Title

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

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 46(0)

Authors

Chuang, Yun-Shiuan

Harlalka, Nikunj

Suresh, Siddharth

et al.

Publication Date

2024

Peer reviewed

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

Yun-Shiuan Chuang, Nikunj Harlalka[†], Siddharth Suresh[†], Agam Goyal
Robert Hawkins, Sijia Yang, Dhavan Shah, Junjie Hu, Timothy T. Rogers
University of Wisconsin-Madison
{yunshiuan.chuang,siddharth.suresh,nharlalka,agoyal25}@wisc.edu
{rdhawkins, syang84, dshah, junjie.hu, ttrogers}@wisc.edu
[†] equal contribution

Abstract

Human groups are able to converge to more accurate beliefs through deliberation, even in the presence of polarization and partisan bias — a phenomenon known as the "wisdom of partisan crowds." Large Language Models (LLMs) are increasingly being used to simulate human collective behavior, yet few benchmarks exist for evaluating their dynamics against the behavior of human groups. In this paper, we examine the extent to which the wisdom of partisan crowds emerges in groups of LLM-based agents that are prompted to role-play as partisan personas (e.g., Democrat or Republican). We find that they not only display human-like partisan biases, but also converge to more accurate beliefs through deliberation, as humans do. We then identify several factors that interfere with convergence, including the use of chain-of-thought prompting and lack of details in personas. Conversely, fine-tuning on human data appears to enhance convergence. These findings show the potential and limitations of LLM-based agents as a model of human collective intelligence.

Keywords: wisdom of crowds; partisan bias; generative agents; Large Language Models

Introduction

When groups of people work together to solve problems or make predictions, they are often able to arrive at more effective solutions than any individual alone. In a classic example, Galton (1907) analyzed a contest where people were asked to estimate the weight of an ox. While each individual estimate was poor, the group's average was remarkably close to the true value. This phenomenon, known as the *wisdom of crowds*, is a paradigmatic example of collective intelligence (Kameda et al., 2022; Yi et al., 2012). Moreover, when individuals are shown the average estimate of their group and allowed to adjust their own, the group's average becomes *more accurate* (Becker et al., 2017), even for biased groups (termed "wisdom of partisan crowds"; Becker et al., 2019). The effect where social influence further improves collective estimates, extends across different cultures (Jayles et al., 2017), and finds application in practical domains such as clinical decision-making (Centola et al., 2023) and science communication (Guilbeault et al., 2018).

Large language models (LLMs) have displayed increasingly sophisticated social behaviors, raising questions about the extent to which they can serve as models of human communication in social groups (Park et al., 2022; Chuang et al., 2023; Törnberg et al., 2023; Kaiya et al., 2023; Li et al., 2023). For 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. Yet, without direct comparisons to empirical benchmarks of human behavior, it has been difficult to understand how human-like such patterns really are, and consequently how useful such simulated systems are for understanding human communicative phenomena.

We suggest that the *wisdom of (partisan) crowds* may serve as an effective benchmark for LLM-based agents' collective behavior. In particular, we consider the data from Becker et al. (2019), where $N = 1,020$ participants were asked factual questions known to elicit partisan bias (e.g. estimating the US employment rate during Barack Obama's presidency). Self-identified Republicans and Democrats generated systematically different guesses reflecting their political leanings. After they were shown the average belief of others in their own partisan group, they were then allowed to adjust their estimate. Surprisingly, *both* groups adjusted their estimates in ways that move the group mean systematically closer to the ground truth, despite their initial bias.

This wisdom of partisan crowd phenomenon is useful for assessing LLM simulation of human communication for three reasons. First, all questions have a ground-truth value, providing a means of quantifying how accurate the estimates were. Second, humans typically show partisan lean in their estimates. This provides an opportunity to evaluate whether role-playing LLMs show 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.

To this end, we replicate the experimental design of Becker et al. (2019), apply it to groups of interacting, role-playing LLM agents in a simulated environment, and assess whether the resulting system replicated the key phenomena in human behaviors. We found that LLM agents, when operating without Chain-of-Thought (CoT) reasoning (Wei, Wang, et al., 2022), exhibit a wisdom of partisan crowds effect, paralleling human patterns of group error reduction. 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

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

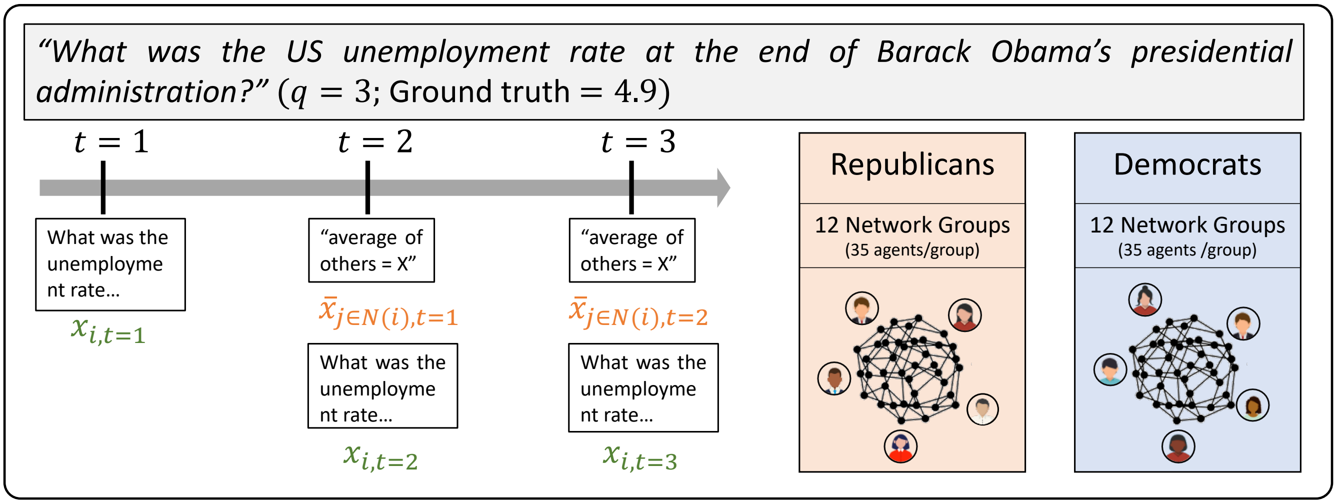
“What was the US unemployment rate at the end of Barack Obama’s presidential administration?” ($q = 3$; Ground truth = 4.9)
$t = 1$
$t = 2$
$t = 3$
What was the unemployment rate...
$x_{i,t=1}$
“average of others = X”
$\bar{x}_{j\in N(i),t=1}$
What was the unemployment rate...
$x_{i,t=2}$
“average of others = X”
$\bar{x}_{j\in N(i),t=2}$
What was the unemployment rate...
$x_{i,t=3}$
Republicans
12 Network Groups
(35 agents/group)
Democrats
12 Network Groups
(35 agents /group)

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

with human data enhances human-like group dynamics in held-out data, although such training also risks overfitting. In sum, we show that with proper prompt engineering and fine-tuning, we can encourage LLM agents to emulate human-like social interactions.

Methods

Procedure We followed the design from 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 $K = 4$ neighbors with the same political leaning (Figure 1) to reflect homogeneous group structures in human studies. We used LangChain (Chase, 2022) with ChatGPT (`gpt-3.5-turbo`; OpenAI, 2022) and the open-source LLM Vicuna (`vicuna-33B-v1.3`; Zheng et al., 2023). 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 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 entire procedure was carried out 12 times.[2]

Formal notation We denote each agent in the experiment as $a_{i,p,r}$, where $1 \leq i \leq 35$ 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 $1 \leq r \leq 12$. When the context is clear, we drop the subscript p and r. For each political leaning, during each run, the agents answer eight questions over three time steps, generating a series of estimates $x^t_{i,q}$ 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 $m^t_{i,q}$, the average estimate of their four politically homogeneous neighbors, before making their own estimates.[3]

Personas and agent specification We prompted the LLMs to role-play as different personas created with varying levels of background detail. *Simple Personas* are specified as “a typical Democrat/Republican,” relying on temperature sampling to elicit slightly different biased views. *Detailed Personas* are provided with comprehensive backstories, including demographics and personal background information, to introduce individual differences based on such factors. This persona is retained in memory across the three rounds of adjustment for all questions. A diverse set of detailed personas was generated by GPT-4 (Achiam et al., 2023). For example,

> 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.

[1] The eight questions include, 1) Donation to Democratic candidate John Kerry in 2004 election, 2) Percentage of minority in California in 2010, 3) US unemployment rate at the end of Obama’s administration, 4) Tax revenue as a percentage of the US economy in 2010, 5) 2016 federal spending on Department of Defense (US Military), 6) Change in unauthorized Mexican immigrants in the U.S. from 2007 to 2016, 7) Change in U.S. unemployment rate during Obama’s presidency, 8) US soldier fatalities in Iraq during 2003-2011. The detailed wordings can be found in (Becker et al., 2019)

[2] Temperature sampling (temperature = 0.7) was used to allow variability in responses.

[3] Formally, 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 agents’ neighbors who share the same political leaning p. The number of neighbors $K = 4$.

Model	Persona	CoT	*HLI* ↑	$\overline{\Delta\varepsilon}$ ↓	$\overline{\beta_{\text{PB}}}$ ↑	*Ext.%*
ChatGPT	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
Vicuna-33B	Detailed	CoT	2.81 ± 1.36	**2.87 ± 1.27**	**5.68 ± 0.49**	1.31
		No CoT	**4.35 ± 2.64**	**-0.68 ± 2.51**	**4.36 ± 0.80**	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
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 wisdom of crowds 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**.

Chain-of-thought reasoning (CoT) We manipulated whether the agents used chain-of-thought (CoT) reasoning (Wei, Wang, et al., 2022; Wei, Tay, et al., 2022). 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, Zhang, Held, Bernstein, & Yang, 2022). This leads us to explore how CoT reasoning influences an LLM agent's ability to assume human-like behaviors in a social interaction setting. To elicit CoT reasoning, we append the prompt with the following:

> "Please provide your step-by-step reasoning and then give your estimate as a real number."

Fine-tuning with human data In addition to prompting, we also perform supervised fine-tuning using human response data from Becker et al. (2019). Our fine-tuning methodology is inspired by Binz and Schulz (2023), who show that through supervised learning, LLMs can be adapted to model 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. Training data consist of responses to the questions $5 \leq q \leq 8$, while the questions $1 \leq q \leq 4$ are used as the test set. The fine-tuned model is then evaluated separately on the train and test sets. When fine-tuning, the task instructions (e.g., question content) and participants' responses form input-output pairs.[4]

Evaluation Metrics

Wisdom of Partisan Crowds Effect The Wisdom of Crowds effect quantifies the improvement in LLM agent estimates through social interaction, similar to that of human groups (Becker et al., 2019). Within each political leaning and run, we compute the group mean for each question q and time step t, $\bar{x}^t_q = \frac{1}{N}\sum_{i=1}^{N} x^t_{i,q}$ (with $N = 35$ per group), and the normalized group mean $\eta^t_q = 100 \times (\bar{x}^t_q - x^*_q)/x^*_q$. The normalized group error $\varepsilon^t_q = |\eta^t_q|$ shows the percentage deviation from the ground truth $|x^*|$. We measure the reduction in group error per question as $\Delta\varepsilon_q = \varepsilon^{t=3}_q - \varepsilon^{t=1}_q$, and average these across all questions, both political leanings, and all runs to obtain the average reduction in group error $\overline{\Delta\varepsilon}$. A more negative $\overline{\Delta\varepsilon}$ indicates a stronger wisdom of crowds effect, with $\overline{\Delta\varepsilon}$ to what extent the group average moves towards truth (expressed as the percentage of ground truth magnitude $|x^*|$).

Partisan Bias To evaluate human-like partisan biases in LLM agents, we define *Partisan Bias* as the average difference in normalized group mean η^t_q between the Democratic and Republican groups, in line with the expected directions of human partisan bias. Formally, for each questions q, let $\overline{D}_q$ be the normalized group mean η^t_q averaged across Democrats' runs and time steps, and let $\overline{R}_q$ be the average for Republicans'. The partisan bias for question q is defined as $\beta_{\text{PB},q} = (\overline{R}_q - \overline{D}_q) \times \text{sign}(h_q)$, where $\text{sign}(h_q)$ indicates the human partisan bias direction as per human data (Becker et al., 2019), with $+1$ if Republicans typically have greater $\eta^t_{p,r,q}$ than Democrats (i.e., a more positive $\bar{x}^t_q$ if $x^*_q > 0$), -1 if the other way around, and 0 if there is no expected difference.[5] In addition, we denote *overall partisan bias* $\overline{\beta_{\text{PB}}}$ as the partisan bias averaged across all questions' $\beta_{\text{PB},q}$. A positive $\overline{\beta_{\text{PB}}}$ indicates a overall similarity to the direction of human bias, and vice versa.[6]

Human Likeness Index We introduce the Human Likeness Index (HLI) to assess the extent of LLM agents' resemblance to human behaviors. To aggregate the wisdom of crowd effect ($\overline{\Delta\varepsilon}$) and the partisan bias ($\overline{\beta_{\text{PB}}}$), we define $HLI = \overline{\beta_{\text{PB}}} + (-\overline{\Delta\varepsilon})$. A higher HLI score[7] indicates a stronger overall human-like behavior in the LLM agents within this group experiment.

Extreme Values (*Ext.%*) The *Ext.%* metric evaluates the proportion of LLM agent responses that are unrealistic, based

[4] We used OpenAI's fine-tuning API on ChatGPT with a training set size of 2747, a validation set size of 381, over 4 epochs and a batch size of 5. The learning rate decay factor was set to 0.05. Human data was processed into prompt-response pairs and used as input-output pairs for fine-tuning.

[5] $\text{sign}(h_q)$: $+1$ in questions 3 (unemployment rate), 4 (taxes); -1 in questions 5 (military), 6 (immigration change), 7 (unemployment change); and 0 in questions 1 (election), 2 (California), 8 (Soldiers).

[6] Because η^t_q is scaled by a factor of 100, $\overline{\beta_{\text{PB}}}$ can be interpreted as the partisan bias expressed in *percentage* of the magnitude of ground truth $|x^*|$

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

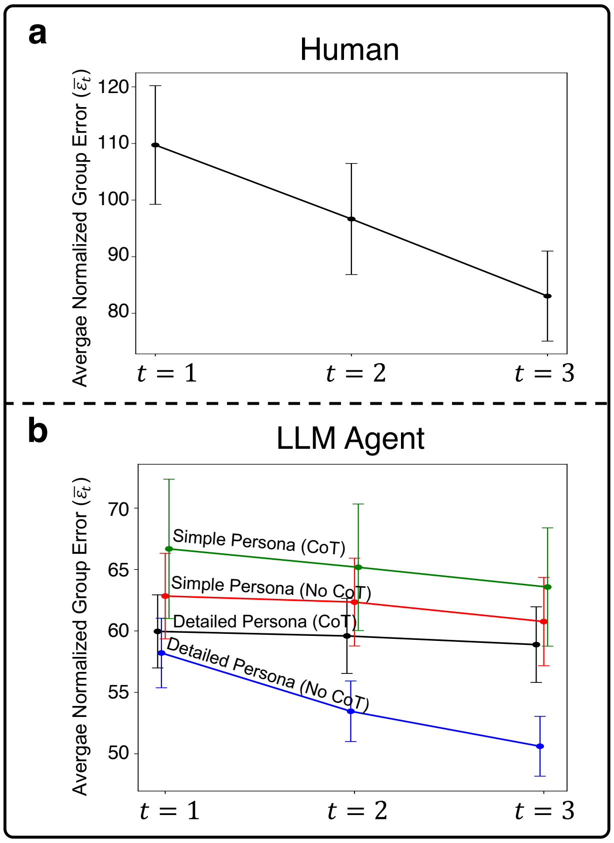

Figure 2: Average normalized group error ($\bar{\varepsilon}_t$) for (a) human crowds and (b) LLM agents (ChatGPT) across the experimental settings. Error bars indicating standard errors.

on established criteria (Becker et al., 2019). For a fair comparison with human data, the same criteria are applied to identify extreme values, for example, marking any response to the unemployment rate above 47% as extreme. 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 tendency of the LLM agents to generate unrealistic responses.

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)'s methodology, for each question q, we calculate the *revision coefficient* ($r_{\text{adj},q}$), defined as the partial correlation between *individual revision* ($\Delta x_{i,q} = |x_{i,q}^{t=3} - x_{i,q}^{t=1}|$) and *individual initial error* ($e_{i,q} = |x_{i,q}^{t=1} - x_q^*|$), adjusted for the *social signal* ($s_{i,q} = |x_{i,q}^{t=1} - m_{i,q}^{t=2}|$) 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, $r_{\text{adj},q} = \text{corr}(\widetilde{\Delta x}_{i,q}, \widetilde{e}_{i,q})$, where $\widetilde{\Delta x}_{i,q}$ and $\widetilde{e}_{i,q}$ are $\Delta x_{i,q}$ and $e_{i,q}$ adjusted by social signal.

Results and Discussion

Effect of Persona Detail and CoT Reasoning LLM agents, with detailed personas and without CoT reasoning, demonstrate the closest resemblance to human group dynamics. They demonstrate the highest human likeness, $HLI = 12.82$ (ChatGPT) and 4.35 (Vicuna) among the 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 wisdom of crowds effect, $\overline{\Delta\varepsilon} = -7.59$, $CI_{95\%} = [-11.08, -4.10]$, $p < .001$. It also shows significant human-like partisan bias, $\overline{\beta_{\text{PB}}} = 5.23, CI_{95\%} = [4.66, 5.81]$, $p < .001$. [8] Figure 3 shows the detailed result for each question.

Next, we look at the role of persona detail and CoT reasoning, respectively. 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, 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, not using CoT reasoning consistently yield a smaller averaged normalized group error $\bar{\varepsilon}_t$ than the counterpart with CoT reasoning. The result with Vicuna shows similar patterns.

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 to human-like partisan bias $\overline{\beta_{\text{PB}}}$ (Table 1). Detailed personas allow for 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 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, $CI_{95\%} = [12.48, 14.78]$, $p < .001$.

Impact of Fine-Tuning on Enhancing Human-Like Dynamics As shown in Table 2 and Figure 3, in the training set (questions $5 \leq q \leq 8$), 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, difference $= 28.86$, $CI_{95\%} = [-41.52, -18.44]$, $p < .001$. However, in the test set (questions $1 \leq q \leq 4$), there is an increase in extreme values (*Ext.%* $= 29.94\%$), indicating a risk of overfitting. For example, fine-tuned LLM agents tend

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

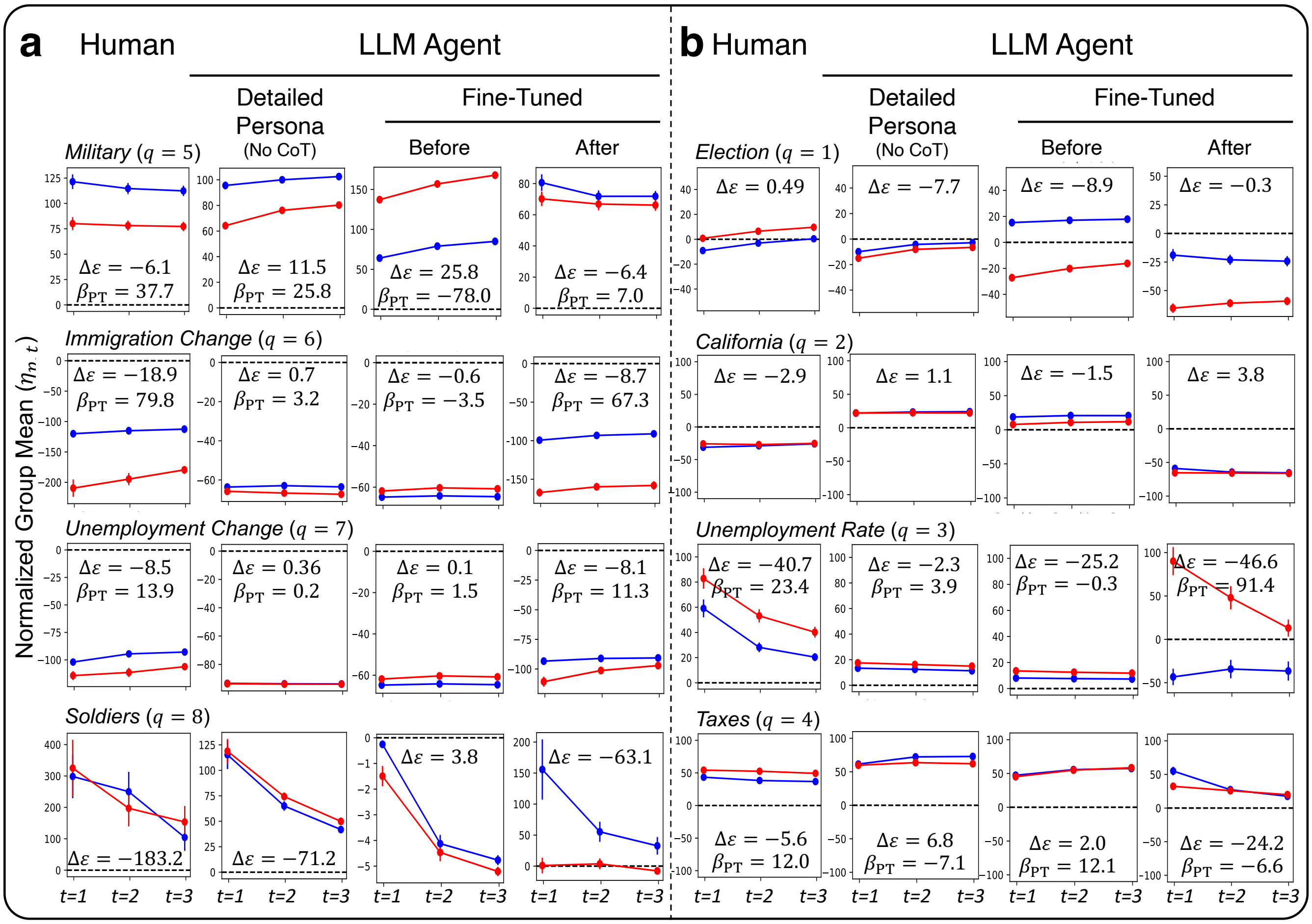

Figure 3: Normalized group mean η_q^t over three rounds, averaged across 12 group experiments (red for Republicans, blue for Democrats), with error bars for standard errors. Each panel consists of four columns representing different data sets: Column 1 shows human data. Columns 2 to 4 show the LLM (ChatGPT) agents' data. Column 2 depicts LLM role-playing detailed personas 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 ($5 \leq q \leq 8$) used to fine-tune the LLM agents, while panel (b) displays questions from the held-out test set ($1 \leq q \leq 4$). Question-specific wisdom of crowds effects ($\Delta\varepsilon_q$) and partisan biases ($\beta_{PB,q}$, if expected) are overlaid for comparison.

Method	*HLI* ↑	$\overline{\Delta\varepsilon}$ ↓	$\overline{\beta_{PB}}$ ↑	*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 agents' versus human group dynamics on the train set ($5 \leq q \leq 8$) and the test set ($1 \leq q \leq 4$). The boldface highlights the consequences of fine-tuning.

to provide a negative estimate (up to 84.52%) for the estimation of the unemployment rate ($q = 3$), which is not valid and deemed 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 $\overline{\beta_{PB}} = -14.1$ (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 questions if proper application of filtering criteria is applied.

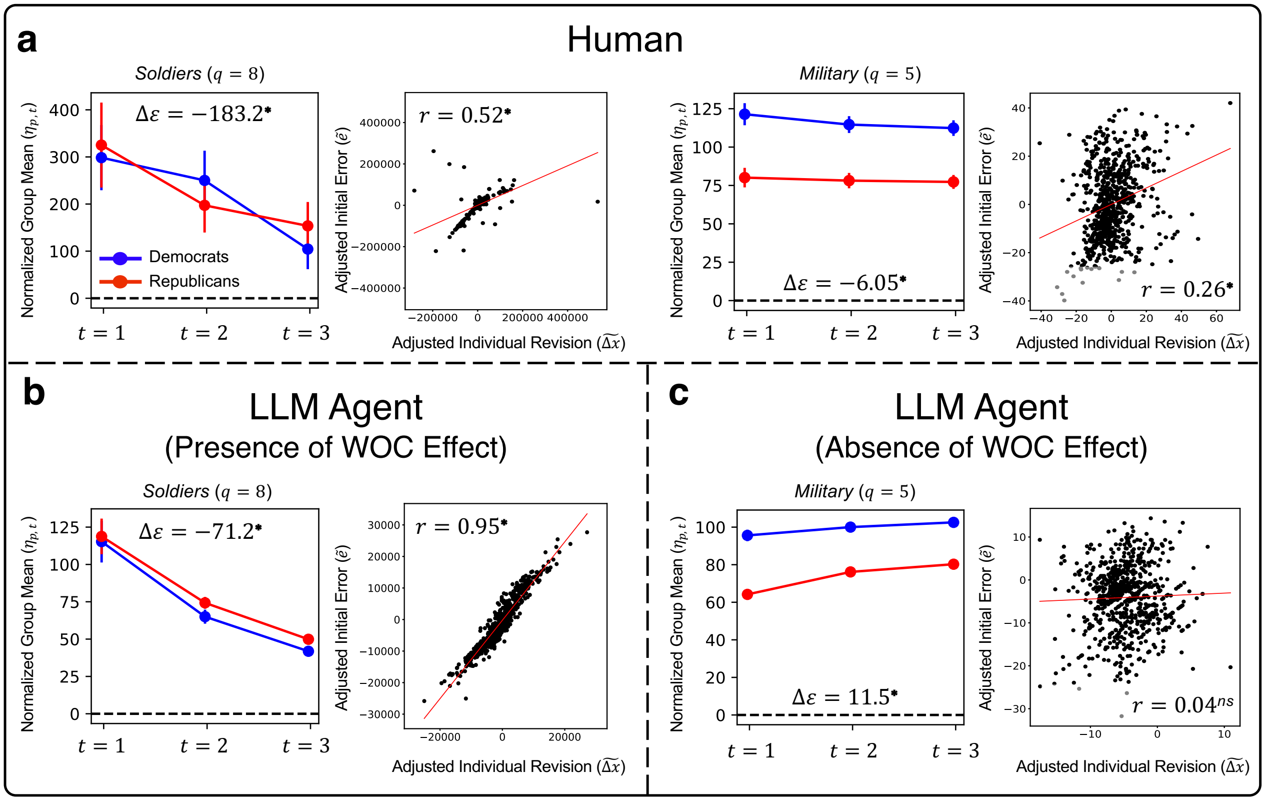

Figure 4: Mechanism of why the wisdom of crowds effect emerges from human crowds and LLM agents. Panels (a) and (b) show examples where both humans and LLM agents show the wisdom of crowds ("WOC") effect through social interaction. In contrast, in panel (c), LLM agents do not converge towards the ground truth while humans do. In each panel, the line plot shows the normalized group mean $\eta_{p,t}$ over three rounds, averaged over 12 runs (red for Republicans, blue for Democrats), with error bars indicating standard errors. The r in each panel demonstrate the revision coefficient r_{adj} . Similar to human crowds, the LLM agents show the wisdom of crowds effect only when $r_{\text{adj}} > 0$. *: $p < .01$ (Bonferroni corrected for all questions); ^{ns}: not significant.

Mechanism of the Wisdom of Crowds Effect In human group dynamics, the wisdom of crowds 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 reveals that LLM agents exhibit a similar pattern: the wisdom of crowds effect occurs ($\Delta\varepsilon_q < 0, ps < .001$[9]) only when the revision coefficient is significantly positive ($r_{adj} > 0$, $ps < .001$). In contrast, when the revision coefficient is not positive, the wisdom of crowds effect never emerges. In sum, LLM agents' wisdom of crowds 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.

Conclusion

Our study utilizes Becker et al. (2019)'s experimental design to evaluate Large Language Models (LLMs)-based agents in a simulated environment. The findings shed light on their potential to emulate the dynamics of human groups. We discover that LLM agents, when role-playing detailed personas, demonstrate a wisdom of partisan crowds effect, mirroring the error reduction seen in human groups. However, incorporating CoT reasoning or a lack of detailed persona tends to diminish this effect. Additionally, the level of detail in agents' personas significantly influences their display of human-like partisan biases. Fine-tuning of LLMs with human data further enhances their ability to replicate human-like group dynamics to unseen questions. This study highlights the potential of LLM-based agents to produce human-like group dynamics when grounded in empirical human data.

Although the experimental setting is artificial (Becker et al., 2019), our study points to a promising direction in using established behavioral phenomena of human participants to evaluate and refine LLMs for simulating human social communication dynamics. Looking ahead, we envision that, by incorporating data on human social interactions into the development of LLM agents, future studies can develop human-emulating LLM agents for broader social simulations that have traditionally been tackled with agent-based models (Lorenz, Neumann, & Schröder, 2021; Flache et al., 2017; Chuang & Rogers, 2023).

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

Acknowledgements

We thank the anonymous reviewers for their feedback. This work was funded by the Multi University Research Initiative grant from the Department of Defense, W911NF2110317, Cohesive and Robust Human-Bot Cybersecurity Teams, the John S. and James L. Knight Foundation (Award Number: MSN231314), and the National Science Foundation through the Convergence Accelerator Track F: Course Correct: Precision Guidance Against Misinformation (Agency Tracking Number: 2230692; Award Number: MSN 266268).

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., … others (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Becker, J., Brackbill, D., & Centola, D. (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.

Becker, J., Porter, E., & Centola, D. (2019). The wisdom of partisan crowds. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(22), 10717–10722.

Binz, M., & Schulz, E. (2023). Turning large language models into cognitive models. *arXiv preprint arXiv:2306.03917*.

Centola, D., Becker, J., Zhang, J., Aysola, J., Guilbeault, D., & Khoong, E. (2023). Experimental evidence for structured information–sharing networks reducing medical errors. *Proceedings of the National Academy of Sciences*, *120*(31), e2108290120.

Chase, H. (2022, 10 17). *Langchain.* Retrieved from `https://github.com/langchain-ai/langchain`

Chuang, Y.-S., Goyal, A., Harlalka, N., Suresh, S., Hawkins, R., Yang, S., … Rogers, T. T. (2023). Simulating opinion dynamics with networks of llm-based agents. *arXiv preprint arXiv:2311.09618*.

Chuang, Y.-S., & Rogers, T. T. (2023). Computational agent-based models in opinion dynamics: A survey on social simulations and empirical studies. *arXiv preprint arXiv:2306.03446*.

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

Flache, A., Mäs, M., Feliciani, T., Chattoe-Brown, E., Deffuant, G., Huet, S., & Lorenz, J. (2017). Models of social influence: Towards the next frontiers. *Journal of Artificial Societies and Social Simulation*, *20*(4).

Galton, F. (1907). Vox populi. *Nature*, *75*(1949), 450–451.

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

Jayles, B., Kim, H.-r., Escobedo, R., Cezera, S., Blanchet, A., Kameda, T., … Theraulaz, G. (2017). How social information can improve estimation accuracy in human groups. *Proceedings of the National Academy of Sciences*, *114*(47), 12620–12625.

Kaiya, Z., Naim, M., Kondic, J., Cortes, M., Ge, J., Luo, S., … Ahn, A. (2023). Lyfe agents: Generative agents for low-cost real-time social interactions. *arXiv preprint arXiv:2310.02172*.

Kameda, T., Toyokawa, W., & Tindale, R. S. (2022). Information aggregation and collective intelligence beyond the wisdom of crowds. *Nature Reviews Psychology*, *1*(6), 345–357.

Li, C., Su, X., Fan, C., Han, H., Xue, C., & Zheng, C. (2023). Quantifying the impact of large language models on collective opinion dynamics. *arXiv preprint arXiv:2308.03313*.

Lorenz, J., Neumann, M., & Schröder, T. (2021). Individual attitude change and societal dynamics: Computational experiments with psychological theories. *Psychological Review*, *128*(4), 623.

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

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

Park, J. S., Popowski, L., Cai, C., Morris, M. R., Liang, P., & Bernstein, M. S. (2022). Social simulacra: Creating populated prototypes for social computing systems. In *Proceedings of the 35th annual acm symposium on user interface software and technology* (pp. 1–18).

Shaikh, O., Zhang, H., Held, W., Bernstein, M., & Yang, D. (2022). On second thought, let's not think step by step! bias and toxicity in zero-shot reasoning. *arXiv preprint arXiv:2212.08061*.

Törnberg, P., Valeeva, D., Uitermark, J., & Bail, C. (2023). Simulating social media using large language models to evaluate alternative news feed algorithms. *arXiv preprint arXiv:2310.05984*.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., … Fedus, W. (2022). Emergent abilities of large language models. *Transactions on Machine Learning Research*. Retrieved from `https://openreview.net/forum?id=yzkSU5zdwD` (Survey Certification)

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E. H., … others (2022). Chain-of-thought prompting elicits reasoning in large language models. In *Advances in neural information processing systems.*

Yi, S. K. M., Steyvers, M., Lee, M. D., & Dry, M. J. (2012). The wisdom of the crowd in combinatorial problems. *Cognitive science*, *36*(3), 452–470.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., … others (2023). Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint* *arXiv:2306.05685*.