Forecasting with Societies of Agents

Nathaniel Li, Yury Orlovskiy, Evan Ellis
Anish Kachinthaya, Harbani Jaggi, Haokun Zheng[1]

[1]UC Berkeley

Abstract

Recent developments in in large language models (LLMs) have enabled the creation of agents that can simulate human behavior and interaction. While these agents have shown promise in various domains, their application to forecasting remains largely unexplored. We present an approach that leverages societies of LLM-powered agents to predict outcomes in complex decision-making scenarios. Our system creates diverse agent roles (judges, jurors, prosecutors, etc.) that interact through structured dialogue to simulate real-world deliberation processes. By allowing agents to collaboratively evaluate evidence and resolve disagreement through structured reasoning, our system provides interpretable and transparent forecasts across diverse scenarios, including policy debates and bargaining simulations. The system achieves an accuracy of 60% when predicting outcomes on Manifold Markets, an outcome-betting website. Our approach's flexibility enables rapid iteration and intervention testing, allowing researchers to explore how varying assumptions or inputs affect predicted outcomes, as well as how different deliberation strategies can improve accuracy. We qualitatively demonstrate the framework's ability to simulate nuanced interactions and produce forecasts accompanied by multi-faceted and evidence-backed reasoning chains. Unlike previous forecasting approaches that rely on single-model predictions or human expert ensembles, our method enables rapid simulation of multiple scenarios and stakeholder interactions, allowing for dynamic testing of different interventions and their impacts on predicted outcomes. We release our code and simulation templates at https://github.com/festusev/AgentSocieties. In addition, we release a rolling benchmark of simulations based on open Manifold Markets questions at https://huggingface.co/datasets/evanellis/manifold. Our video presentation is shared at https://drive.google.com/file/d/1lKNq3euphXMfLStahjoePxeuxbbrQ6YW/view?usp=sharing, and our slides are shared at https://docs.google.com/presentation/d/1OhfeZgszoOKx3qNmN2_YyLJodlgzfCv3Kzx4U2UPKOE/edit?usp=sharing.

1 Introduction

The ability to accurately predict outcomes of complex social scenarios – from legal proceedings to international trade disputes – remains an unsolved challenge. While large language models (LLMs) have demonstrated remarkable capabilities in tasks requiring reasoning and domain expertise, their potential for simulating multi-agent interactions in forecasting scenarios remains largely unexplored [Su et al., 2024]. Traditional approaches to forecasting, both with humans or ML systems, often rely on individual expert opinions or simplified statistical models, failing to capture the dynamic nature of human decision-making processes that ultimately determine outcomes.

We present a novel framework for automated forecasting using societies of LLM agents that simulate key stakeholders in decision-making scenarios (Figure 1). Our system orchestrates multiple agents – including judges, jurors, prosecutors, and negotiators – in structured dialogues that mirror real-

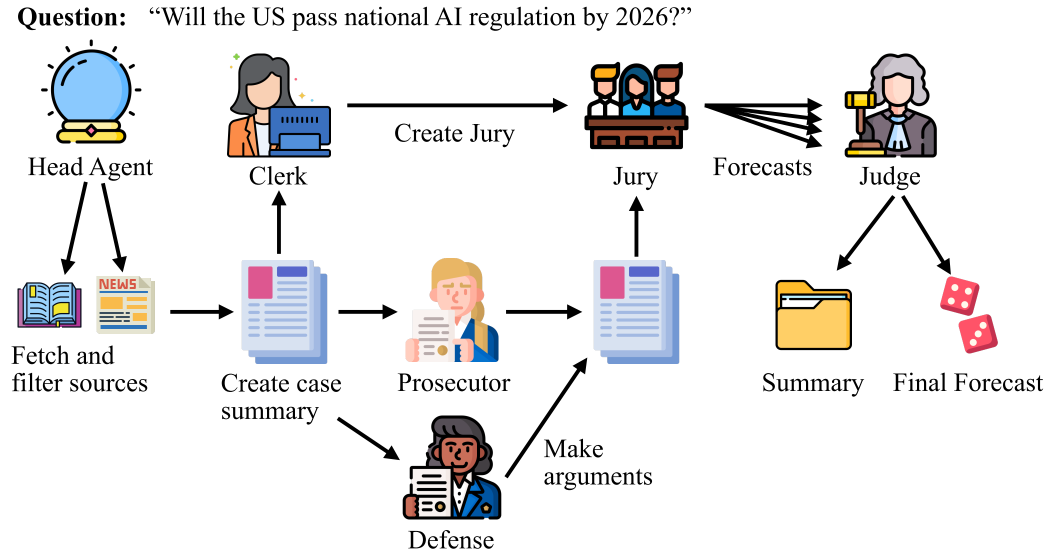

Figure 1: Our pipeline for forecasting with societies of agents, beginning with an initial setup coordinated by the head agent, followed by arguments by the prosecution and defense, and finally a deliberation amongst the jury.

world deliberative processes. Each agent maintains distinct characteristics and objectives while engaging in complex interactions such as presenting evidence, bargaining, and voting on final outcomes (Section 2). This approach leverages a fundamental advantage of machine intelligence: the ability to rapidly simulate thousands of interaction scenarios with varying initial conditions and environmental factors [Glockner et al., 2023, Gao et al., 2023]. Towards actualizing this proposal, agent frameworks [Wu et al., 2023a] have emerged as powerful tools for building multi-agent systems, enabling developers to create collaborative AI environments where multiple specialized agents work together to solve complex tasks. Among these, we use AutoGen, which stands out as an open-source framework that provides a flexible multi-agent conversation infrastructure.

The framework's implementation centers on a flexible architecture that supports diverse forecasting scenarios without requiring scenario-specific customization (Section 3). A head agent coordinates the simulation, managing turn-taking and information flow between specialized agents that represent different stakeholders. This design enables researchers to systematically study the impact of various interventions on predicted outcomes.

Initial results demonstrate that our agent societies achieve reasonable across multiple domains, including legal proceedings, trade negotiations, and policy debates. We achieve a 61% accuracy in forecasting binary outcomes, with little systematic bias from uncertainty (Section 4). To facilitate further research and development in agent-based forecasting, we are releasing our implementation and scenario templates.

2 Related Work

Forecasting has long been a challenging domain in machine learning, with approaches ranging from traditional statistical models to advanced neural networks. The emergence of agent-based systems and large language models (LLMs) has opened new avenues for tackling forecasting problems [Halawi et al., 2024, Hsieh et al., 2024, Schoenegger et al., 2024].

Seminally, Park et al. [2023] introduced "societies" of LLM agents capable of simulating human-like behaviors in a sandbox environment, demonstrating the ability of LLMs to create agents with emergent social dynamics, including work, leisure, and culture. Furthering this research, Gu et al. [2024] demonstrates a framework which demonstrates that language shapes the collective behavior of multi-agent LLM societies through interactive debate. Towards more practical applications, Wu et al. [2023b] introduced a platform for developing LLM applications using multiple conversing agents, while Chan et al. [2023] proposed benchmarks for evaluating the outputs and potential feedback loops through multi-agent LLM debate.

Some researchers have explored more specialized or speculative applications of LLM agents. Wang et al. [2023], [FAIR] have developed superhuman ML systems, employing language model agents to outperform humans in playing Minecraft and Diplomacy, respectively. Chen et al. studied consensus-seeking behaviors in LLM-driven multi-agent systems, providing insights into inter-agent negotiation dynamics. Following up on this work, Noh and Chang [2024], Abdelnabi et al. [2024] focus on evaluating LLMs' negotiation abilities, with the latter examining the influence of personality traits on negotiation outcomes.

Like all other powerful technologies, LLM agents are dual-use – they can be leveraged for both benefit and harm. Indeed, Rivera et al. [2024] highlights concerning escalation patterns in military and foreign policy decision-making scenarios where nation-states are replaced with language model agent simulations, emphasizing the need for caution in deploying AI agents in sensitive / national-security relevant application. In this project, we aim to employ language model agents for a specific application – developing simulations of important societal phenomena that are otherwise too costly to conduct in persona. Indeed, LLM agents now demonstrate near-or-exceeding-human performance on forecasting, showing real potential for improving epistemics with real world impact.

3 Method

3.1 Multi-Agent System Architecture

We propose a multi-agent framework for forecasting binary outcomes through simulated legal proceedings. Our system architecture comprises specialized agents that mirror key roles in legal deliberation, operating in a coordinated temporal sequence to generate probabilistic forecasts.

The framework is orchestrated by a head agent (Agent 0) that serves as both coordinator and primary decision-maker. The head agent initializes the process by retrieving and analyzing relevant news articles and legal documentation. Each agent in the system is instantiated with specific capabilities defined through a configuration dictionary $C = \{\text{web\_retrieval} : \text{bool}, \text{llm} : \text{bool}\}$, determining their access to external information and language model capabilities.

3.2 Information Retrieval and Ranking

Our system implements a sophisticated article processing pipeline comprising three main components: retrieval, content extraction, and ranking. The initial retrieval phase utilizes the DuckDuckGo search API with a normalized query mechanism that removes punctuation and converts text to lowercase for consistent search patterns. Content extraction employs BeautifulSoup for HTML parsing with selective element removal. The ranking mechanism implements a dual-criteria scoring system. For each article A, we compute a composite score S:

$$S(A) = \frac{R(A) + O(A)}{2} \tag{1}$$

where $R(A) \in [1, 10]$ represents the relevance score and $O(A) \in [1, 10]$ represents the objectivity score. Articles are processed in chunks of size $k = 2000$ tokens, with a maximum of three chunks per article to maintain computational efficiency while preserving content integrity.

3.3 Jury Selection

The jury selection process is managed by a specialized clerk agent that optimizes for demographic diversity across multiple dimensions. The selection algorithm considers various demographic factors including age, ethnicity, geography, and profession, denoted as vector D. The system incorporates domain-specific knowledge weights through an expertise relevance score E, while also accounting for socioeconomic distribution S across background and education levels. Geographic balance G is maintained through careful consideration of urban and rural perspective ratios. The clerk agent generates structured jury profiles, with each juror being instantiated with a system message that encodes their unique perspective and expertise, which subsequently influences their probability estimation process.

3.4 Temporal Process Flow

The simulation operates on a discrete time-step basis $T = \{0, ..., 27\}$, with each step mapping to atomic agent actions. The process begins with the head agent executing information retrieval and ranking during steps zero through two. At step three, agent instantiation occurs with capability configuration C, followed by the judge agent establishing trial parameters at step four. The prosecution and defense present their arguments from steps five through twenty-two. Jury deliberation and probability computation take place from steps twenty-three through twenty-six, culminating in the final forecast synthesis at step twenty-seven.

Here is an example dialogue between the agents:

PROSECUTOR: Based on the evidence summarized, there are compelling reasons to assert that the probability of the S&P 500 index closing higher on December 12 than on December 11 is significantly elevated. Here are the key points that support this assertion...

DEFENSE: While the evidence suggests that there could be factors leading to a higher close for the S&P 500 on December 12, it's essential to recognize that these factors also come with significant potential risks and underlying uncertainties that imply a lower probability of a higher close. Here are several key points that support the argument for a low probability forecast for the S&P 500 to close higher...

ALICE MONTGOMERY (JUROR): Based on the presented evidence, I would assign a probability of 70% to the question: Will the S&P 500 stock index close higher on December 12 than it closed on December 11?...

.
.
.

JUDGE: After reviewing the jury forecasts from five different perspectives regarding the likelihood that the S&P 500 index will close higher on December 12 than it did on December 11, we can summarize their reasoning and compute the median probability forecast as follows:...
Thus, the final summary of the jury forecasts indicates a median probability forecast of 70% that the S&P 500 stock index will close higher on December 12 than it did on December 11.

The Jurors, each with a different role and background, vote on probabilities which the Judge summarizes into the final forecast. We found that the median probability forecast performed worse than taking the mean, so our results are with the mean of the Juror's forecasts.

3.5 Implementation Details

Our implementation utilizes the GPT-4 language model for agent reasoning, with each agent configured through a ConversableAgent class with specific system messages and capability flags. The framework maintains a hierarchical control structure through a ScenarioConfig class. Agent interactions are managed through a message-passing system where each communication is logged and stored in a variable store V for reference and analysis. The head agent maintains control flow through explicit turn management and speaker selection.

3.6 Web-Based Information Processing

The article processing pipeline implements a two-stage metadata extraction process involving URL analysis and content analysis. Through this pipeline, the system processes article content using tiktoken encoding with chunk_size $= 2000$ and max_chunks $= 3$ to optimize for GPT-4 context window limitations while maintaining content coherence.

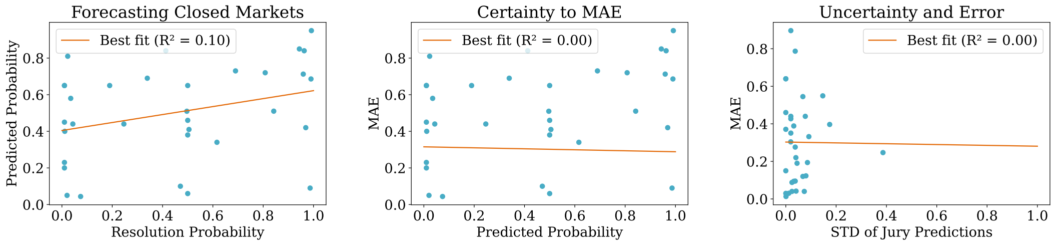

Figure 2: **Evaluation on resolved markets.** We evaluate our forecaster on 33 resolved Manifold Markets. We achieve a 61% accuracy, and a 30% mean absolute error (MAE). The left plot shows the correlation between the final resolution probability and our forecasted probability. The middle plot shows that there is little relationship between the prediction probability and the error, indicating that our method is unbiased. The right plot shows no correlation between the uncertainty of the jurors and the error.

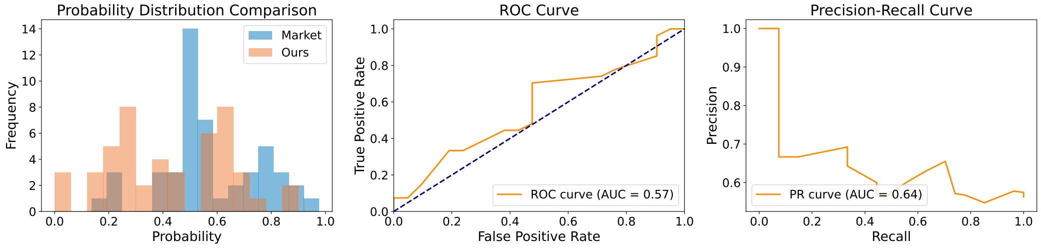

Figure 3: **Comparison with human forecasts in unresolved markets.** We compare our agent-based forecasting system and Manifold Markets predictions across probability distributions, ROC curves (AUC = 0.57), and precision-recall curves (AUC = 0.64).

3.7 Evaluation Methodology

The system's forecasting capability is evaluated using binary outcome scenarios with ground truth labels. Our evaluation framework examines multiple performance dimensions: the alignment of forecasts with actual outcomes, the degree of consensus among jurors, the effective utilization of retrieved information, and the consistency of agent interactions across multiple runs. Each scenario execution maintains comprehensive logging of agent interactions, information retrieval results, and decision processes, enabling thorough post-hoc analysis and system optimization.

4 Evaluation

4.1 Evaluation on Forecasting Tasks

We evaluate on Manifold Markets, a website which hosts prediction markets which users can bet on. We created a dataset of 33 resolve markets with custom scenarios for each. We then ran the society of agents forecaster on these scenarios. Results are shown in Figure 2. We achieved a 61% accuracy and a 30% mean absolute error (MAE) over these scenarios, providing the first baseline of its kind on Manifold Markets.

To investigate whether our system was biased, we compared the predicted probability to the mean absolute error and found that there was no correlation. This suggests that ensembles of agents do not become less accurate as their forecasted probability grows or shrinks. We also evaluated whether the uncertainty of the ensemble pool correlated with the error, and again found no correlation. These results are promising: it appears that LLMs in this deliberation process do not exhibit a kind of systematic failure mode, as we had expected. Instead, the avenues for improvement appear to be in search and context management.

4.2 Alignment with Market Sentiment

For markets that will not resolve in the near future, we perform evaluations to check the alignment between model forecasting predictions and market sentiment. Our analysis, visualized in Figure 3, reveals notable differences between human prediction patterns and our agent-based system's forecasts. The probability distribution comparison shows that human predictions on Manifold Markets tend to cluster around 50%, suggesting a tendency toward uncertainty or hedging behavior in human forecasters. In contrast, our agent-based system produces more varied predictions distributed across the probability spectrum, though rarely exceeding 90%, indicating more decisive but still bounded confidence in its forecasts.

The moderate alignment between our system's predictions and market sentiment is reflected in the ROC curve (AUC = 0.57) and precision-recall curve (AUC = 0.64). These metrics suggest that our system's predictions align with market sentiment approximately half the time, indicating independent reasoning rather than mere replication of crowd wisdom. This divergence is particularly interesting as it suggests our system is not simply learning to mimic human prediction patterns but rather developing its own forecasting methodology through structured agent interactions.

Examining specific cases reveals varying degrees of alignment with market sentiment. In the case of the alleged UHC CEO assassination prediction, our system's 85% probability forecast closely matched the market certainty of 83.97%, demonstrating strong alignment. This consensus emerged from clear evidence including possession of the murder weapon, documented stalking behavior, and explicit motives. However, in other cases, we observe notable divergences. For the Elena Ferrante identity prediction, our system assigned a 5% probability while the market showed higher uncertainty with 76.42% certainty, suggesting our agents' structured analysis of literary and cultural factors led to a more decisive conclusion than the crowd's collective judgment.

The Russia-Argentina inflation prediction presents another interesting case of divergence, where our system produced a 90% probability forecast compared to the market's 70.27% certainty. Our system's higher confidence emerged from a detailed analysis of economic indicators, with 9 out of 11 jurors assigning probabilities of 85% or higher based on concrete factors such as the projected inflation rates (9.2% vs 123%). This case illustrates how our multi-agent framework can arrive at more decisive predictions through explicit reasoning about quantitative evidence.

These examples highlight a key pattern in our system's behavior: while market predictions often reflect aggregate human intuition and uncertainty, our system tends to produce more decisive forecasts through explicit reasoning chains and structured evidence evaluation. The observed divergences from market sentiment, particularly in cases requiring analysis of complex quantitative or domain-specific evidence, demonstrate the potential value of complementing human prediction markets with agent-based forecasting systems. Even when our predictions differ significantly from market sentiment, they are accompanied by transparent reasoning processes and specific evidence consideration that can provide valuable complementary perspectives to human forecasters.

5 Discussion

Our analysis of resolved markets provides insights into the performance of our agent-based forecasting system. With an overall accuracy of 61% and a mean absolute error (MAE) of 0.3, the results indicate that while the system shows promise in probabilistic forecasting, there is room for improvement. These metrics reflect the system's capability to produce reasonably reliable predictions, though not consistently outperforming expectations in all scenarios. This study serves at the first baseline for this task.

An intriguing finding from the resolved market analysis is the apparent lack of correlation between jury uncertainty and forecast error. One might expect that higher uncertainty, as expressed by greater variance among agent jurors, would correspond to larger errors in predictions. However, our results suggest otherwise: predictions with high jury consensus (low uncertainty) do not necessarily exhibit lower error, nor do predictions with high uncertainty show a consistent trend toward greater error. This observation underscores the complexity of forecasting in uncertain environments and suggests that factors beyond internal agent confidence contribute significantly to prediction accuracy.

In regards to our findings for alignment with market sentiment, they show the distinct characteristics and potential value of agent-based forecasting systems in comparison to human-driven prediction markets. The divergence in prediction patterns between our system and human forecasters provides insights into the complementary roles these approaches can play in probabilistic forecasting. Human prediction markets excel at aggregating diverse perspectives, capturing a wide range of intuition and heuristics. However, they are susceptible to biases, uncertainty, and social dynamics. In contrast, our system provides a rigorous, evidence-driven alternative that offers transparency in decision-making and consistency in analyzing quantitative and qualitative data. The moderate alignment with market sentiment, combined with the system's capability to produce independently reasoned predictions, positions it as a complementary tool for improving forecasting accuracy.

6 Conclusion

We introduce a novel approach for forecasting that creates societies of LLM agents to make predictions. Our method is a structured ensemble where "lawyer" agents offer competing arguments to a jury which produces the final set of forecasts. We outperform the accuracy of human forecasters on Manifold Markets in closed competitions. In ongoing competitions, our forecasts loosely align with human predictors, but our method tends to produce more decisive predictions. This framework is flexible, and can easily be extended to more complex modes of deliberation.

Future work may explore different types of agent interaction. It may also ensemble different LLM models, or perform N rollouts of the deliberation process and aggregate the results. We expect that more advanced versions of search and context management will also improve accuracy. Optimizing evidence aggregation mechanisms and exploring hybrid systems that combine machine-driven analysis with human intuition could further enhance predictive performance. Additionally, adapting agent behavior to specific domains or market types may improve accuracy in context-specific scenarios, paving the way for more robust and actionable forecasting frameworks.

7 Ethics

The development of multi-agent forecasting systems using LLMs poses ethical considerations. We examine four key areas: accountability and transparency, potential for misuse, societal applications, and technical limitations.

Accountability and transparency. Our architecture provides inherent transparency advantages over single-model approaches since the deliberative process between agents creates explicit reasoning chains that can be analyzed and audited. Each agent's role and contributions are clearly documented through the sequential process in which agents interact. However, the LLM reasoning remains particularly opaque and the system's decisions could still reflect biases present in the training data or that emerge from unexpected agent interactions.

We recommend implementing several accountability measures to address this. For one, we recommend comprehensive logging of all agent interactions and decision processes, which is included in our approach and code. We also recommend clear documentation of agent roles and system parameters, regular evaluation against human expert benchmarks, and public release of our code and agent roles for community scrutiny.

Potential for misuse. While our system is designed for general forecasting applications, we acknowledge that there is potential for misuse of our system. For example, bad actors could potentially use similar systems to generate misleading forecasts that influence market behavior or public opinion. We mitigate this risk by emphasizing the limitations of the system and encouraging users to treat forecasts as complementary to existing human judgment.

Another concern is that advanced forecasting capabilities could be used to gain unfair advantages in competitive scenarios or manipulate decision-making processes. We recommend implementing access controls and usage monitoring in deployed systems. We also make our code public to help further research in this field. The system could also be modified to generate convincing but false narratives about future events. To counter this, we emphasize the importance of transparency in the reasoning process between agents to include verifiable evidence in the forecasting process.

Societal implications. There is a risk that decision-makers may over-rely on automated forecasts, potentially diminishing the role of human judgment and expertise. We emphasize that our system should augment rather than replace human decision-making processes. Advanced forecasting tools could exacerbate existing power imbalances if only available to well-resourced organizations. Our open-source approach aims to democratize access to these capabilities while encouraging responsible use.

Technical limitations. We acknowledge several important technical constraints of our approach. Despite providing probability estimates, our system's uncertainty quantification may not fully capture all sources of error and variability. Users should be aware of these limitations when making decisions based on the system's forecasts. The system's performance may vary significantly across different domains and types of forecasting tasks, necessitating careful validation before deployment in new contexts. Furthermore, the quality of forecasts depends heavily on the availability and reliability of input data, and users should be mindful of potential data biases and limitations.

To promote responsible development and deployment of agent-based forecasting systems, we recommend establishing clear guidelines for use, developing robust evaluation frameworks that consider both technical performance and ethical implications, and maintaining ongoing dialogue with stakeholders to understand and address concerns. Through these measures, we aim to realize the benefits of agent-based forecasting while minimizing potential harms.

References

S. Abdelnabi, A. Gomaa, S. Sivaprasad, L. Schönherr, and M. Fritz. Cooperation, competition, and maliciousness: Llm-stakeholders interactive negotiation, 2024. URL https://arxiv.org/abs/2309.17234.

C.-M. Chan, W. Chen, Y. Su, J. Yu, W. Xue, S. Zhang, J. Fu, and Z. Liu. Chateval: Towards better llm-based evaluators through multi-agent debate. *ArXiv*, abs/2308.07201, 2023. URL https://api.semanticscholar.org/CorpusID:260887105.

M. F. A. R. D. T. (FAIR)†, A. Bakhtin, N. Brown, E. Dinan, G. Farina, C. Flaherty, D. Fried, A. Goff, J. Gray, H. Hu, A. P. Jacob, M. Komeili, K. Konath, M. Kwon, A. Lerer, M. Lewis, A. H. Miller, S. Mitts, A. Renduchintala, S. Roller, D. Rowe, W. Shi, J. Spisak, A. Wei, D. Wu, H. Zhang, and M. Zijlstra. Human-level play in the game of <i>diplomacy</i> by combining language models with strategic reasoning. *Science*, 378(6624):1067–1074, 2022. doi: 10.1126/science.ade9097. URL https://www.science.org/doi/abs/10.1126/science.ade9097.

C. Gao, X. Lan, N. Li, Y. Yuan, J. Ding, Z. Zhou, F. Xu, and Y. Li. Large language models empowered agent-based modeling and simulation: A survey and perspectives, 2023. URL https://arxiv.org/abs/2312.11970.

M. Glockner, I. Staliūnaitė, J. Thorne, G. Vallejo, A. Vlachos, and I. Gurevych. Ambifc: Fact-checking ambiguous claims with evidence, 2023. URL https://arxiv.org/abs/2104.00640.

Z. Gu, X. Zhu, H. Guo, L. Zhang, Y. Cai, H. Shen, J. Chen, Z. Ye, Y. Dai, Y. Gao, Y. Hu, H. Feng, and Y. Xiao. Agentgroupchat: An interactive group chat simulacra for better eliciting emergent behavior, 2024. URL https://arxiv.org/abs/2403.13433.

D. Halawi, F. Zhang, C. Yueh-Han, and J. Steinhardt. Approaching human-level forecasting with language models, 2024. URL https://arxiv.org/abs/2402.18563.

E. Hsieh, P. Fu, and J. Chen. Reasoning and tools for human-level forecasting, 2024. URL https://arxiv.org/abs/2408.12036.

S. Noh and H.-C. H. Chang. Llms with personalities in multi-issue negotiation games, 2024. URL https://arxiv.org/abs/2405.05248.

J. S. Park, J. C. O'Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein. Generative agents: Interactive simulacra of human behavior, 2023. URL https://arxiv.org/abs/2304.03442.

J.-P. Rivera, G. Mukobi, A. Reuel, M. Lamparth, C. Smith, and J. Schneider. Escalation risks from language models in military and diplomatic decision-making. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 836–898. ACM, June 2024. doi: 10.1145/3630106.3658942. URL http://dx.doi.org/10.1145/3630106.3658942.

P. Schoenegger, I. Tuminauskaite, P. S. Park, and P. E. Tetlock. Wisdom of the silicon crowd: Llm ensemble prediction capabilities rival human crowd accuracy, 2024. URL https://arxiv.org/abs/2402.19379.

J. Su, C. Jiang, X. Jin, Y. Qiao, T. Xiao, H. Ma, R. Wei, Z. Jing, J. Xu, and J. Lin. Large language models for forecasting and anomaly detection: A systematic literature review, 2024. URL https://arxiv.org/abs/2402.10350.

G. Wang, Y. Xie, Y. Jiang, A. Mandlekar, C. Xiao, Y. Zhu, L. J. Fan, and A. Anandkumar. Voyager: An open-ended embodied agent with large language models. *Trans. Mach. Learn. Res.*, 2024, 2023. URL https://api.semanticscholar.org/CorpusID:258887849.

Q. Wu, G. Bansal, J. Zhang, Y. Wu, B. Li, E. Zhu, L. Jiang, X. Zhang, S. Zhang, J. Liu, A. H. Awadallah, R. W. White, D. Burger, and C. Wang. Autogen: Enabling next-gen llm applications via multi-agent conversation, 2023a. URL https://arxiv.org/abs/2308.08155.

Q. Wu, G. Bansal, J. Zhang, Y. Wu, S. Zhang, E. Zhu, B. Li, L. Jiang, X. Zhang, and C. Wang. Autogen: Enabling next-gen llm applications via multi-agent conversation framework. *ArXiv*, abs/2308.08155, 2023b. URL https://api.semanticscholar.org/CorpusID:260925901.