abstract

Human groups are able to converge on more accurate beliefs through deliberation, even in the presence of polarization and partisan bias — a phenomenon known as the "wisdom of partisan crowds." Similar mechanisms may be crucial for mitigating bias and inaccuracy in Large Language Models (LLMs), yet few benchmarks exist for the collective intelligence of LLM-based agents. In this study, we evaluate the extent to which collectives of Large Language Models (LLMs) display human-like social dynamics. We evaluate how agents' responses evolve through repeated social interaction. Even when LLM agents are prompted to role-play as partisan personas (e.g. Democrat or Republican), we find that they are able to converge to more accurate beliefs, as humans do. However, we also observe several factors that interfere with convergence, including chain-of-thought reasoning and more extreme partisan biases. Conversely, fine-tuning on human data appears to enhance the wisdom of crowds effect. These findings show the potential and limitations of using LLM-based agents as a model of human collective intelligence.

1 Introduction

As Large Language Models (LLMs) have grown more human-like in the behaviors they produce (Park et al., 2022), there has been increasing interest in investigating whether they can be used to better understand emulate human communication in social groups. As one example, Park et al. (2023) used LLMs to construct *generative agents* that interact with each other in a simulated environment: initiating conversations, spreading information, remembering past events, and planning future actions. The resulting vignettes can show remarkably convincing interactions in which information about novel events, such as the planning of a birthday party, diffuses throughout the community of simulated agents. Yet it is difficult to understand how human-like such patterns really are, and consequently how useful such simulated systems are for understanding human communicative phenomena, without replicable empirical benchmarks of human behavior for comparison.

The current paper develops one such benchmark deriving from a phenomenon in the study of human collective intelligence, the *wisdom of (partisan) crowds*. The phenomenon reflects two interesting characteristics of human cognition. First, when estimating real-world quantities related to a politically polarized issue, self-identified Republicans and Democrats often generate systematically different guesses that reflect their political leanings. For example, when asked to estimate the US employment rate during Barack Obama's administration, both groups overestimate relative to the ground truth, but Republicans produce reliably higher over-]estimates, presumably remembering unemployment as worse than it actually was under the Democratic regime.

Second, when shown the mean guess from other members of their preferred party, *both* groups adjust their estimates in ways that move the group mean systematically closer to the ground truth. In a seminal study, Becker et al. (2019) collected data from 1120 participants identifying as either Republican or Democrat. Each was asked factual questions known to elicit partisan bias, and after responding, was shown the average belief of others in their same partisan group (ie other Democrats or Republicans). Participants were then permitted to adjust their estimate, and the same procedure was repeated, yielding a series of three estimates for each respondent. The authors observed that, after each round of feedback, the mean estimate for both groups moved closer to the ground truth.

This phenomenon is useful for assessing LLM simulation of human communication, for three reasons. First, all of the questions have a ground-truth value that is easily searchable in traditional media,

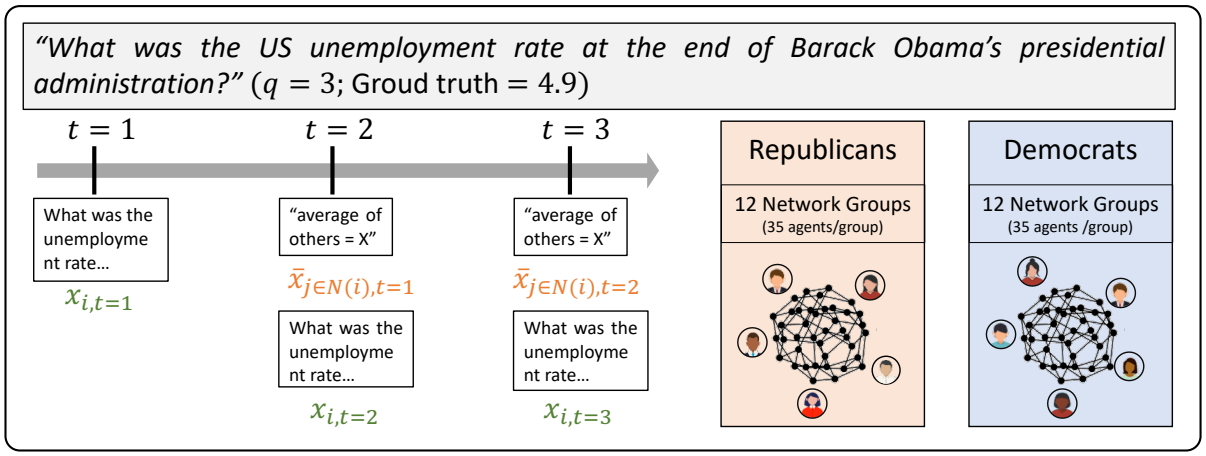

Figure 1: Experimental design comparing social feedback effects on LLM agents' estimations of partisan-biased factual questions. LLM agents role-playing Democrat and Republican update their estimates after considering their peers' average responses. (Becker et al., 2019).

providing a useful means of quantifying how accurate LLM estimates actually are. Second, when tested without role-playing prompts, contemporary LLMs typically get these questions exactly right, whereas humans typically don't know the ground truth and show partisan lean in their estimates. This provides an opportunity to evaluate whether LLMs, when asked to role-play as Democrats or Republicans, show reliably human-like patterns of partisan bias in their responses. Third, The social exchange of information within human partisan groups increased mean accuracy for each while also reducing polarization between groups, providing a reliable dynamic phenomenon in human communication that can be assessed in LLMs.

For these reasons, we replicated the experimental design of Becker et al. (2019), applying it to groups of interacting, role-playing LLM agents in a simulated environment, and assessing whether the resulting system replicated the key phenomena in human behavior. We found that LLM agents, when operating without Chain-of-Thought (CoT) reasoning, exhibit a substantial Wisdom of Partisan Crowds (WOC) effect, closely paralleling human patterns of error reduction in group settings. However, the use of CoT reasoning reduced this effect. We also show that the "depth of persona" created in the role-playing prompt critically influences whether LLM agents exhibit human-like partisanship bias in their estimates. Finally, fine-tuning LLMs with human data enhances human-like group dynamics in held-out data, though such training also risks overfitting. Together the work suggests a promising approach toward using esablished behavioral phenomena from human participants to evaluate and shape the use of LLMs for understanding dynamics of social communication.

2 Methods

2.1 Experimental Procedure

We followed the experimental design of Becker et al. (2019) using LLMs to role-play Democrat and Republican personas. Each LLM agent is embedded in a network structure that governs interactions, connecting to four others sharing the same political leaning (all agents have node degree = 4; see Figure 1) and thus reflecting the homogeneous group structures in human studies. Each LLM agent, powered by LangChain (Chase, 2022) with OpenAI's ChatGPT (`gpt-3.5-turbo`; OpenAI, 2022) and the open-source LLM Vicuna (`vicuna-33B-v1.3`; Zheng et al., 2023), maintains the continuity of each persona's memory throughout the experiment. Over three rounds, these agents were prompted to answer the same eight fact-based questions with known partisan biases as shown in Figure 1. After each round, agents were given the average estimates of their connected peers and were asked to provide their estimates again. Thus at the end of the three rounds, each agent had produced three estimates for each of the eight questions. [1] The full list of questions is provided in §D. The full procedure was conducted 12 times for each group to mirror the 12 groups of human participants in the social conditions in Becker et al. (2019). Temperate sampling (temperature =

[1] The original study by Becker et al. (2019) divided these questions into two separate experiments, with the first four questions belonging to the first experiment and the last four to the second. However, as both sets of questions follow the same experimental procedure, we have merged them into a single analysis in our study.

Model	Role-playing Method	Persona	CoT	$HLI \uparrow$	$\overline{\Delta\varepsilon} \downarrow$	$\overline{\beta_{\text{PB}}} \uparrow$	*Ext.%*
ChatGPT	Vanilla Role-playing	Detailed	CoT	4.45 ± 0.8	**-1.08 ± 0.76**	**3.37 ± 0.25**	0.00
			No CoT	**12.82 ± 1.89**	**-7.59 ± 1.87**	**5.23 ± 0.28**	0.00
		Simple	CoT	-20.13 ± 1.1	**-2.07 ± 0.87**	**-22.2 ± 0.67**	0.00
			No CoT	-21.8 ± 1.77	**-3.11 ± 1.47**	**-24.91 ± 0.98**	0.00
	Induced Partisanship	Simple	CoT	8.3 ± 1.22	**-2.09 ± 0.82**	6.21 ± 0.90	0.00
			No CoT	-31.6 ± 2.64	**-8.47 ± 1.91**	-40.06 ± 1.82	0.00
Vicuna-33B	Vanilla Role-playing	Detailed	CoT	2.81 ± 1.36	**2.87 ± 1.27**	**5.68 ± 0.49**	1.31
			No CoT	**3.67 ± 2.64**	**0.68 ± 2.51**	**4.36 ± 0.8**	1.38
		Simple	CoT	3.36 ± 1.25	**0.59 ± 1.18**	**3.94 ± 0.41**	0.98
			No CoT	-0.63 ± 2.63	**0.49 ± 2.47**	**-0.14 ± 0.91**	5.60
	Induced Partisanship	Simple	CoT	0.17 ± 5.41	**20.68 ± 5.24**	20.85 ± 1.34	13.02
			No CoT	-17.55 ± 2.76	**9.09 ± 2.28**	-8.46 ± 1.56	4.84
Human	-	-	-	66.5 ± 6.79	-33.16 ± 6.74	33.35 ± 0.83	8.37

Table 1: Evaluation of resemblance between LLM agent and human in social interaction setting. The three main human-LLM alignment metrics are, HLI (the more positive, the more human-like, $\overline{\Delta\varepsilon}$ (the more negative, the stronger the WOC effect) and $\overline{\beta_{\text{PB}}}$ (the more positive, the more aligned with human). The black boldface highlights the condition with the highest HLI. The metrics are shown with the standard errors. Notably, when there is **no CoT** reasoning, $\overline{\Delta\varepsilon}$ is always more negative than using **CoT** reasoning. In addition, using a **detailed persona** always leads to a more positive $\overline{\beta_{\text{PB}}}$ than using a **simple persona**.

0.7) was used to allow variability in responses. §F shows the actual prompts. The compute resource for using Vicuna is in §I.

Formal Notations We denote each agent in the experiment as $a_{i,p,r}$, where $i \in [1, 35]$ indexes the agent within a specific run, p denotes political leaning (Democrat or Republican; abbreviated as Dem and Rep hereafter), and r specifies the run index with $r \in [1, 12]$. During each run, agents answer eight questions across three time steps, generating a series of estimates $x^t_{i,p,r,q}$ for question q at time t. Since all eight questions are fact-based, each has a ground-truth value, denoted as x^*_q. Starting at $t \geq 2$, agents are shown the average estimate from their politically homogeneous neighbors before making their own estimates. The average estimate from neighbors for agent $a_{i,p,r}$ at time t for question q is $m^t_{i,p,r,q} = \frac{1}{K}\sum_{j \in \mathcal{N}(i,p,r)} x^{t-1}_{j,p,r,q}$, where $\mathcal{N}(i, p, r)$ is the set of indices for the agent's neighbors who share the same political leaning p. We set the number of neighbors to $K = 4$ in this setup.

2.2 Personas and Agent Specification

Persona Detail As noted in the introduction, without role-playing LLMs can often generate the correct answer for widely-reported matters of fact. To elicit differing responses, and to assess whether LLMs can capture partisan differences in such estimates, we asked the LLMs to role-play as different personas created with varying levels of background detail. *Simple Personas* were specified as "a typical Democrat/Republican," relying on temperature sampling to elicit slightly different biased views. *Detailed Personas* were provided with comprehensive backstories, including demographics and personal background information, to introduce individual differences based on such factors. §C shows the full list of both detailed and simple personas. The detailed personas are generated by an LLM to represent a diverse sample.

Chain-of-thought Reasoning (CoT) Another factor is to decide whether to encourage the agent to employ chain-of-thought (CoT) reasoning (Wei et al., 2022b,a). CoT has demonstrated success as a prompting strategy in solving complex reasoning tasks, such as arithmetic problems. However, recent work also indicates that CoT reasoning may lead to stereotypes and biases (Shaikh et al., 2022). This leads us to explore how CoT reasoning influences an LLM agent's ability to assume human-like behaviors in social interaction setting.

Induced Partisanship We also consider explicitly introducing the expected partisan biases into the personas' prompts. This serves as a validation to see if one can control the partisan bias directly in the prompt. For example, a Republican persona may be reminded of a belief that Obama's presidency was ineffective in reducing unemployment. The induced partisanship encourages the agents to behave according to known human biases as in

Becker et al. (2019). §E shows the full list of induced partisanship for each question. In contrast to this approach, when agents are role-playing without any additional induced biases., we termed it "*vanilla role-playing*".

2.3 Fine-Tuning the LLMs with Human Data

In addition to in-context learning through prompting, we also perform supervised fine-tuning using human response data from the experiment in Becker et al. (2019) to enhance the resemblance of human behaviors in LLM agents. Our fine-tuning methodology is inspired by Binz and Schulz (2023), showing that through supervised learning, LLMs can be adapted to modeling human decision-making behavior in an unseen task. We aim to investigate whether fine-tuning also improves the resemblance of human-like behavior in group interaction settings. We fine-tune two separate LLMs: one for Democrats and another one for Republicans. The training data consists of responses to questions $q \in [5, 8]$, while using questions $q \in [1, 4]$ as the testing set. The fine-tuned model is then evaluated separately on the train set and the test set. When fine-tuning, no persona is provided. Details on how we fine-tuned the LLM (ChatGPT in specific) are in the §H.

2.4 Evaluation Metrics

Our evaluation metrics are designed to assess how *human-like* LLM agent group dynamics is.

Wisdom of Partisan Crowds Effect (WOC) The WOC effect quantifies the improvement in LLM agents' estimates through social interaction, similar to human groups (Becker et al., 2019). For each question q, run r, time step t, and political leaning p, we compute the group mean $\bar{x}^t_{p,r,q} = \frac{1}{N}\sum i = 1^N x^t_{i,p,r,q}$ (with $N = 35$ per group), and the normalized group mean $\eta^t_{p,r,q} = 100 \times (\bar{x}^t_{p,r,q} - x^* q)/x^q$. The normalized group error $\varepsilon p, r, q^t = |\eta^t_{p,r,q}|$ shows the percentage deviation from ground truth $|x^|$. We measure group error reduction per question as $\Delta\varepsilon_{p,r,q} = \varepsilon^{t=3}_{p,r,q} - \varepsilon^{t=1}_{p,r,q}$, and average these across all groups and questions to get the average group error reduction $\overline{\Delta\varepsilon}$. A more negative $\overline{\Delta\varepsilon}$ indicates a stronger WOC effect, with $\overline{\Delta\varepsilon}$ representing the percentage of ground truth size $|x^*|$ by which estimates move towards the truth. We also use the question-specific group error reduction $\Delta\varepsilon_q$ for individual question analysis. For detailed derivation, see §J.1.

Partisan Bias To evaluate the human-like partisan biases in LLM agents, we define *Partisan Bias* as the average difference in normalized group mean $\eta^t_{p,r,q}$ between Democrat and Republican groups, in line with expected human partisan bias directions. Formally, $\overline{\beta_{\text{PB}}} = \frac{1}{Q'TR^2}\sum_{q=1}^{Q'}\sum_{t=1}^{T}\sum_{r'=1}^{R}\sum_{r^\dagger=1}^{R}(\eta^t_{p=Rep,r',q} - \eta^t_{p=Dem,r^\dagger,q}) \times \text{sign}(h_q)$, where $T = 3$ represents the time steps, $R = 12$ the number of runs, $\text{sign}(h_q)$ indicates the human partisan bias direction, with $+1$ if Republicans typically have greater $\eta^t_{p,r,q}$ than Democrats (i.e., a more positive $\bar{x}^t_{p,r,q}$ if $x^* > 0$), as per human data in Becker et al. (2019), -1 if the other way around, and 0 if there is no expected difference (see §C for $\text{sign}(h_q)$ of each question), and $Q' = 5$ the questions where human shows partisan bias (i.e., $\text{sign}(h_q) \neq 0$). A positive $\overline{\beta_{\text{PB}}}$ indicates an resemblance to the direction of human bias, while a negative value suggests a divergence. [2]. In addition, we also denote *question-specific partisan bias* $\beta_{\text{PB,q}} = \frac{1}{TR^2}\sum_{t=1}^{T}\sum_{r'=1}^{R}\sum_{r^\dagger=1}^{R}(\eta^t_{p=Rep,r',q} - \eta^t_{p=Dem,r^\dagger,q}) \times \text{sign}(h_q)$, which is useful when referring to results of a given question.

Human Likeness Index (HLI) We introduce the Human Likeness Index (HLI) to assess the extent of LLM agents' resemblance to human behaviors. To aggregate across the WOC effect ($\overline{\Delta\varepsilon}$) and the partisan bias ($\overline{\beta_{\text{PB}}}$), we define $HLI = \overline{\beta_{\text{PB}}} + (-\overline{\Delta\varepsilon})$.[3] A higher HLI score indicates a stronger overall human-like behavior in LLM agents within this group experiment.

Extreme Values (*Ext.%*) The *Ext.%* metric evaluates the proportion of LLM agent responses that are unrealistic, based on established criteria (Becker et al., 2019). For a fair comparison with human data, the same criteria for identifying extreme values are applied, for instance, marking any unemployment rate response above 47% as extreme. These criteria are detailed in §G. Extreme values are excluded from calculations of *Average Group Error Reduction* ($\overline{\Delta\varepsilon}$) and *Partisan Bias* ($\overline{\beta_{\text{PB}}}$). The *Ext.%* thereby serves as a measure of the LLM agents' tendency to generate unrealistic responses.

[2] Because $\eta^t_{p,r,q}$ is scaled by a factor of 100, $\overline{\beta_{\text{PB}}}$ can be interpreted as the partisan bias expressed in *percentage* of the size of ground truth $|x^*|$

[3] A linear addition makes sense because both $\overline{\beta_{\text{PB}}}$ and $\overline{\Delta\varepsilon}$ are on the same scale. They can both be expressed as the percentage of the size of ground truth value $|x^*|$.

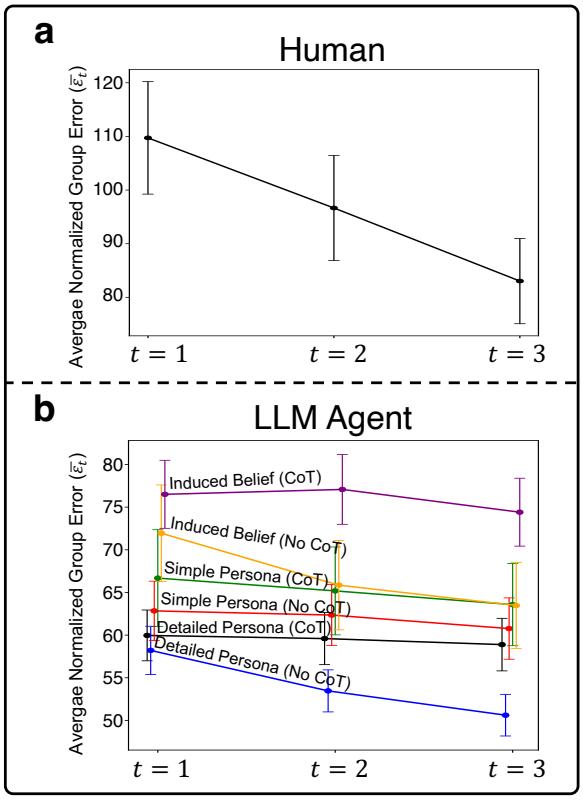

Figure 2: Average Normalized Group Error ($\bar{\varepsilon}_t$) for (a) human crowds and (b) LLM agents (ChatGPT) across six experimental settings. Error bars indicating standard errors.

Revision Coefficient In human crowds, the *mechanism* for why the group mean converges towards the truth through social interaction is that those who are more accurate at their initial estimate tend to be influenced less by the information they received, and thus pull the group distribution towards the truth (Becker et al., 2017). Following Becker et al. (2017)' methodology, for each question q, we calculate the *revision coefficient* ($r_{\text{adj},q}$), defined as the partial correlation between *individual revision* ($\Delta x_{i,p,r,q} = |x^{t=3}_{i,p,r,q} - x^{t=1}_{i,p,r,q}|$) and *individual initial error* ($e_{i,p,r,q} = |x^{t=1}_{i,p,r,q} - x^{*}_{p}|$), adjusted for the *social signal* ($s_{i,p,r,q} = |x^{t=1}_{i,p,r,q} - m^{t=2}_{i,p,q,r}|$) that each individual receives. Adjusting for the social signal is important as individuals with higher initial errors often receive stronger social feedback as they deviate from the rest. Formally, $r_{\text{adj},q} = \text{corr}(\widetilde{\Delta x}_{i,p,r,q}, \widetilde{e}_{i,p,r,q})$, where $\widetilde{\Delta x}_{i,p,r,q}$ and $\widetilde{e}_{i,p,r,q}$ are $\Delta x_{i,p,r,q}$ and $e_{i,p,r,q}$ adjusted by social signal, respectively. Please refer to §J.2 for detailed derivation.

3 Results

3.1 Effect of Persona Detail, CoT Reasoning, and Induced Partisanship

Role-Playing without Chain-of-Thought Reasoning LLM agents with detailed personas, but without CoT reasoning, demonstrate the closest resemblance to human group dynamics. They demonstrate the highest human likeness, $HLI = 12.82$ (ChatGPT) and 3.67 (Vicuna) among the six experimental settings (Table 1). Figure 2 visualizes the wisdom of partisan crowd results of LLM agents (ChatGPT). These agents converge significantly towards the ground truth after social interaction, quantified by a significant WOC effect, $\overline{\Delta\varepsilon} = -7.59$, $CI_{95\%} = [-11.08, -4.10]$, $p < .001$. It also shows a significant human-like partisan bias, $\overline{\beta_{\text{PB}}} = 5.23, CI_{95\%} = [4.66, 5.81]$, $p < .001$. [4] Figure 3 shows the detailed result for each question. Next, we look into the role of persona detail, CoT reasoning, and induced partisanship, respectively. The result with Vicuna is shown in Figure 6 in §B.

Absence of CoT Reasoning Encourages Human-like WOC Effect As shown in Table 1, without CoT reasoning, the agents' error reduction through social interaction is consistently greater than with CoT reasoning, $\overline{\Delta\varepsilon}$ (without CoT) < $\overline{\Delta\varepsilon}$ (with CoT), difference $= 4.63$, $CI_{95\%} = [2.10, 7, 20]$, $p < .001$. For example, when role-playing detailed persona, the LLM agents (ChatGPT)' error reduction $\overline{\Delta\varepsilon} = -7.59$ when there is no CoT reasoning, as opposed to $\overline{\Delta\varepsilon} = -1.08$ with CoT reasoning, difference $= 6.52$, $CI_{95\%} = [2.59, 10.72]$, $p < .001$. In addition, as shown in Figure 2b, without CoT reasoning always yield a smaller averaged normalized group error $\bar{\varepsilon}_t$ than the counterpart without CoT reasoning.

Detailed Persona and CoT Reasoning Encourages Human-like Partisan Bias The depth of persona detail and the use of CoT reasoning significantly increase the LLM agents' resemblance of human-like partisan bias $\overline{\beta_{\text{PB}}}$ (Table 1). Detailed personas enables a more human-like partisan bias across the two LLMs and across the two CoT reasoning conditions, $\overline{\beta_{\text{PB}}}$ (detailed persona) > $\overline{\beta_{\text{PB}}}$ (simple persona), difference = 15.48, $CI_{95\%} = [14.63, 16.36]$, $p < .001$. On the other

[4] The p-values and 95% Confidence Intervals ($CI_{95\%}$) are derived from bootstrapping with 1000 resamplings (Efron, 1992).

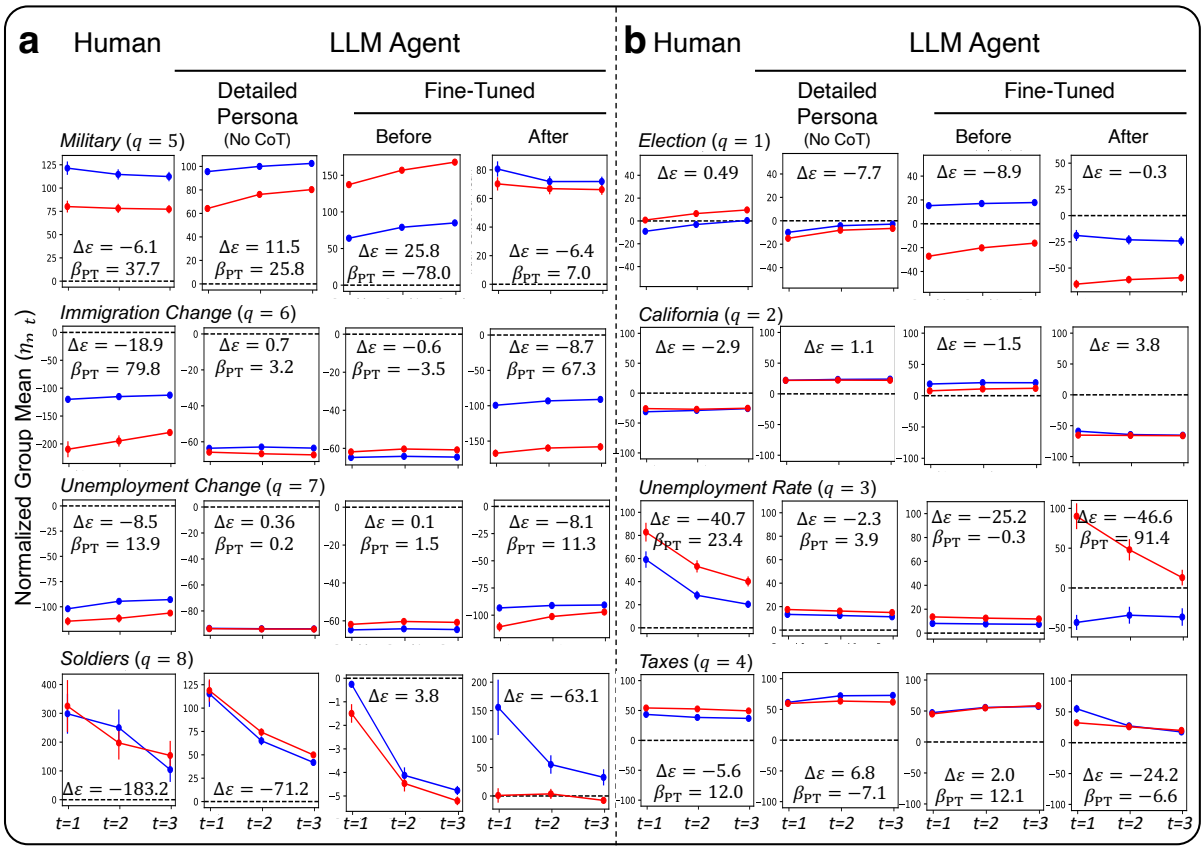

Figure 3: Trajectories of normalized group mean $\eta_{p,t}$ over three rounds, averaged across 12 group experiments (red for Republicans, blue for Democrats), with error bars for standard errors. Each panel consists of four columns representing different data sets: Column 1 shows human data. Columns 2 to 4 shows LLM (ChatGPT) agents' data. Column 2 depicts LLM vanilla role-playing with a detailed persona and without CoT reasoning (the configuration with the highest HLI); Column 3 presents LLM results before fine-tuning; and Column 4 illustrates LLM after fine-tuning. Panel (a) includes questions from the training set ($q \in [5, 8]$) used for fine-tuning the LLM agent, while Panel (b) displays questions from the hold-out test set ($q \in [1, 4]$). Question-specific WOC effects ($\Delta\varepsilon_q$) and partisan biases ($\beta_{\text{PB,q}}$, if expected) are overlaid for comparison.

hand, the use of CoT reasoning also enables a more human-like partisan bias across the two LLMs and across all conditions, $\overline{\beta_{\text{PB}}}$ (CoT) > $\overline{\beta_{\text{PB}}}$ (no CoT), difference = 13.64, $CI_{95\%} = [12.48, 14.78]$, $p < .001$.

LLM Agent's CoT Reasoning is Different from LLM's CoT Reasoning Our results indicate a noticeable difference in the behavior of CoT reasoning when LLM agents are role-playing versus without role-playing. To understand why CoT reasoning fails to improve accuracy, we investigate whether CoT reasoning functions differently when LLMs are engaged in role-playing. Without CoT reasoning, when asked about the unemployment rate during Obama's administration, ChatGPT directly refers to factual data, i.e., *"According to the Bureau of Labor Statistics, the unemployment rate at the end of Barack Obama's presidency was 4.7%"*. In contrast, when an LLM agent role-play as a Democrat or a Republican, their CoT reasoning reflects their persona's perspective. For example, a Republican agent reasons: "I remember hearing a lot about the unemployment rate during Obama's presidency, and I know that it was a big issue. I think it was still pretty high at the end of his term. I'm going to guess that it was around 8% at the end of 2016."

This example illustrates how, in role-playing scenarios, LLM agents' CoT reasoning is more influenced by their personas than factual accuracy.

Results of Role-playing with Induced Partisanship: Introducing explicit partisanship biases into LLM agents does not result in an increase in their Human Likeness Index (HLI) beyond what was achieved by agents role-playing with detailed personas without CoT reasoning (Table 1). While this approach, when paired with CoT reasoning, leads to high partisan bias, $\overline{\beta_{\text{PB}}} = 6.21$,

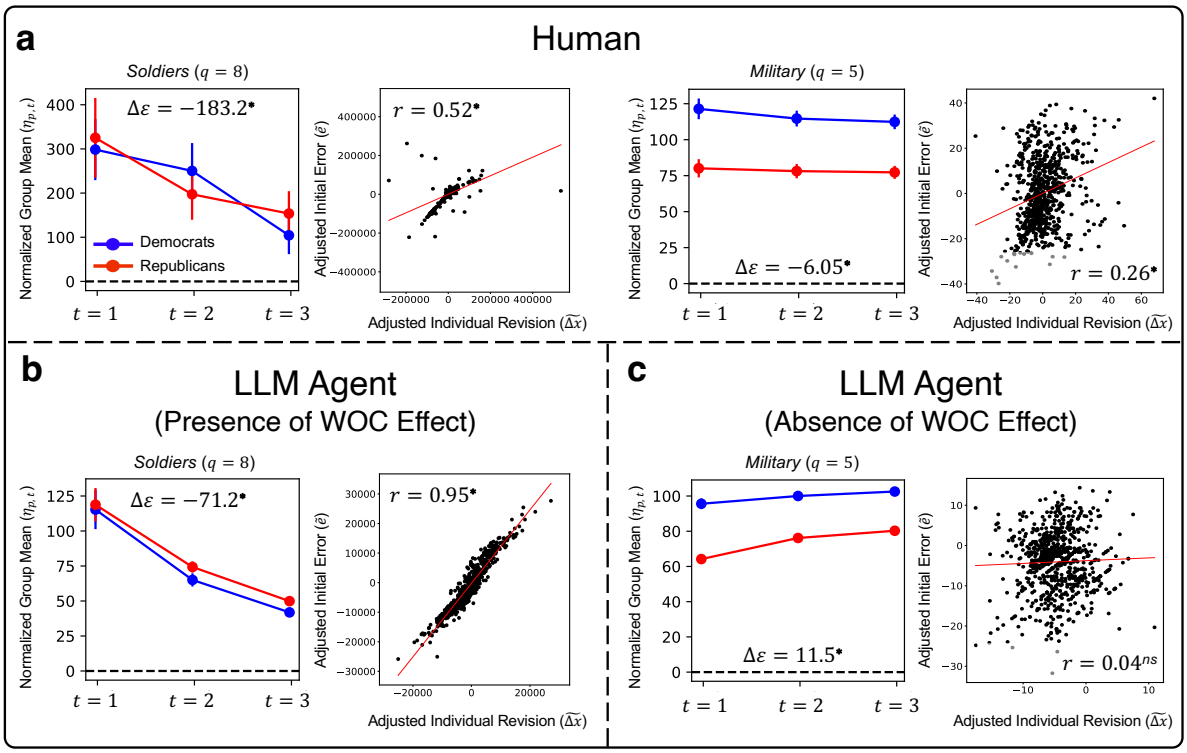

Figure 4: Mechanism of why the WOC effect emerges from crowds of Human crowds and LLM agents. Panel (a) and (b) shows the examples where both humans and LLM agents show the WOC effect through social interaction (i.e., the question-specific WOC effect $\Delta\varepsilon_q < 0$). In contrast, in panel (c), LLM agents do not converge towards the ground truth while humans do. In each panel, the line plot shows the normalized group mean $\eta_{p,t}$ trajectory over three rounds, averaged across 12 runs (red for Republicans, blue for Democrats), with error bars indicating standard errors. The in each panel demonstrate the revision coefficient, the correlation r_{adj} between the adjusted initial individual error $\widetilde{e}_{i,p,r,q}$ and adjusted estimate revisions $\widetilde{\Delta x}_{i,p,r,q}$ (§2.4). Similar to human crowds, the LLM agents show the WOC effect only when $r_{\text{adj}} > 0$. The results of the full set of questions are shown in Figure 5 (§A). *: $p < .01$ (Bonferroni corrected for all questions); ^{ns}: not significant.

$CI_{95\%} = [4.39, 8.02]$, $p < .001$ (ChatGPT) and 20.85, $CI_{95\%} = [18.24, 23.59]$, $\overline{\beta_{\text{PB}}} = p < .001$ (Vicuna), it adversely reduces the resemblance in the WOC effect, $\overline{\Delta\varepsilon} = -2.09$, $CI_{95\%} = [-3.73, -0.48], p = .002$ (ChatGPT) and $\overline{\Delta\varepsilon} = 20.68$, $CI_{95\%} = [11.15, 31.94], p > .05$ (Vicuna). On the other hand, in scenarios without CoT reasoning, introducing partisanship even *reverses* the resemblance of partisan bias, where $\overline{\beta_{\text{PB}}} < 0$ for both LLMs (ChatGPT: $\overline{\beta_{\text{PB}}} = -40.06, CI_{95\%} = [-43.64, -36.33], p < .001$; Vicuna: $\overline{\beta_{\text{PB}}} = -8.46, CI_{95\%} = [-11.45, -5.23], p < .001$. These results suggest that the explicit introduction of partisanship into prompts does not consistently enhance human-like behavior in group dynamics, when compared to vanilla role-playing with detailed persona.

3.2 Impact of Fine-Tuning on Enhancing Human-Like Dynamics

We also investigate whether fine-tuning LLM agents with human response data enhances their resemblance of human group dynamics.

Method	$HLI \uparrow$	$\overline{\Delta\varepsilon} \downarrow$	$\overline{\beta_{\text{PB}}} \uparrow$	*Ext.%*
Before Fine-tuning				
Train	-33.95 ± 1.58	7.27 ± 1.18	-26.68 ± 1.04	0.00
Test	-0.11 ± 0.75	2.31 ± 0.73	2.2 ± 0.14	0.00
After Fine-tuning				
Train	**50.11** ± 6.18	-21.59 ± 6.12	28.53 ± 0.89	0.09
Test	**31.97** ± 3.77	-14.1 ± 3.02	17.87 ± 2.26	**29.94**
Human				
Train	97.95 ± 13.02	-54.15 ± 12.97	43.8 ± 1.20	8.11
Test	29.83 ± 2.21	-12.16 ± 2.07	17.67 ± 0.78	8.64

Table 2: Evaluation of fine-tuned LLM alignment with human group dynamics on the train set ($q \in [5, 8]$) and the test set ($q \in [1, 4]$). The boldface highlights the consequences of fine-tuning.

As shown in Table 2 and Figure 3, in the training set (questions $q \in [5, 8]$), the human likeness index (HLI) increases to 50.11 (from -33.95 before fine-tuning), partisan bias $\overline{\beta_{\text{PB}}}$ increases to 28.53 from -26.68, difference $= 55.20, CI_{95\%} = [52.55, 58.00], p < .001$, and the wisdom of crowds effect $\overline{\Delta\varepsilon}$ changes to -21.59 from 7.27, dif-

ference $= 28.86, CI_{95\%} = [-41.52, -18.44], p < .001$. However, in the test set (questions $q \in [1, 4]$), there is an increase in extreme values (*Ext.%* $= 29.94\%$), indicating a risk of overfitting. For example, the fine-tuned LLM agents tend to provide a negative estimate (up to 84.52%) for unemployment rate estimation ($q = 3$), which is not valid and deemed as extreme values, presumably because there is a similarly worded question about the *change* in unemployment rate where many humans provide negative estimates (46.90% of responses). Nonetheless, after filtering out extreme responses, the fine-tuned models continue to show strong human-like behavior, with an enhanced HLI of 31.97 (increased from 0.11 before tuning), and $\overline{\beta_{\text{PB}}} = -14.1$ (increased from 2.31, difference$=$ $16.42, CI_{95\%} = [-22.10, -10.17], p < .001$) and $\overline{\Delta\varepsilon} = -14.1$ (changed from 2.31, difference$=$ $15.67, CI_{95\%} = [11.37, 20.47], p < .001$). These findings suggest that fine-tuning can greatly enhance the human-like qualities of LLM agents, and even generalize well to unseen question if proper application of filtering criteria is applied.

3.3 Mechanism of the WOC Effect

In human group dynamics, the wisdom of crowds (WOC) effect arises when individuals with initially accurate estimates are less influenced by others, as indicated by a *positive revision correlation coefficient* (Becker et al., 2017). This mechanism is distinct from situations where error reduction is uniform across the group. Our analysis, detailed in Figure 4 and 5 reveals that LLM agents exhibit a similar pattern: the WOC effect occurs ($\Delta\varepsilon_q < 0, ps < .001$[5]) only when the revision coefficient is significantly positive ($r_{adj} > 0$, $ps < .001$). In contrast, when the revision coefficient is not positive, the WOC effect never emerges. In sum, LLM agents' WOC effect emerges through the same mechanism as human crowds, where those with more precise initial estimates exert greater influence on the group's final consensus.

4 Related Work

Simulating Human-like Group Dynamics with LLM-based Agents Research in applying LLM-based agents for social simulation is expanding (Park et al., 2023, 2022; Kaiya et al., 2023; Törnberg et al., 2023; Li et al., 2023). However, these behaviors are not evaluated against actual human data and thus how human-like they are remains unclear. Similarly, Törnberg et al. (2023) simulate social media platforms using LLMs and agent-based modeling to evaluate the effect of different news feed algorithms, yet the comparison with human interactions is also absent. Park et al. (2022) demonstrates that LLM-driven agents can produce *human-like* posts on platforms like Reddit that. However, their work doesn't test the agents' ability to represent demographic-specific behaviors such as political leanings. This highlights a critical gap in the field: evaluate the human likeness of demographic-based LLM agents in social interaction setting.

Wisdom of Crowds Through Social Interaction Research on the wisdom of crowds (WOC) effect has consistently shown that social interaction can refine group estimations in human crowds Becker et al. (2017) demonstrate that collective accuracy increases with information exchange in decentralized networks, and Becker et al. (2019) show that it also manifests in politically homogeneous crowds. In addition, they identify the mechanism for why WOC emerges through social interaction: individuals with larger initial errors adjust their estimates more, contributing to group wisdom. In addition, the effect is robust across cultures (Jayles et al., 2017). It has also been applied to real-world application settings, such as clinical decision-making (Centola et al., 2023) and science communication (Guilbeault et al., 2018). These studies validate the use of social information to improve collective intelligence as a robust benchmark to evaluate human-like behaviors in social interaction context.

5 Conclusion

This study highlights the significant potential of Large Language Models (LLMs) in simulating human group dynamics, particularly in politically sensitive contexts. Our research underscores the influence of persona detail and Chain-of-Thought (CoT) reasoning on the resemblance of LLM agents to human behaviors. Agents with detailed personas, without using CoT reasoning, show a closer alignment with human group dynamics. On the other hand, the use of explicit induced partisanship. Additionally, fine-tuning LLMs with human data show promising results, with a caveat of the risk of overfitting. These findings contribute insights to the development of LLM agents with human-like social bevaviors.

[5] Bonferroni corrected for both the p values for both $\Delta\varepsilon_q < 0$ and $r_{adj} > 0$

Limitations

Two-party political leaning Political orientation often extends beyond mere affiliation with two primary parties. In this research, we focused solely on personas identifying as either Democratic or Republican, a decision influenced by the fact that the human study by (Becker et al., 2019) included participants who supported only these two parties.

Non-exhaustive prompting strategies The current study only uses CoT prompting (Wei et al., 2022b), and doesn't explore other prompting strategies like (Brown et al., 2020; Yao et al., 2023) due to limited time and budget. Previous work has shown the effect of prompting on model performance (Srivastava et al., 2022; Suresh et al., 2023).

Fine-tuning only closed-source models Even thought we evaluate Vicuna-33B, an open-sourced model, in our experiment, we only fine-tuned ChatGPT and did not fine-tune Vicuna-33B.

Task limited to a number guessing game In this experiment, we employed a number guessing game to test the wisdom of crowds effect. This approach was chosen to emulate a similar task conducted previously by (Becker et al., 2019), aiming to investigate how collective estimations or judgments made by a group can lead to accurate or insightful outcomes.

Ethics Statement

We will release the code base exclusively for research purposes.

U.S.-Centric Perspectives The authors, based in the U.S., acknowledge that their perspective, including references to minority groups, is shaped by this context and may not be globally applicable. The data that the study is based on (Becker et al., 2019) also recruited participants in the U.S.

Minority personas induced in LLMs models may not accurately represent real-world views We tried to induce minority personas in LLMs but we acknowledge that the LLMs have inherent biases due to the data that they have been trained on (Christiano et al., 2017; Ziegler et al., 2019; Feng et al., 2023), and it is hard to steer these models to reflect personas of people belonging to minority demographics. Moreover, we also failed to recruit 'LLM agents' in a systematic way to include more diverse demographics.

References

Joshua Becker, Devon Brackbill, and Damon Centola. 2017. Network dynamics of social influence in the wisdom of crowds. *Proceedings of the National Academy of Sciences of the United States of America*, 114(26):E5070.

Joshua Becker, Ethan Porter, and Damon Centola. 2019. The wisdom of partisan crowds. *Proceedings of the National Academy of Sciences of the United States of America*, 116(22):10717–10722.

Marcel Binz and Eric Schulz. 2023. Turning large language models into cognitive models. *arXiv preprint arXiv:2306.03917*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Damon Centola, Joshua Becker, Jingwen Zhang, Jaya Aysola, Douglas Guilbeault, and Elaine Khoong. 2023. Experimental evidence for structured information–sharing networks reducing medical errors. *Proceedings of the National Academy of Sciences*, 120(31):e2108290120.

Harrison Chase. 2022. Langchain.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Bradley Efron. 1992. Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics: Methodology and distribution*, pages 569–593. Springer.

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

Douglas Guilbeault, Joshua Becker, and Damon Centola. 2018. Social learning and partisan bias in the interpretation of climate trends. *Proceedings of the National Academy of Sciences*, 115(39):9714–9719.

Bertrand Jayles, Hye-rin Kim, Ramón Escobedo, Stéphane Cezera, Adrien Blanchet, Tatsuya Kameda, Clément Sire, and Guy Theraulaz. 2017. How social information can improve estimation accuracy in human groups. *Proceedings of the National Academy of Sciences*, 114(47):12620–12625.

Zhao Kaiya, Michelangelo Naim, Jovana Kondic, Manuel Cortes, Jiaxin Ge, Shuying Luo, Guangyu Robert Yang, and Andrew Ahn. 2023. Lyfe agents: Generative agents for low-cost real-time social interactions. *arXiv preprint arXiv:2310.02172*.

Chao Li, Xing Su, Chao Fan, Haoying Han, Cong Xue, and Chunmo Zheng. 2023. Quantifying the impact of large language models on collective opinion dynamics. *arXiv preprint arXiv:2308.03313*.

OpenAI. 2022. Introducing ChatGPT. https://openai.com/blog/chatgpt. [Accessed 13-10-2023].

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 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.

Joon Sung Park, Lindsay Popowski, Carrie Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2022. Social simulacra: Creating populated prototypes for social computing systems. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, pages 1–18.

Omar Shaikh, Hongxin Zhang, William Held, Michael Bernstein, and Diyi Yang. 2022. On second thought, let's not think step by step! bias and toxicity in zero-shot reasoning. *arXiv preprint arXiv:2212.08061*.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

Siddharth Suresh, Kushin Mukherjee, Xizheng Yu, Wei-Chun Huang, Lisa Padua, and Timothy T Rogers. 2023. Conceptual structure coheres in human cognition but not in large language models.

Petter Törnberg, Diliara Valeeva, Justus Uitermark, and Christopher Bail. 2023. Simulating social media using large language models to evaluate alternative news feed algorithms. *arXiv preprint arXiv:2310.05984*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022a. Emergent abilities of large language models. *Transactions on Machine Learning Research*. Survey Certification.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed H Chi, Quoc V Le, Denny Zhou, et al. 2022b. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.

A Results of Revision Coefficient

Figure 5 detailed in the analysis of revision coefficients. Specifically, the wisdom of crowds effect emerges only when the revision coefficient is positive, mirroring human behavior. Like human crowds, LLM agents enhanced collective accuracy when individuals with initially more accurate estimates are less influenced by peer opinions, thereby steering the group toward more accurate collective estimates.

B Vicuna Results

Figure 3 shows the result for human versus ChatGPT-based LLM agents. Here, we shows the results for human versus Vicuna-based LLM agents in Figure 6. As shown, Vicuna-based LLM agents show the WOC effect ($\overline{\Delta\varepsilon} < 0$) for all questions except the question $q = 6$, $p < .001$ (Bonferroni corrected).

C List of Personas

C.1 Detailed Persona Condition

In this section, we list the personas of the agents interacting in our experiment. 35 personas were used in the experiment for both Democrats and Republicans. For brevity purposes, a subset of them is included here. The full set of personas will be released along with the codebase at the time of publication. The list of personas is generated using this prompt (PARTY is either Republican or Democrat) with GPT4 (OpenAI, 2022):

```
Create detailed descriptions for 35
individuals who identify as PARTY
and are above the age of 18, with
varying degrees of political leaning
(from Lean PARTY to Strong PARTY).
Ensure that their demographics
are diverse and reflective of
the actual distribution of the US
population, including factors such as
age, gender, ethnicity, education
level, and occupation.  Provide
comprehensive background information
for each person.  Use the format
below.  Name:[] Political leaning:
[] Age:  [] Gender:  [] Ethnicity:
[] Education:  [] Occupation:  []
Background:  []
```

Democrats

Name: Isabella Johnson
Political leaning: Strong Democrat
Age: 67
Gender: Female
Ethnicity: White
Education: Bachelor's Degree in Education
Occupation: Retired Teacher
Background: Isabella is from Portland, Oregon, and spent her career advocating for public education and teachers' rights. She is passionate about social justice, healthcare, and environmental issues. Isabella is widowed with two grown children and enjoys birdwatching and painting in her free time.

Name: Jamal Brown
Political leaning: Lean Democrat
Age: 51
Gender: Male
Ethnicity: African American
Education: Bachelor's Degree in Finance
Occupation: Financial Analyst
Background: Jamal grew up in Detroit, Michigan, and became politically active during the 2008 recession. He supports policies promoting economic fairness and equal opportunities. Jamal is divorced with one child and enjoys playing golf and attending jazz concerts.

Name: Karen Patel
Political leaning: Moderate Democrat
Age: 34
Gender: Female
Ethnicity: Indian American
Education: Master's Degree in Computer Science
Occupation: Software Engineer
Background: Karen was born in San Jose, California, and is a strong advocate for affordable housing and tech industry regulations. She also supports women's rights and STEM education. Karen is married with two young children and enjoys hiking and coding projects in her free time.

Name: Larry Jackson

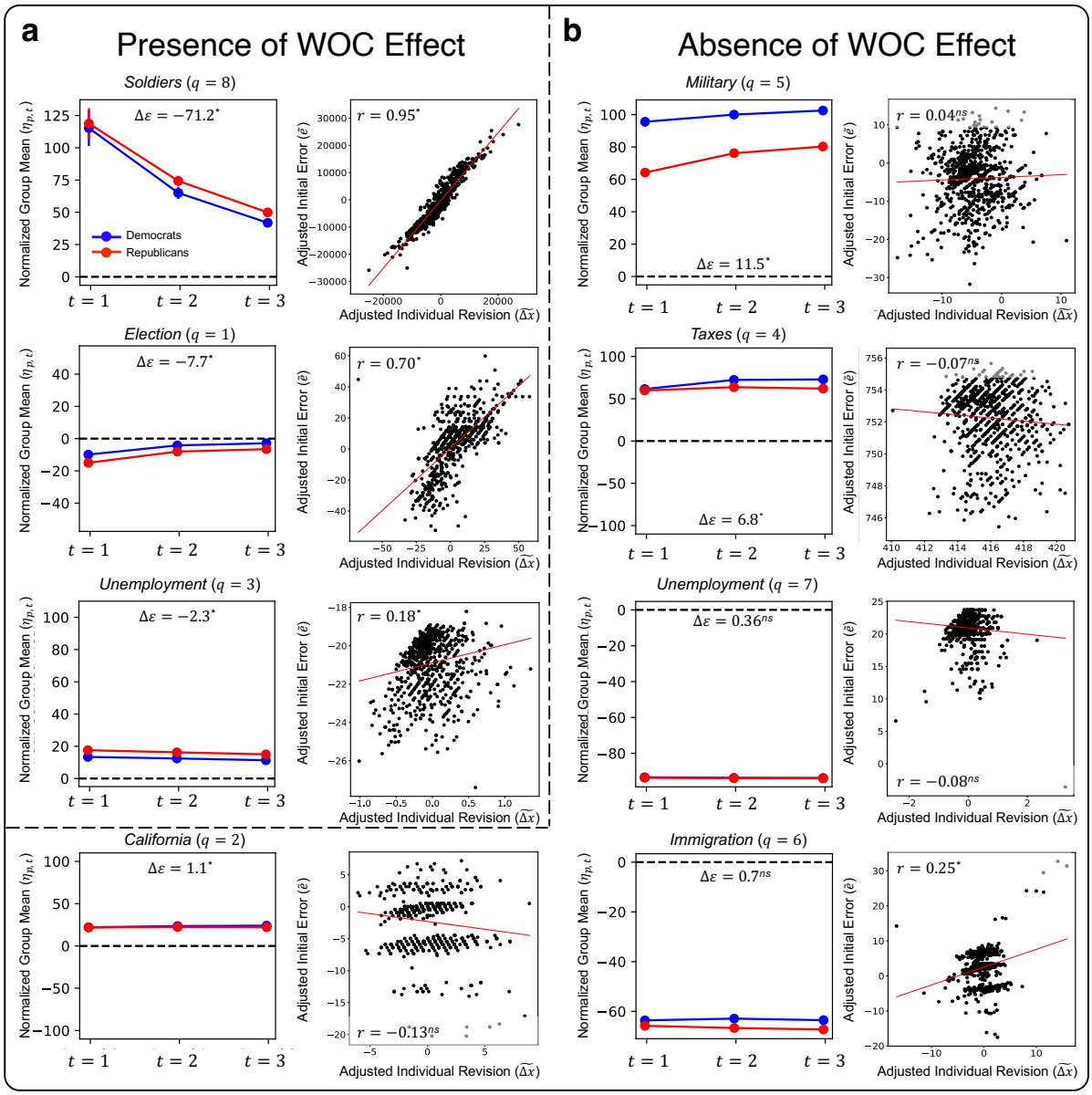

Figure 5: Analysis of the mechanism of LLM agents' wisdom of crowds (WOC) effect at the individual level. Panel (a) shows the questions where LLM agents exhibit the WOC effect ($\Delta\varepsilon_q < 0$). Panel (b) shows the questions where LLM agents do not show the WOC effect. Within each panel, the questions are ordered by their revision correlation r_{adj}. In each panel, the line plot shows the normalized group mean $\eta_{p,t}$ trajectory over three rounds, averaged across 12 runs (red for Republicans, blue for Democrats), with error bars indicating standard errors. The in each panel demonstrate the revision coefficient, the correlation r_{adj} between the adjusted initial individual error $\widetilde{e}_{i,p,r,q}$ and adjusted estimate revisions $\widetilde{\Delta x}_{i,p,r,q}$ (§2.4). The LLM agents show the WOC effect only when $r_{\text{adj}} > 0$ (panel a). *: $p < .01$ (Bonferroni corrected for all questions); ^{ns}: not significant.

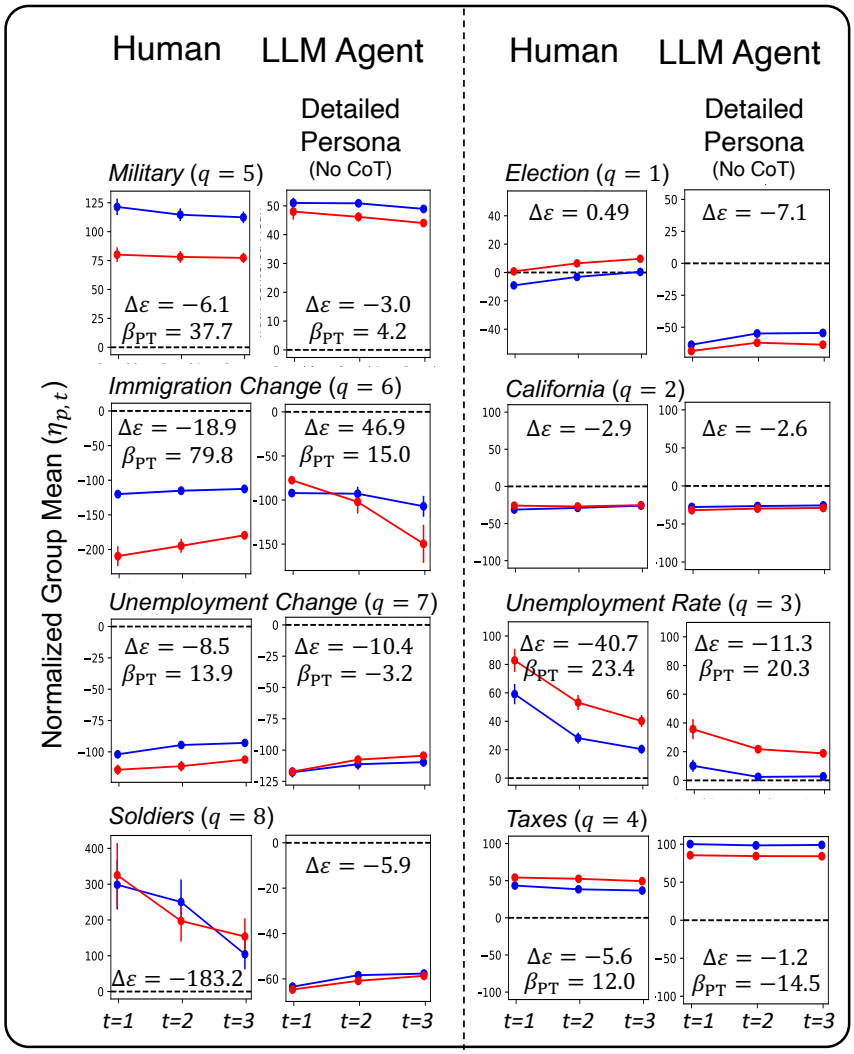

Figure 6: Trajectories of normalized group mean $\eta_{p,t}$ over three rounds, averaged across 12 group experiments (red for Republicans, blue for Democrats), with error bars for standard errors. Each panel consists of two columns representing different data sets: Column 1 shows human data. Columns 2 shows LLM (Vicuna) agents' data. We separate questions into two panels to facilitate comparison with Figure 3, where panel (a) includes questions $q \in [5, 8]$, and (b) displays questions $q \in [1, 4]$. Question-specific WOC effects ($\Delta\varepsilon_q$) and partisan biases ($\beta_{\text{PB,q}}$, if expected) are overlaid for comparison.

Political leaning: Strong Democrat
Age: 42
Gender: Male
Ethnicity: White
Education: Bachelor's Degree in Sociology
Occupation: Nonprofit Fundraiser
Background: Larry hails from Madison, Wisconsin, and is deeply involved in local politics. He is passionate about income inequality, racial justice, and LGBTQ+ rights. Larry is married with no children and enjoys traveling and volunteering for progressive causes.

Name: Monica Rodriguez
Political leaning: Lean Democrat
Age: 48
Gender: Female
Ethnicity: Puerto Rican
Education: Associate Degree in Nursing
Occupation: Registered Nurse
Background: Monica grew up in New York City, New York, and supports policies that improve healthcare access and quality. She is also concerned about affordable housing and education reform. Monica is a single mother of two and enjoys salsa dancing and cooking.

Republicans

Name: Charles Matthews
Political leaning: Lean Republican
Age: 38
Gender: Male
Ethnicity: African American
Education: Bachelor's degree in Business Administration
Occupation: Small business owner
Background: Charles was born and raised in Atlanta, Georgia, where he attended a public university before starting his own business. He has a wife and two children. Charles supports limited government intervention and lower taxes, and he values entrepreneurship and self-reliance.

Name: Alice Thompson
Political leaning: Moderate Republican
Age: 29
Gender: Female
Ethnicity: Caucasian
Education: Master's degree in Public Health
Occupation: Epidemiologist
Background: Alice grew up in a small town in Ohio before moving to Philadelphia for her studies. She is passionate about public health and believes in data-driven policies. She leans Republican due to her fiscal conservatism and support for individual rights.

Name: Juan Rodriguez
Political leaning: Strong Republican
Age: 45
Gender: Male
Ethnicity: Hispanic
Education: High school diploma
Occupation: Construction worker
Background: Juan, originally from Mexico, migrated to Texas with his family when he was a child. A father of three, he believes in traditional family values, hard work, and limited government. He is a staunch advocate for securing the nation's borders.

Name: Sarah Chang
Political leaning: Lean Republican
Age: 23
Gender: Female
Ethnicity: Asian American
Education: Bachelor's degree in Environmental Science
Occupation: Environmental Consultant
Background: Sarah was born and raised in California. She supports free-market solutions to environmental issues and believes in responsible resource management. Sarah leans Republican due to her fiscally conservative views and her opposition to excessive government regulation.

Name: Robert Klein

```
Political leaning: Moderate Republican
Age: 64
Gender: Male
Ethnicity: Caucasian
Education: Bachelor's degree in Engineering
Occupation: Retired engineer
Background: Robert, a native of Pennsylvania, worked for a major manufacturing company for over 30 years. He is a Vietnam War veteran and a strong supporter of the Second Amendment. Robert believes in fiscal responsibility, limited government, and a strong national defense.
```

C.2 Simple Persona Condition

Democrats

```
A typical Democrat in the USA.
```

Republicans

```
A typical Republican in the USA.
```

D Full List of Questions

Below is the full list of questions ($q \in [1, 8]$) in the experiment, along with the ground truth x_q^* and the sign of human partisan bias direction $\text{sign}(h_q)$ observed in human data (Becker et al., 2019).

1. In the 2004 election, individuals gave $269.8 million to Republican candidate George W. Bush. How much did they give to Democratic candidate John Kerry? (Answer in millions of dollars—e.g., 1 for $1 million.) [$x_1^* = 224.6$, $\text{sign}(h_1) = 0$]

2. According to 2010 estimates, what percentage of people in the state of California identify as Black/African-American, Hispanic, or Asian? (Give a number from 0 to 100.) [$x_2^* = 60.2$, $\text{sign}(h_2) = 0$]

3. What was the US unemployment rate at the end of Barack Obama's presidential administration—i.e., what percentage of people were unemployed in December 2016? (Give a number from 0 to 100.) [$x_3^* = 4.9$, $\text{sign}(h_3) = 1$]

4. In 1980, tax revenue was 18.5% of the economy (as a proportion of GDP). What was tax revenue as a percentage of the economy in 2010? (Give a number from 0 to 100.) [x_4^* : 14.6, $\text{sign}(h_4) = 1$]

5. For every dollar the federal government spent in fiscal year 2016, about how much went to the Department of Defense (US Military)? Answer with a number between 0 and 100. [$x_5^* = 15$, $\text{sign}(h_5) = -1$]

6. In 2007, it was estimated that 6.9 million unauthorized immigrants from Mexico lived in the United States. How much did this number change by 2016, before President Trump was elected? Express your answer as a percentage of change. [$x_6^* = -27.8$, $\text{sign}(h_6) = -1$]

7. How much did the unemployment rate in the United States change from the beginning to the end of Democratic President Barack Obama's term in office? Express your answer as a percentage of change. [$x_7^* = -46$, $\text{sign}(h_7) = -1$]

8. About how many US soldiers were killed in Iraq between the invasion in 2003 and the withdrawal of troops in December 2011? [$x_78* = 4486$, $\text{sign}(h_8) = 0$]

E List of Induced Partisanship

Below is the full list of induced partisanship corresponding to the questions in the experiment ($q \in [1, 8]$). We did not induce a bias for certain questions because (Becker et al., 2019) found that there is no significant human partisan bias for those questions.

1. Did not induce a bias for this question.

2. Did not induce a bias for this question.

3. Republicans tend to estimate a higher unemployment rate than democrats.
   (a) **Democrats:** "*You believe that Barack Obama did a good job in reducing the US unemployment rate.*"
   (b) **Republicans:** "*You believe that Barack Obama did a poor job in reducing the US unemployment rate.*"

4. Republicans tend to estimate higher tax revenue than democrats.
   (a) **Democrats:** "*You believe that tax rates are not as high as they should be in general.*"
   (b) **Republicans:** "*You believe that tax rates are too high in general.*"

5. Democrats tend to estimate a higher military budget than Republicans.
   (a) **Democrats:** "*You believe that the US federal budget spent on the US military is too high in general.*"
   (b) **Republicans:** "*You believe that the US federal budget spent on the US military is not as high as it should be in general.*"
6. Republicans tend to estimate higher number of immigrants than democrats
   (a) **Democrats:** "*You believe that the unauthorized immigrants from Mexico were not a major national crisis before President Trump was elected.*"
   (b) **Republicans:** "*You believe that the US federal budget spent on the US military is not as high as it should be in general.*"
7. Republicans tend to estimate a higher unemployment rate than democrats.
   (a) **Democrats:** "*You believe that Barack Obama did a good job in reducing the US unemployment rate.*"
   (b) **Republicans:** "*You believe that Barack Obama did a poor job in reducing the US unemployment rate.*"
8. Did not induce a bias for this question.

F Full list of Prompts

In this section, we detail the prompts we use at each time step in the experiments.

```
Round 1: Role play this person.
{AGENT_PERSONA}
{INDUCED_BIAS}

Let's play a game where you'll be asked a single question, and you must provide an answer. This game has 3 trials, allowing you 2 chances to revise your response. Keep in mind that it's a group game, played concurrently with other participants. After you submit your first answer, you'll be given the average of other players' initial responses. Following your second submission, you'll receive the average of their second-round responses. At the end of the game, the more accurate your final answer is compared to the actual truth, the more money you will earn.

Now, {AGENT_NAME}, in this first round of the game, you are asked to answer the following question.

{QUESTION_CONTENT}

You must give an answer even if you are not sure.

Use the following format:

My Reasoning: [YOUR ({AGENT_NAME}'s) STEP-BY-STEP REASONING]

My Final Answer: [YOUR ({AGENT_NAME}'s) ESTIMATE (A REAL NUMBER)]
```

```
Round 2: Now, {AGENT_NAME}, in this second round of the game, you are asked again to answer the following question.

{QUESTION_CONTENT}

You must give an answer even if you are not sure.

This time, you are provided with other players in the first round of the game, who are all {AGENT_PARTY}. Their average answer: {FEEDBACK}

Use the following format:

My Reasoning: [YOUR ({AGENT_NAME}'s) STEP-BY-STEP REASONING]

My Final Answer: [YOUR ({AGENT_NAME}'s) ESTIMATE (A REAL NUMBER)]
```

```
Round 3: Now, {AGENT_NAME}, in
```

```
this third round of the game, you
are asked again to answer the
following question.

{QUESTION_CONTENT}

You must give an answer even if
you are not sure.

This time, you are provided
with other players in the first
round of the game, who are all
{AGENT_PARTY}.  Their average
answer:  {FEEDBACK}

Use the following format:

My Reasoning: [YOUR
({AGENT_NAME}'s) STEP-BY-STEP
REASONING]

My Final Answer:  [YOUR
({AGENT_NAME}'s) ESTIMATE (A
REAL NUMBER)]
```

In the prompt outline above, the highlighted sections are subject to change based on the configuration of the prompt:

AGENT_PERSONA: If we are using the detailed personas condition for a prompt, this placeholder is replaced with one of the descriptions from §C.

INDUCED_BIAS: This placeholder is replaced with a bias corresponding to the party of the agent and the question. A list of these biases can be found in §E.

AGENT_NAME: For the detailed persona condition, this placeholder is derived from the agent persona. For the simple persona condition, this is replaced with a generic name such as "*r_1*" for republicans and "*d_1*" for democrats.

My Reasoning [YOUR (AGENT_NAME's) STEP-BY-STEP REASONING]: If the agent do not use chain-of-thought (CoT) reasoning in its responses, this section is excluded from the prompt at each time step. otherwise, it is included.

AGENT_PARTY: This placeholder is replaced with either "*democrats*" or "*republicans*" depending on the agent's party affiliation.

FEEDBACK: For the second and third round prompts, this placeholder is replaced with the mean response of the agent's neighbours for the previous round.

QUESTION_CONTENT: This placeholder is replaced with one of the question from §D.

G Criteria for Extreme Values

For all questions in the experiment ($q \in [1, 8]$), we use the same criteria as in the human study (Becker et al., 2019)[6]. The criteria are as follows:

1. Election donations for John Kerry in 2004: Answers above \$2246 million ($10 \times x^*$) or below 0 are considered extreme.

2. California demographics: Answers above 602% ($10 \times x^*$) or below 0% are considered extreme.

3. Unemployment rate at the end of Obama's presidency: Answers above 49% ($10 \times x^*$) or below 0% are considered extreme.

4. Tax revenue as a percentage of GDP in 2010: Answers above 146% ($10 \times x^*$) or below 0% are considered extreme.

5. Military spending: Answers above 100% or below 0% are considered extreme.

6. Immigration rate changes: Answers above 1000% or below -1000% are extreme.

7. Unemployment rate changes: Answers above 1000% or below -1000% are extreme.

8. Soldier deaths: Answers above 1 million and answers below 0 are extreme.

These criteria are based on the realistic ranges expected for these measures, as per the guidelines in Becker et al. (2019).

H Fine-tuning Details

OpenAI's ChatGPT (`gpt-3.5-turbo-0613`) was fine-tuned using the human data from (Becker et al., 2019) to investigate the change in human-AI alignment in a group interaction setting. Two separate models were fine-tuned: one for Democrats and another one for Republicans.

These were the hyper-parameters used in the fine-tuning:

- Training Set Size: 2747

[6] Becker et al. (2019) use a log function to normalize the responses for $q \in [1, 4]$ without removing extreme values. In our study, we do not do this since LLM agents may occasionally return negative values. Instead, we follow their "alternative normalization procedire" outlined in their supplementary materials. They show that different normalization procedures do not yield significant differences.

- Validation Set Size: 381
- Number of Epochs: 4
- Batch Size: 5
- Learning Rate Decay Factor: 0.05

Human data was collected for 12 groups. For each agent $a_{i,p,r}$, a list of 4 neighbours $\mathcal{N}(i,p,r)$ was provided. We recreated the feedback provided to agents in rounds 2 and 3 by taking the average of their neighbours' estimates in the previous round, given by $m^t_{i,p,r,q} = \frac{1}{4}\sum_{j\in\mathcal{N}(i,p,r)} x^{t-1}_{j,p,r,q}$. The data for each agent was separated into 3 prompt-response pairs, corresponding to a specific round in the experiment. Each pair included the prompt and its corresponding response from the respective round, as well as those from earlier rounds. These pairs were passed into OpenAI's fine-tuning API and the experiment was run on the resulting models.

I Compute Resources

For running the Vicuna-33B model, we use a Linux server with 1 GPU card (A100-SXM4-80GB), taking about 12 hours to run the full experiment

J Detailed Definition of Notations

This appendix section elaborates on the specific definitions and the derivation process of key metrics and terms used throughout our study.

J.1 Reduction in Group Error Through Social Interaction

The primary indicator of human-like performance in this group experiment setting is the ability to improve estimates over time through social interaction, as per (Becker et al., 2019). To quantify this effect, we define several key terms:

Group Mean ($\bar{x}^t_{p,r,q}$): For each run r and each political leaning p, the group mean at each time step t for question q is denoted as $\bar{x}^t_{p,r,q}$. It represents the average of the estimates provided by all agents in the group at that time step and is expressed as $\bar{x}^t_{p,r,q} = \frac{1}{N}\sum_{i=1}^{N} x^t_{i,p,r,q}$, where $N = 35$ is the total number of agents in the group.

Group Error ($\delta^t_{p,r,q}$): The group error for each run r, political leaning p, and time step t for question q is denoted as $\delta^t_{p,r,q}$. This metric is calculated as the difference between the group mean and the ground truth value: $\delta^t_{p,r,q} = |\bar{x}^t_{p,r,q} - x^*_q|$. It represents the raw error of the group's collective estimate in comparison to the actual value for each question.

Normalized Group Mean ($\eta^t_{p,r,q}$): The normalized group mean is defined as $\eta^t_{p,r,q} = 100 \times (\bar{x}^t_{p,r,q} - x^*_q)/(x^*_q)$, where x^*_q is the ground truth value for question q. This metric provides a relative measure of the group's average estimate compared to the ground truth, normalized by the scale of each question. We further scale by 100 to express the group mean as the *percentage* of ground truth x^*; $+100$ means overestimating by one unit of x^*, and -100 means underestimating by one unit of x^*. Note that, unlike Becker et al. (2019), who use log functions for normalization for some questions, we adopt their "alternative normalization procedure" due to occasional negative values from LLM agents, as outlined in their supplementary materials. They demonstrate that various normalization methods yield similar results.

Normalized Group Error ($\varepsilon^t_{p,r,q}$): The normalized group error, $\varepsilon^t_{p,r,q}$, measures the absolute deviation of the normalized group mean from the ground truth value and is defined as $\varepsilon^t_{p,r,q} = |\eta^t_{p,r,q}| = 100 \times |(\bar{x}^t_{p,r,q} - x^*_q)/x^*_q|$. Note that we also scale by 100, so $\varepsilon^t_{p,r,q}$ can be interpreted as the error in terms of the percentage of the ground truth value x^*. For example, $\varepsilon^t_{p,r,q} = 50$ means that the group mean is deviating from the ground truth x^* by 50% of $|x^*|$.

Average Normalized Group Error ($\overline{\varepsilon_t}$): To evaluate group performance across political leanings, runs, and questions, we calculate the *average normalized group error*, denoted as $\overline{\varepsilon_t}$. This metric represents the average normalized group error at a specific time step t. Formally, $\overline{\varepsilon_t} = \frac{1}{P\cdot R\cdot Q}\sum_p\sum_{r=1}^{R}\sum_{q=1}^{Q}\varepsilon^t_{p,r,q}$, where $P = 2$ represents the number of political leanings, R is the total number of runs, and $Q = 8$ is the number of questions.

Group Error Reduction ($\Delta\varepsilon_{p,r,q}$): The change in normalized group error from the initial to the final estimate is quantified as $\Delta\varepsilon_{p,r,q} = (\varepsilon^{t=3}_{p,r,q} - \varepsilon^{t=1}_{p,r,q})$, indicating error reduction for each run. Note that because $\varepsilon^t_{p,r,q}$ is already scaled by 100, $\Delta\varepsilon_{p,r,q}$ can be interpreted as the error reduction in terms of the percentage of the ground truth value x^*. For example, $\Delta\varepsilon_{p,r,q} = -50$ means that the group error $\delta^t_{p,r,q}$ reduces by 50% of the size of the ground truth $|x^*|$. In other words, the group mean

$x^t_{p,r,q}$ moves towards the ground truth x^* by 50% of the size of the ground truth $|x^*|$)

Average Group Error Reduction ($\overline{\Delta\varepsilon}$): To quantify the change in group error over the course of the experiment, we calculate the *average group error reduction*, denoted as $\overline{\Delta\varepsilon}$. This is calculated by averaging $\Delta\varepsilon_{p,r,q}$ across all political leanings, runs, and questions. Formally, $\overline{\Delta\varepsilon} = \frac{1}{P \cdot R \cdot Q} \sum_p \sum_{r=1}^{R} \sum_{q=1}^{Q} \Delta\varepsilon_{p,r,q}$, where $P = 2$ represents the number of political leanings, $R = 12$ the total number of runs, and $Q = 8$ the number of questions. $\overline{\Delta\varepsilon}$ reflects the *wisdom of partisan crowds* effect in LLM agents. A more negative value of $\overline{\Delta\varepsilon}$ indicates a stronger wisdom of partisan crowds effect.

J.2 Deriving the Revision Coefficient

To compute the revision coefficient, we follow the methodology outlined by Becker et al. (2017), focusing on the relationship between individual revisions, initial errors, and the social signal. The process involves two main steps: regression to obtain residuals and computing the partial correlation.

Obtaining Residuals: The first step involves using Ordinary Least Squares (OLS) regression to adjust individual revisions and errors based on the social signal. This is done to isolate the effect of an individual's initial accuracy on their subsequent revision, independent of the strength of the social influence they experience.

1. **Adjusting Individual Revisions:** For each question q, across all individual i, political leaning p, and run r, we perform a regression of their individual revision $\Delta x_{i,p,r,q}$ against the social signal $s_{i,p,r,q}$. For a given question q, the regression equation is:

$$\Delta x_{i,p,r,q} = a_{1,q} \cdot s_{i,p,r,q} + b_{1,q} + \epsilon_{1,i,p,r,q}$$

   where $a_{1,q}$ and $b_{1,q}$ are regression coefficients, and $\epsilon_{1,i,p,r,q}$ is the residual. The residual $\epsilon_{1,i,p,r,q}$ represents the adjusted individual revision $\widetilde{\Delta x}_{i,p,r,q} = \epsilon_{1,i,p,r,q}$.

2. **Adjusting Initial Errors:** Similarly, we regress the individual initial error $e_{i,p,r,q}$ against the same social signal:

$$e_{i,p,r,q} = a_{2,q} \cdot s_{i,p,r,q} + b_{2,q} + \epsilon_{2,i,p,r,q}$$

   where $a_{2,q}$ and $b_{2,q}$ are coefficients, and $\epsilon_{2,i,p,r,q}$ is the residual. The residual $\epsilon_{2,i,p,r,q}$ becomes the adjusted initial error $\widetilde{e}_{i,p,r,q} = \epsilon_{2,i,p,r,q}$.

Computing the Partial Correlation: With the residuals obtained, the revision coefficient ($r_{\text{adj},q}$) is computed as the Pearson correlation between the adjusted individual revisions and adjusted initial errors:

$$\begin{aligned} r_{\text{adj},q} &= \text{corr}(\Delta\widetilde{x}_{i,p,r,q}, \widetilde{e}_{i,p,r,q}) \\ &= \text{corr}(\epsilon_{1,i,p,r,q}, \epsilon_{2,i,p,r,q}) \end{aligned}$$

This partial correlation reflects the extent to which individuals with higher initial accuracy are less influenced by social signals in revising their estimates, after controlling for the strength of the social signal itself.