Are Multi-Agents Really Useful for Modeling Human Interaction Behaviors?

A Case Study on the Wisdom of Crowds Phenomenon

Chevrier Lucas, Arribas Hugo
Université Paris Dauphine – PSL

Abstract

This work explores the capacity of Large Language Models (LLMs), organized as multi-interactive agents, to reproduce collective intelligence phenomena, particularly the wisdom of crowds. Through a partial replication of the experiment in the article *adaptive social networks promote the wisdom of crowds* (Almaatouq et al.), we evaluate whether LLM agents can produce collective judgments more accurate than those of isolated agents, and under what conditions network structure influences the quality of these judgments. Our study reveals that, contrary to initial results, dynamic networks do not systematically outperform static networks. Agents fail to consistently select the highest-performing peers, calling into question the effectiveness of simulated social adaptation. However, all network configurations offer better performance than isolated agents, partially validating the wisdom of crowds hypothesis. Our results highlight the technical and methodological limitations of LLMs in faithfully simulating human social dynamics, particularly due to contextual memory constraints, visualization capabilities, and trust modeling. This research contributes to discussions on the heuristic scope of LLM agents in social simulation and opens perspectives for more realistic modeling of human interactions using artificial intelligence.

1 Introduction

In 1907, Sir Francis Galton asked 787 villagers to estimate the weight of an ox. None of them found the correct answer. Yet, by averaging their estimates, Galton obtained a result almost identical to the actual weight of the animal. This is one of the first observations of the "wisdom of crowds" phenomenon. At a symposium in 2008 at the Collège de France, Jon Elster defined the wisdom of crowds as follows: "principles and mechanisms (...) by which a group may be better able to solve problems, obtain more correct information, or make more accurate predictions than each of its members taken individually, including experts within the group." To provide evidence of this phenomenon and observe it, social and behavioral sciences, among others, have multiplied experiments. These works have identified and discussed variables favoring, or not, the emergence of this collective intelligence. However, these methods present limitations. They are usually observed through experimental methods. But the law of sampling dictates that conclusions are more robust when the number of observations is higher. Yet, it is costly and difficult to conduct such experiments on a large scale. Moreover, behavioral data is often difficult to access, as it is concentrated in the hands of private actors. To overcome these obstacles, we propose to mobilize, to a certain extent, Large Language Models (LLMs), that is, computer programs operating on the basis of machine learning and neural networks to generate text or images. In particular, multi-agent approaches, or networked agents, offer interesting opportunities for modeling interacting economic agents. This is, in any case, the central hypothesis of our research. We wish to contribute to scientific reflection on the capacity of multi-agent LLM models to simulate human behaviors, taking the wisdom of crowds as a case study. In doing so, we seek to produce evidence in favor of—or against—the heuristic scope of these models for modeling collective intelligence phenomena. We thus propose a partial reproduction of the experiment by Almaatouq et al. (2020), following the methodology of experimental replications proposed by John Horton. The original study observes the positive influence of network adaptivity on the construction of a collective judgment closer to reality. Two experiments were conducted with participants recruited via Amazon Mechanical Turk, who had to estimate correlations from scatter plots. Variables such as available information and interaction structure were adjusted to test the effect of network adaptability. We reproduce here part of this experiment and compare our results to those of the reference article, in order to draw conclusions on the research question that concerns us: To what extent can human behaviors be mobilized using LLM multi-agents?

Our demonstration will be organized as follows: a literature review (Section 2) will precede the presentation of our method (Section 3). We will then present the results obtained (Section 4), before discussing their scope and the limitations of our approach (5).

2 Literature Review

The concept of "wisdom of crowds" postulates that the aggregation of judgments from a large number of individuals can surpass individual expertise. This intuition is confirmed in many domains. From fungi (Tero et al.) to human thinking exercises (Hong and Page), through the behavior of fish schools (Couzin et al.), numerous experiments have provided evidence of the phenomenon. Nevertheless, some works have nuanced the influence of the collective on the quality of judgments when they focused on the determining variables of the wisdom of crowds. Social influence in this regard can harm this collective wisdom: when individuals adjust their opinions to conform to the group, the diversity of opinions decreases, weakening the quality of collective judgment (Lorenz et al.).

However, some research indicates that social interaction, particularly in decentralized networks, can on the contrary improve collective estimates. Becker et al. (2017) showed that, in such configurations, information exchange allowed for better judgment accuracy (Becker et al.). Thus, the network structure directly influences the quality of collective predictions. From a more recent perspective, researchers suggest that LLM language models are capable of convincingly imitating human communication (Horton). These models can serve to create generative agents that interact in simulated environments: they initiate conversations, disseminate information, remember past events, and plan their actions. Just as homo oeconomicus is a theoretical model in economic sciences, these agents could constitute a "homo sillicus" for experimenting with economic behaviors. However, the fidelity of these models to human dynamics remains uncertain. It has been observed that, without "Chain-of-Thought" (CoT) reasoning, LLMs exhibit a WOC (Wisdom of Partisan Crowds) effect that mimics the error reduction dynamics observed in humans. Research on the use of LLMs in social simulation is intensifying, but comparisons with human data are rare. Törnberg et al. (2023) thus simulated social networks with LLMs to study the effect of news feed algorithms, without however validating their results by human comparison (Törnberg et al.). More recently, Yang et al. (2024) introduced OASIS, a scalable simulation framework using up to one million LLM-driven agents to replicate social phenomena on platforms like Reddit and X. Their work stands out by empirically comparing simulated dynamics (e.g., polarization, herd behavior) with human data, offering a robust benchmark for validating LLM-based social simulations (Yang et al.).

In this project, we aim to participate in these recent discussions by proposing a comparative study of data from observed human behaviors, and data from the modeling of these behaviors by LLM multi-agents. To do this, we reproduce the study by Almaatouq et al. (2020). It explores the effects of adaptive networks on the wisdom of crowds. It highlights two mechanisms: global adaptation, where network structure evolves to strengthen the influence of the most performing members; and local adaptation, where the most accurate individuals better resist social influence, increasing the weight of their initial estimation. These two phenomena are described as determining variables in the emergence of the wisdom of crowds.

3 Methodology

3.1 Experimental Design

In this study, we conducted a partial replication of Almaatouq et al.'s (2020) experiment to evaluate the capacity of LLM agents, organized in a network, to reproduce wisdom of crowds dynamics. All source code and experimental configurations used in this replication are accessible at: `https://github.com/LucasChevrierGit/OASN`. Using the GPT-4o mini model, we designed three experimental configurations: isolated agents (without interaction), static network (each agent is fixedly connected to two others), and dynamic network (connections evolve at the initiative of the agents themselves). Each group consists of 6 agents, playing 10 successive rounds. In each round, agents receive an image representing a scatter plot and must estimate the correlation value between the points in the plot. After an initial estimate, each agent goes through three sub-rounds during which they can consult the past responses of two other agents. Note that they can access their performance in the previous round, and decide to adjust their own estimate at any time. In dynamic networks, agents freely choose the two peers they wish to follow in each round, based on past performances visible in the conversation and those of the current round, explicitly displayed. An exogenous shock is introduced halfway through, by modifying the difficulty level of the images

for certain agents, to test the resilience of networks to a sudden change in environment. Agent performance is evaluated based on the mean absolute error between their final estimate and the actual correlation value. The comparison of the three configurations—isolated agent, static network, dynamic network—allows for analyzing the effects of interaction, adaptability, and peer choice on the emergence of collective wisdom. However, important methodological constraints should be noted: a reduced number of participants (6 instead of 12), a shorter experiment duration (10 rounds instead of 20), and structural limitations related to the very functioning of LLMs, which particularly affect the representation of memory and the continuity of interactions.

Experiment 1 (E_1): Varies Network Plasticity; Holds Feedback. In the first experiment (E_1, $N = 719$), each group was randomized to one of the following three treatment conditions: 1) a solo condition, where each individual solved the sequence of tasks in isolation; 2) a static network condition, in which 12 participants were randomly placed in static communication networks with a fixed degree (i.e., each participant had exactly three neighbors); and 3) a dynamic network condition, in which participants at each round were allowed to select up to three neighbors to communicate with. Across all conditions, at each round, the participants were initially asked to submit an independent guess. Then, those in static and dynamic network conditions entered a social-exposure stage, where they could observe the answers of their network peers, update their own, and see their peers' updated beliefs in real time. After submitting a final guess, the participants in all conditions were provided with performance feedback. Finally, those in the dynamic network condition were allowed to revise which peers to follow (up to three neighbors) in subsequent rounds (see *SI Appendix*, Fig. S3 for the experimental design and *SI Appendix*, Figs. S4–S8 for the online platform screenshots).

Figure 1: Extract of the Almaatouq's et al. (2020) methodology. Focus on the experimentation we replicated.

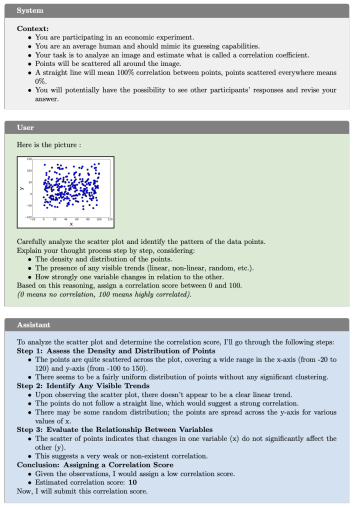

Figure 2: First stage discussion (initial estimate)

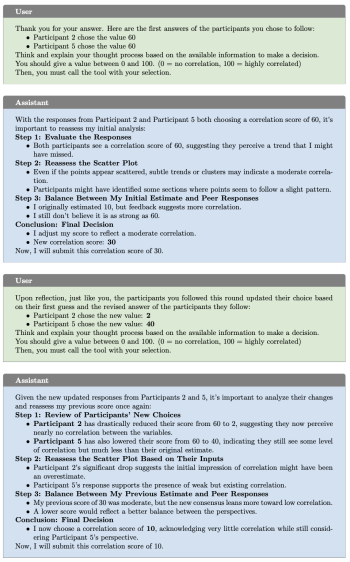

Figure 3: Second stage discussion with 2 sub rounds (advised choice)

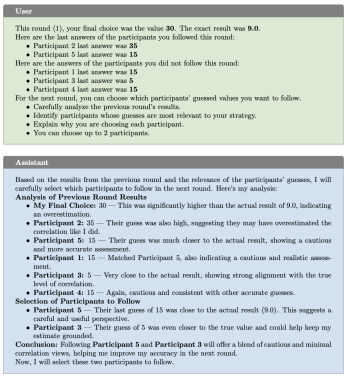

Figure 4: Third stage discussion (choosing participants for next round)

3.2 Role-Based Prompting Framework

To simulate human decision-making and social interaction in our experiment, each LLM agent operates within a structured prompting framework, composed of three distinct roles: system, user, and assistant. The system prompt provides the agent with general context and behavioral instructions such as acting like an average human participant and estimating correlation coefficients from scatter plots. The user prompt simulates the role of the experiment facilitator, delivering the scatter plot and asking the agent to explain its reasoning and submit an estimate. The assistant response corresponds to the agent's behavior: it analyzes the image, provides a detailed justification for its estimate, and interacts with simulated peer feedback when applicable. Crucially, whenever a response is requested from an agent via the OpenAI API, the entire discussion history is systematically included in the prompt. This enables the model to recall its previous actions and those of others, simulating a form of short-term memory essential for modeling sequential decision-making and social influence.

3.3 Condition-Specific Interaction Flows

The structure of these exchanges varies depending on the experimental condition. In the solo self-feedback configuration (Figure 1), the prompt sequence includes only the image and the agent's own estimation, without any social interaction or opportunity for revision. In the static network condition (Figures 1 and 2), agents also receive feedback and are exposed to a fixed subset of peers'

responses, allowing them to revise their judgment across multiple sub-rounds. The dynamic network condition (Figures 1, 2 and 3) adds an additional layer of complexity: agents not only observe peer responses and performances but are also prompted to select which participants to follow in the next round. This selection is informed by performance data, including both historical and current round accuracy, enabling adaptive peer choice based on perceived reliability.

To ensure that each agent produces a structured and actionable output, every API call also included a function-calling tool. This mechanism required the model to return both a natural language explanation of its reasoning and a precise numerical response in a predefined format.

3.4 Prompt Variations and Confidence Reasoning

To explore the effects of prompt design on agent behavior, two different prompting setups were employed in the experiment. Both versions followed the same structural logic, involving role-based interactions (system, user, and assistant) and multi-round dynamics. However, they differed in the degree of instruction provided to the agents—specifically regarding self-confidence and decision revision.

In the baseline prompting configuration, the initial user prompt during the first estimation round included four main tasks: (1) observe the image, (2) analyze the scatter plot based on visual patterns, (3) reason through the relationship between variables, and (4) assign a correlation score between 0 (no correlation) and 100 (perfect correlation). In the advised-choice stage, when agents were exposed to peers' responses—the prompt simply instructed them to reflect on the available information, update their estimate accordingly, and call the function with their revised score.

In the extended prompting configuration, additional lines were added to simulate self-assessment and adaptive confidence. During the initial estimate, agents were prompted not only to assign a correlation score, but also to explicitly express how confident they were in the accuracy of their estimate given the image. In the advised-choice stage, the prompt included an additional instruction: agents were told that any update to their score should consider both their own confidence level and the estimates of the other participants. This adjustment was intended to encourage more deliberate weighting of self-generated judgments versus socially derived information, and to better simulate human-like decision calibration under uncertainty.

4 Results

The results of the replication are mixed and reveal key differences between human behavior and that of LLM-based agents.

General results A first and striking observation is the significantly higher variance in estimation errors across rounds when compared to human participants. This is true for both prompting configurations. While humans tend to produce relatively stable performances when exposed to similarly difficult visual tasks, LLM agents show unpredictable fluctuations in accuracy, even between scatter plots that are visually and statistically close. This inconsistency highlights the jagged technological frontier of LLM capabilities: their capacity to deliver accurate responses appears not only non-linear, but also highly sensitive to subtle and sometimes opaque variations in visual inputs. In other words, the reliability of an agent's estimate is not always anchored in an interpretable logic, making it difficult to trace performance regressions or improvements to clear factors.

As a direct consequence of this unpredictability, the introduction of an exogenous shock, implemented at round five, by changing the image difficulty levels for each agent had a relatively

modest impact on the overall group dynamics. In the original human-based experiment, increased task difficulty disrupted previously stable estimation patterns, creating a measurable gap in learning behavior. In contrast, LLM agents in our study already displayed considerable noise in their performance prior to the shock, such that the additional difficulty did not systematically degrade accuracy. In fact, for some agents, performance improved post-shock, likely due to random variance rather than adaptive behavior.

To mitigate this limitation, increasing the number of experimental runs while randomizing the order of images for each agent could have helped smooth out individual fluctuations and produce more generalizable trends. However, the original experiment relied on a fixed image sequence for all participants, and the goal of this replication was to mirror the original protocol as closely as possible to ensure methodological comparability.

In both prompting configurations, AI-agent networks are better than individual agents, which is evidence of a certain wisdom of crowds. The best of the agents is indeed worse than the average of the group's choices.

Baseline prompting configuration The major difference with the original experiment is that groups in the dynamic network are on average worse than the static network. These results go against both the conclusions of the paper and intuition. If given the opportunity, agents should prioritize following the choice of the most successful participants. In our case, AI agents (with baseline prompting) choose the previously most successful agent in only 62 percent of cases (among their two potential "neighbors"). Moreover, the hypothesis validated by the paper of a capacity of dynamic networks to better support shocks is also not in line with our results. The shock, characterized in the article by a modification of difficulties (associated with the displayed images) for each agent, does not modify the general trend. The static network remains more efficient. The error is lower and the learning after the modification of the system by the shock is better.

Extended prompting configuration Forcing the agents to express their confidence in both their own estimates and those of their peers—had a significant impact on behavior across both network configurations. Most notably, the learning trajectory of agents in the dynamic network improved substantially compared to the baseline configuration, selecting the most successful agent in 68 percent of cases (against 62 percent in the baseline configuration). While the first five rounds still reflected higher average errors in the dynamic condition than in the static one, the trend shifted following the introduction of the exogenous shock at round five. From that point onward, dynamic networks consistently outperformed static networks in terms of accuracy.

These results align more closely with the findings of the original study, which hypothesized that dynamic networks are better suited to adapt to environmental changes. The explicit incorporation of confidence as a decision-making factor appears to support this adaptivity, as agents became more selective in who they followed and revised their estimates more cautiously. The dynamic networks were not only more resilient post-shock, but also demonstrated more consistent error reduction round after round, suggesting a form of collective learning. In contrast, the performance of the static network declined after the shock under this prompting configuration.

These findings help to mitigate the limitations observed in the baseline setup. They suggest that adding simple mechanisms for self-assessment and confidence-based reasoning can enhance the quality and adaptiveness of multi-agent LLM networks, bringing them closer to human-like behavior in social estimation tasks.

Limitations These results should however be interpreted with caution. The original experiment included 239 games for the control experimental condition (solo self-feedback) and 20 games for each of the other two conditions, each game comprising 20 rounds. In comparison, our replication

was limited to 3 games per condition, with only 10 rounds each. The number of participants was also reduced from 12 to 6. This reduction in the number of rounds limits the possibility of learning both individually and collectively. Moreover, each agent could only follow the estimates of two other participants, which considerably restricts access to the information necessary to refine their own estimates. Finally, the weakness of the sample makes the value of the statistical results more uncertain. As a result, comparing our two prompting configurations, both with each other and with the original experiment—should be approached with caution, as the observed differences may in part reflect sampling variability or context-specific dynamics rather than systematic effects.

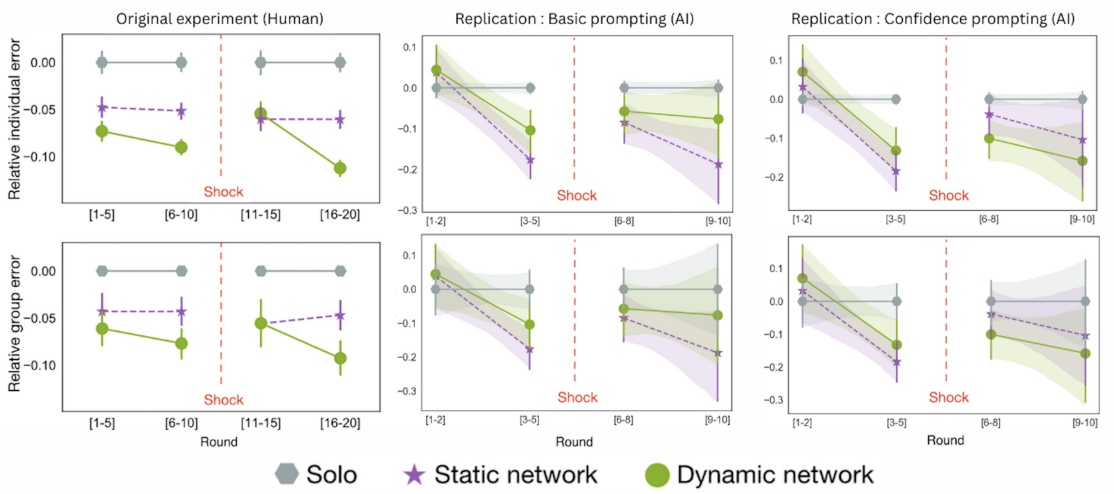

Figure 5: Comparative results of the respective experimentation.

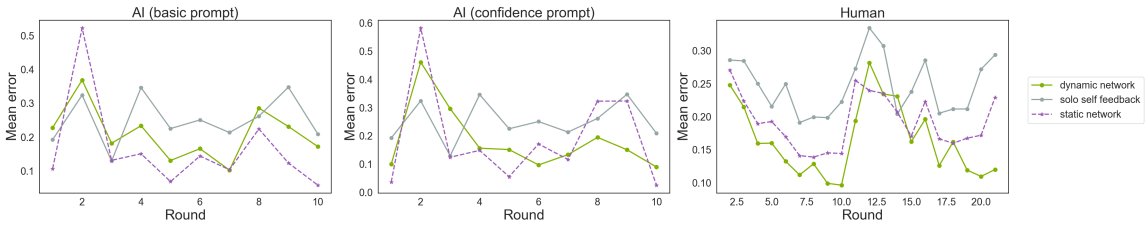

Figure 6: Comparative representation of the overall performances of both human and multi-agents LLL during the experimentations.

5 Discussion

Impact of Prompting and Experimental Structure Inadequate prompting constitutes one of the main factors limiting the performance of LLMs to produce better results. In the case of our replication, this limitation is clearly visible. The first two rounds of the experiment , for both the

static and dynamic networks, are less effective than individual agents. This trend gradually reverses, suggesting a learning phase of the game rules. This initial information deficit is accentuated by a major difference between the original experiment and our protocol. The original experiment relies on the ability of participants to see the modification of the choice of people they follow in real time. This dynamic made visible the degree of confidence associated with each choice. For example, a small variation can signify strong assurance, and vice versa. To try to reproduce this aspect, the choice was to define several sub-rounds (3), each round allowing the agent to see the choice of the previous round.

Moreover, to limit the number of tokens per request—and thus the overall cost of an experiment—the image whose correlation value the model had to find was only provided during the initial estimate. During the "discussion" phase with other models, the image is removed from the conversation, limiting the agents' ability to "reason", but especially to put into perspective the estimates of the participants they follow with their own. The extended prompting configuration, which required the agent to assess its confidence in both its own estimate and those of its peers, helped partially mitigate this limitation. By explicitly prompting the model to reflect on the reliability of each response, the additional layer of reasoning compensated, at least in part, for the absence of the image during the revision phase. This self-assessment mechanism provided the LLM with contextual cues to evaluate the relative value of competing estimates, encouraging more selective and deliberate decision-making despite the lack of visual reference.

Addressing this memory deficit is an important direction for improving LLM-based multi-agent models. Researchers have begun to experiment with adding memory architectures on top of LLMs to more closely mimic human-like recall and forgetting. For example, Park et al. (2023) implemented a memory module for generative agents: a structured database of an agent's experiences (stored in natural language) indexed by time and importance (Park et al.). Their agents could retrieve relevant past events and even generate higher-level "reflections" about those experiences (e.g. inferring personal preferences or lessons learned), which then informed future decisions. This design proved effective at maintaining character consistency and social coherence over long periods. Such mechanisms are directly applicable to our context: with a memory module, an agent in the wisdom-of-crowds simulation could recall its own past guesses, remember peer interactions, and refine its strategy (or its choice of whom to follow) accordingly. It could accumulate a sense of which other agent tends to know what they're doing, or which information was revealed to be misleading. Incorporating a long-term memory store would thus likely enhance the reasoning across rounds that is currently partly missing. Contrary to having only the discussion as the history of action and strategy taken during previous rounds, a memory module would force the agent to reflect on past information enabling better control over what data the agent is using to generate responses.

Technical Constraints Related to the Context Window This limitation related to the number of tokens is not only a budgetary limitation but also a technical one, due to the context window of LLMs. The context window refers to the maximum amount of text (or "tokens") that the model can take into account at once to generate a response. It is the short-term memory of the model for a query. It includes:

– The user's prompt
– Previous instructions (conversation history)
– The model's response

The model used for the replication, GPT-4o mini, benefits from a context window of 128,000 tokens, compared to about 16,000 for ChatGPT-3.5. In our experiment, a single image requires up to 50,000 tokens, which mechanically limits the number of images that can be kept in the

conversation history. A solution considered could have been to reduce the quality of the images to decrease their token cost. However, in the original protocol, human agents only had access to the image during a single round of conversation. If this principle were applied analogously to GPT-4o mini, the constraint related to the context window would not constitute an immediate obstacle. Nevertheless, the representation of visual memory for an LLM remains a challenge. One approach would be to keep the image in the model's history, but this solution remains costly in tokens and limiting. It is therefore necessary to explore alternatives to simulate human photographic memory more efficiently.

6 Conclusion

This study assessed the capacity of large language models (LLMs), deployed as multi-agent systems, to replicate the wisdom of crowds observed in human interactions. By partially replicating the experimental protocol of Almaatouq et al. (2020), we tested whether dynamic networks of LLM agents could generate more accurate collective judgments than static networks or isolated agents.

In its baseline configuration, the experiment revealed several discrepancies between human and LLM agent behavior. Notably, dynamic networks failed to outperform static ones, both in stable conditions and after the introduction of exogenous shocks. This finding challenges the assumption that interaction flexibility inherently improves collective intelligence in LLM-based systems. However, the introduction of an extended prompting configuration, which explicitly incorporated self-confidence and trust reasoning into the agents' decision-making process, significantly improved outcomes. Under this configuration, agents in dynamic networks became more selective and adaptive, closer to the behavior observed in human groups. This prompt modification allowed dynamic networks to outperform static ones in the latter half of the experiment, aligning more closely with the original study's conclusions.

Nevertheless, several limitations constrain the generalizability of our results: a reduced number of agents and rounds, memory constraints, and limited access to visual information during social exchanges. These structural factors introduce bias and highlight the need for caution in interpreting the results. Future work could explore more robust memory architectures, integrate iterative reasoning techniques such as Chain-of-Thought, and employ larger-scale experiments to better evaluate general trends.

These results highlight the critical importance of prompt design in shaping multi-agent behaviors. Even simple additions like confidence evaluation can enable more nuanced and human-like decision strategies in artificial agents. Nevertheless, technical limitations remain. Reduced sample sizes, a limited number of rounds, memory constraints, and the inability to retain or process visual information across rounds all limit the realism and generalizability of our findings.

References

[Almaatouq et al.] Almaatouq, A., Noriega-Campero, A., Alotaibi, A., Krafft, P. M., Moussaid, M., and Pentland, A. Adaptive social networks promote the wisdom of crowds. 117(21):11379–11386.

[Becker et al.] Becker, J., Brackbill, D., and Centola, D. Network dynamics of social influence in the wisdom of crowds. 114(26).

[Couzin et al.] Couzin, I. D., Krause, J., Franks, N. R., and Levin, S. A. Effective leadership and decision-making in animal groups on the move. 433(7025):513–516.

[Hong and Page] Hong, L. and Page, S. E. Groups of diverse problem solvers can outperform groups of high-ability problem solvers. 101(46):16385–16389.

[Horton] Horton, J. J. Large language models as simulated economic agents: What can we learn from homo silicus? Version Number: 1.

[Lorenz et al.] Lorenz, J., Rauhut, H., Schweitzer, F., and Helbing, D. How social influence can undermine the wisdom of crowd effect. 108(22):9020–9025.

[Park et al.] Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., and Bernstein, M. S. Generative agents: Interactive simulacra of human behavior.

[Tero et al.] Tero, A., Takagi, S., Saigusa, T., Ito, K., Bebber, D. P., Fricker, M. D., Yumiki, K., Kobayashi, R., and Nakagaki, T. Rules for biologically inspired adaptive network design. 327(5964):439–442.

[Törnberg et al.] Törnberg, P., Valeeva, D., Uitermark, J., and Bail, C. Simulating social media using large language models to evaluate alternative news feed algorithms. Version Number: 1.

[Yang et al.] Yang, Z., Zhang, Z., Zheng, Z., Jiang, Y., Gan, Z., Wang, Z., Ling, Z., Chen, J., Ma, M., Dong, B., Gupta, P., Hu, S., Yin, Z., Li, G., Jia, X., Wang, L., Ghanem, B., Lu, H., Lu, C., Ouyang, W., Qiao, Y., Torr, P., and Shao, J. OASIS: Open agent social interaction simulations with one million agents.