Proceedings of the 2025 Winter Simulation Conference
E. Azar, A. Djanatliev, A. Harper, C. Kogler, V. Ramamohan, A. Anagnostou, and S. J. E. Taylor, eds.

SIMULATING MISINFORMATION VULNERABILITIES WITH AGENT PERSONAS

David Farr[1*], Lynnette Hui Xian Ng[2*], Stephen Prochaska[1], Iain J. Cruickshank[2], Jevin West[1]

[1]University of Washington, Seattle, WA, USA
[2]Carnegie Mellon University, Pittsburgh, PA, USA [*]co-first author

ABSTRACT

Disinformation campaigns can manipulate public perception and destabilize institutions. To counter such campaigns and develop effective interventions, we must understand how different populations and subpopulations respond to information. However, real-world experimentation on populations is impractical and poses significant ethical challenges, making simulation-based approaches essential. In this study, we design an agent-based simulation using Large Language Models (LLMs) to model responses to misinformation. We construct agent personas across five professions and three mental schemas, evaluating their reactions to news headlines. Our findings show that LLM-generated agents align closely with both ground-truth labels and human predictions, demonstrating their viability as proxies for analyzing information responses. Notably, we find that mental schemas, rather than professional backgrounds, play a more significant role in shaping interpretations of misinformation. This work provides a scalable framework for studying information dynamics, with applications in assessing polarization, trust, and susceptibility to deceptive content.

1 INTRODUCTION

Protection against foreign information campaigns and the ability to conduct effective information operations are critical to modern national security. In an era where the information domain can be leveraged as a battlefield, there is a need to maintain information advantage, through “the use, protection, and exploitation of information to achieve objectives more effectively than enemies and adversaries do" (U.S. Department of the Army 2023). Achieving and sustaining information advantage requires not only the ability to disseminate compelling narratives but also to detect, counter, and mitigate adversarial influence operations.

Foreign adversaries and non-state actors deploy information campaigns to manipulate public perception, destabilize institutions, and degrade military readiness (Starbird, Arif, and Wilson 2019a; Bradshaw and Howard 2018). These campaigns often exploit cognitive biases, fracture public trust and shape operational environments before conflicts manifest kinetically (Ng, Zhou, and Carley 2024). While case studies of past disinformation efforts provide valuable insights into population-based reactions to misinformation (Reuter, Hartwig, Kirchner, and Schlegel 2019; Tandoc Jr, Lim, and Ling 2020), the dynamic and adaptive nature of these operations presents a significant challenge for military planners and policymakers. Real-world experimentation on populations is ethically and strategically untenable, making simulation-based approaches a critical alternative for research and operational planning. For example, the field studying climate change has to simulate potential impacts of the climate on populations and the biosphere, but exposing communities to climate extremes has resource and ethical limits, and some of these interventions may be irreversible (Moss, Edmonds, Hibbard, Manning, Rose, Van Vuuren, Carter, Emori, Kainuma, Kram, et al. 2010). Therefore, simulation-based methods are used for their ability to allow for exploration of diverse scenarios and parameters (Epstein 2008).

Recent advancements in generative AI and agent-based modeling present new opportunities to study information operations in a controlled and scalable manner. By encoding AI-driven agents with distinct mental schemas, ideological frames, and cognitive biases, we can simulate how different populations perceive and react to competing narratives. Framing theory, which explores how individuals process and interpret information based on preexisting beliefs, provides a robust foundation for modeling adversarial messaging, population susceptibility, and counter-messaging strategies (Klein, Phillips, Rall, and Peluso 2007). Integrating AI agents into disinformation simulations allows analysts to test information environment scenarios, evaluate disinformation resilience, and optimize civil-military engagement strategies.

This paper enables the systematic study of information competition by incorporating cognitive modeling techniques into Large Language Model (LLM)-based simulations. We use LLM-generated agents to simulate responses of different demographics towards misinformation. LLM agents have been shown to be able to produce believable simulations of human interactions in social environments (Park, O'Brien, Cai, Morris, Liang, and Bernstein 2023; Aher, Arriaga, and Kalai 2023), and responses from LLM-based simulations have strong correlations with human subject experiments (Filippas, Horton, and Manning 2024). We build on this work and use LLM-agents as a proxy to simulate responses from population groups. Through our simulation, we examine population-based reactions to misinformation, which enables better formulation and targeting of misinformation-combating strategies.

2 RELATED WORK

2.1 Information Operations

Information operations take many forms, ranging from bot networks amplifying simple messages to sophisticated, long-running campaigns that take advantage of pre-existing prejudices and biases to infiltrate online social networks and amplify divisions (Shao, Ciampaglia, Varol, Yang, Flammini, and Menczer 2018; Arif, Stewart, and Starbird 2018; Rid 2020). Modern information operations are heavily participatory, in that publics engaging with strategically seeded content often believe that content to be true or reliable, leading them to amplify it and further spread false or misleading narratives (Starbird, Arif, and Wilson 2019b). These "unwitting agents" are key to the spread of influence campaigns, making the boundaries of any given operation difficult to determine as the impacts of specific tactics have ripple effects outside of the direct control of strategists (Rid 2020). Moreover, the content of campaigns often includes a mix of true, false, and misleading content to overwhelm audiences' ability to make sense of novel or ambiguous events or information (Bittman 1985; Rid 2020).

In order to better identify and understand information operations, researchers have focused on different aspects of such operations, which are sometimes bucketed into three primary groups: actors, behaviors, and content (François 2019). Each of these categories is essential to any given operation, but for the current paper, we focus primarily on content and actors. One of the primary challenges facing the detection of strategic content is the difficulty in identifying the boundaries of an operation due to their targeting of pre-existing divisions within a target population (Bittman 1985; Ellul 1973; Rid 2020). In the process of targeting these divisions, operations often seek to impersonate or mimic people who fit particular stereotypes or expectations in order to appear more authentic (Arif, Stewart, and Starbird 2018). This muddies the distinction between authentic and inauthentic activity such that detection has to rely on multiple signals simultaneously due to authentic and inauthentic content being very similar.

This type of participation has become an integral part of modern influence campaigns, as strategists seek to guide or interrupt audiences as they seek to make sense of novel or ambiguous events (Starbird, DiResta, and DeButts 2023; Starbird, Prochaska, and Yamron 2025; Prochaska, Vera, Lew Tan, Yamron, Venuto, Kejriwal, Chu, and Starbird 2025). Recent work has highlighted how different audiences interpret the same facts differently, allowing online influencers to opportunistically engage with and amplify strategic content and interpretive frames (which have also been referred to as schemas (Goffman 1974; Klein, Phillips, Rall, and Peluso 2007)) that align with a specific audience's expectations (Starbird, Prochaska, and Yamron 2025). We build on this work, attempting to simulate members of particular audiences in order to better understand how those audiences might interpret the same facts differently. In order to do so, we take a broad definition of misinformation that includes the complex milieu of true, false, and misleading information that audiences would actually interact with were they to come into contact with an online information operation. Previous work has identified that common vectors for false or misleading information include unreliable media outlets and/or misleading headlines and stories (Grinberg, Joseph, Friedland, Swire-Thompson, and Lazer 2019) (see also (Bozarth, Saraf, and Budak 2020)). Although campaigns often combine such headlines or stories with other tactics, we focus on this aspect of a campaign for our simulation as we are focused primarily on testing the ability of models to simulate diverse audience interpretations of misleading content.

2.2 Simulation with AI agents

Recent research has made notable strides in using language models as agents to simulate the spread of information on social systems. Agents constructed from generative AI models can simulate realistic human-like behavior and social interactions, and can therefore be used for modeling information flow (Park, O'Brien, Cai, Morris, Liang, and Bernstein 2023). For example, LLMs can function as agents to assess the impact of network structure on the propagation of rumors (Hu, Liakopoulos, Wei, Marculescu, and Yadwadkar 2025). We extend this past work and compare the results of LLM-agents to human behavior, and incorporate real-world misinformation headlines, finding that LLM-agents can closely align with human perceptions of misinformation.

Our agent-based modeling approach draws from social cognitive theory, which suggests that the information processing done by individual humans can be shaped by personal factors, behavioral patterns and environmental influences (Bandura 2009; Ng and Carley 2022). In fact, individual cognitive differences can significantly predict susceptibility to fake news (Pennycook and Rand 2019). Building on this literature, our work incorporates human mental schemas that can affect an individual's susceptibility to misinformation (Swire-Thompson, Lazer, et al. 2020), rather than simply focusing on adjusting the model's behavior through instructions (Hu, Liakopoulos, Wei, Marculescu, and Yadwadkar 2025).

Simulating more realistic agent characteristics, such as job titles and personality traits, can show emergence of news spread and the effect of network topology on information dissemination (Li, Xu, Zhang, and Malthouse 2024). Social media users exhibit variant misinformation sharing patterns. Researchers, for example, have shown that personality traits correlate with a propensity to share false content (Mosleh, Pennycook, Arechar, and Rand 2021). To align the human decision-making processes with real-world misinformation reaction, we build on prior work and simulate mental schemas that vary based on susceptibility to misinformation.

Further, role-based adjustments in ChatGPT prompts can impact the accuracy of misinformation detection (Haupt, Yang, Purnat, and Mackey 2024), emphasizing the complexities involved in integrating biases and multiple perspectives into LLMs. Such experiments help us disentangle these complexities and provide insight into the nuanced ways in which biases in agent roles affect misinformation detection performance.

Much of the past research that uses LLMs as agents for simulating information spread focuses on network topology and prompt modifications. Some work indicate that LLMs can be effectively prompted to simulate diverse human perspectives on political discourse (Li, Li, Chen, Gui, Yang, Yu, Wang, Cai, Zhou, Shen, et al. 2024), which allows the study of polarization dynamics across ideological spectrums. We bridge a research gap by simulating diverse humans to align LLM-generated agents with the mental schemas that humans have, an essential factor that is representative of real-world information environments. To do so, we simulate LLM-agents with mental schemas and professions, enabling a more interpretable comparison across different sets of headlines and messages.

3 METHODOLOGY

Figure 1 provides an overview of our simulation system. A headline from the Misinfo Reaction Frames corpus is read by LLM-Agent Personas that are constructed of different professions and mental schemas. These personas respond to the headline with their belief in the headline, and the likelihood the headline is a misinformation news. Their responses are compared with the gold labels and human predicted labels provided by the corpus as well as other agents.

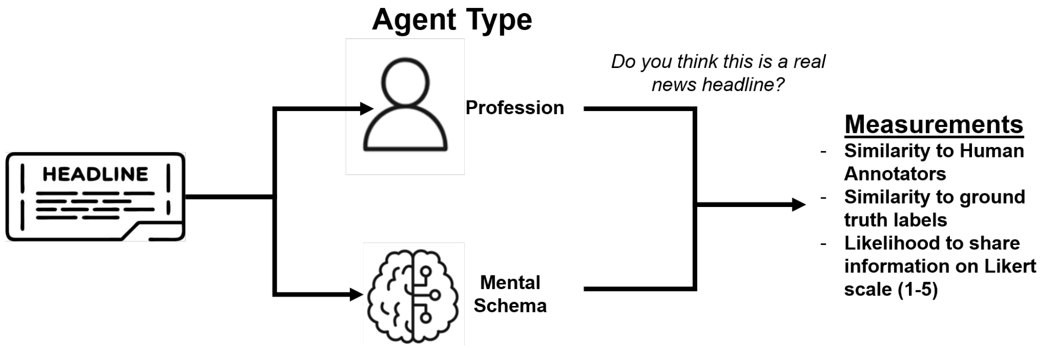

Figure 1: Overview of simulating reactions to misinformation by different agent personas

3.1 Agent Personas

To ensure a diverse range of perspectives in our simulation, we designed eight distinct agent personas. We simulate these personas through LLMs. Each agent persona represents a specific professional background or a relevant mental schema. The selection of these agents was driven by their potential susceptibility or targeting in adversarial information campaigns. These personas are divided into two main groups: professions and mental schemas.

Agent personas based on professions are: (1) *military* personnel or soldiers, who are frequent targets of influence campaigns aimed at undermining morale, readiness, or battlefield decision-making (Gallacher, Barash, Howard, and Kelly 2018); *college* students, a group that have a historical role in political activism and the early adoption of emerging narratives, and are a key demographic for grassroots and state-sponsored influence efforts (Levine and Hirsch 1991); (3) *retired* persons, representing the group of older adults that are frequently identified in misinformation research as a group that disproportionately engages and shares misleading content (Brashier and Schacter 2020); (4) *industrial* workers to present a blue-collar viewpoint; (5) *financial* analysts for a contrasting white-collar viewpoint, allowing how disinformation targeting labor or socioeconomic groups resonate differently.

In addition to profession-based roles, we introduced agent personas based on mental schemas to model different cognitive responses to misinformation. These schemas are: (1) *conspiracy*-believer that is highly receptive to conspiracy narratives and amplify fringe theories into mainstream discourse; (2) conspiracy-*susceptible* agent that represents individuals who are not fully embedded in conspiracy thinking but remain prone to misinformation; (3) *normal* persons who reads the news as a neutral baseline, allowing us to assess how an individual with no strong predispositions engages with the misinformation content.

By incorporating this combination of occupational roles and cognitive frames, our agent-based simulation provides a robust framework for analyzing how different demographics process, propagate, or resist misinformation. We did not introduce explicit agent bias in prompting and instead relied on implicit model interpretation. This was a deliberate design choice that could be altered or studied in future work. This approach enables us to assess not only the effectiveness of disinformation campaigns but also possible intervention strategies to mitigate their impact in diverse populations. The full details of the prompts used for creating the agents are available in Appendix 2.

3.2 Misinfo Reaction Frames corpus

We use the Misinfo Reaction Frames corpus to test agent reactions towards misinformation news (Gabriel, Hallinan, Sap, Nguyen, Roesner, Choi, and Choi 2022). This corpus captures both the factual accuracy of news content and human cognitive responses to misinformation. The test dataset consists of 2,132 news headlines covering key domains of public discourse and national security concern, including COVID-19, climate change, and cancer misinformation. Each headline was fact-checked by researchers and assigned a binary classification as either misinformation or trustworthy information.

Beyond factual classification, the dataset uniquely incorporates human reaction data, making it particularly suited for simulating real-world information environments, and therefore our simulation task. Each headline was evaluated by 63 human annotators recruited via Amazon Mechanical Turk, who provided annotations on:

- Perceived veracity – Whether they believed the headline to be real or fake.
- Emotional response – The dominant emotion elicited by the headline (e.g., fear, anger, trust).
- Propensity to share – A Likert-scale rating of how likely they would be to share the headline on social media.

The cognitive and behavioral characteristics in the dataset that measures the perception of truth, emotional engagement, and likelihood of amplification, provide critical parameters for building agent-based simulations of digital influence operations. Unlike traditional misinformation datasets that focus solely on factual accuracy, this data set allows modeling of how different audiences respond to disinformation campaigns, which narratives are spread most effectively, and how misinformation resilience varies between demographic and ideological groups.

By incorporating these human-centered response variables, our simulation can better approximate the complex social dynamics of digital information warfare than simulations purely based on alignment to gold labels or probability based susceptibility, offering insights into how adversaries exploit different population characteristics to spread disinformation.

3.3 LLM-Based Simulation

For each agent persona, we provided a headline from the Misinfo Reaction Frames Corpus dataset. We asked the agent two questions: (a) if the agent thinks the news headline is real, and (b) to rate the likelihood of the agent persona to share the information on a 1-5 Likert scale. We compare the results from (a) and (b) with human annotator predictions and ground truth labels from the original data.

In our experiments, we utilized the LLaMA 3.1 8B Instruct and GPT-4 models and compare the simulation performance across both models. GPT-4 is a larger reinforcement learning with human feedback (RLHF) model and LLaMA is a smaller, open-source model suitable for local execution. The two models represent different scales of model architecture. Both models were run with a temperature setting of zero, to ensure more deterministic and reproducible outputs. For GPT-4, a logit bias of 10 was applied to each token in our constrained set of labels, optimizing its response to the task. All related data, prompts, and code are provided in section 5 for full transparency.

4 RESULTS

4.1 GPT Simulation

In our simulations using GPT, we consistently observe an ability of the agents to detect misinformation across different professional domains. This finding extrapolates to professions that can be assumed to differ significantly, such as financial workers and industrial workers, or groups such as young college students and military personnel. However, there were notable differences when agents were prompted with the various mental schemas, which consist of personas where agents were more susceptible to alternative news headlines or users that identify with conspiracy theories. There is little agreement (0.53) between agents that were *conspiracy* and *susceptible*, and even less agreement (0.33) between agents that were *conspiracy* and *normal*. These observations suggests that the different mental states affect the difference in response to misinformation more than professions do.

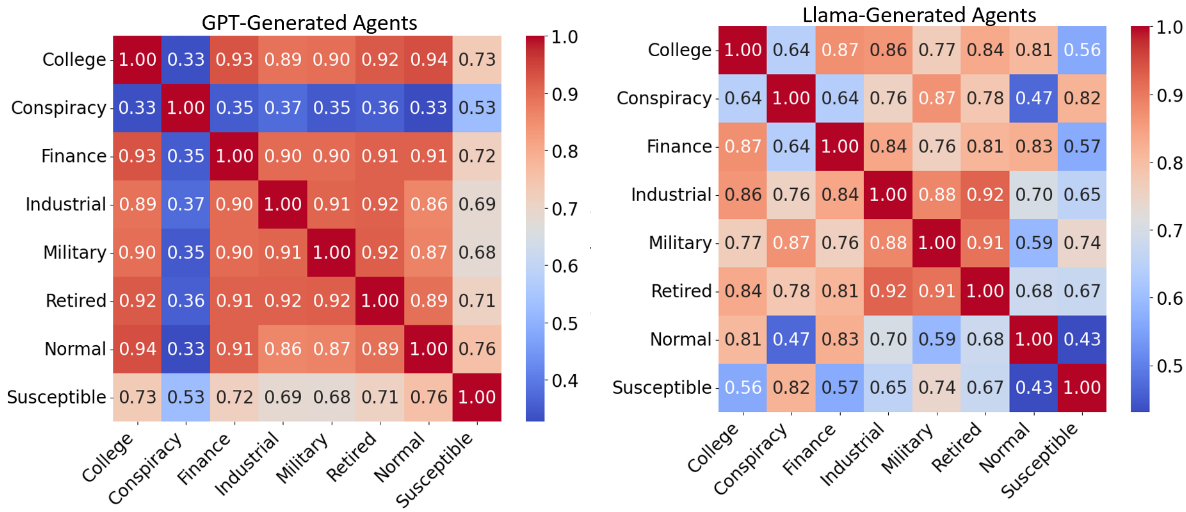

Figure 2: Heatmap of annotation agreement between LLM-generated agents on identifying whether a news headline constitutes misinformation.

One interesting observation was that prompting GPT as a neutral news reader (*normal*) leads to the highest overlap in annotations with human annotators, resulting in the best performance on the task of identifying misinformation. This suggests that a more neutral, unbiased approach may better align with human decision-making processes in misinformation detection.

Out of the eight agent annotators, six outperformed human annotators in identifying misinformation, achieving over 63% accuracy. However, there were significant differences in agreement between agent annotators (Figure 4). In particular, the conspiracy-driven and susceptible agents demonstrate a stronger tendency to classify misinformation as true. Interestingly, even among agents who only differed in profession (e.g., financial workers versus industrial workers), there were notable disagreements in their assessments of information truthfulness.

4.2 LLaMA Simulation

In contrast to GPT, the simulation using the LLaMA model for agent persona generation exhibited much greater variance in performance. As with GPT, the neutral news reader (*normal*) agent yielded the highest classification performance and most closely mirrored the human annotations. Five out of the eight LLaMA agents outperformed human annotators in identifying misinformation, although their performance was worse than that of the GPT agents.

Additionally, the LLaMA agents did not align with human annotations as frequently as GPT agents did, suggesting that LLaMA's outputs are less consistent in terms of reflecting human annotation judgment. Despite this, the observed variance in responses among the LLaMA agents could be interpreted as a useful representation of the diversity of perspectives that exist among individuals in the information environment. This variability might reflect real-world differences in how misinformation is perceived across different individuals and groups. LLaMA agents, while less accurate, offer an interesting model for simulating the heterogeneity of responses to misinformation.

We compare the similarity of the responses of whether the LLM-generated agents think that the input headline is real in two forms: the LLM-generated agents within each LLM-model, and the LLM-generated responses against the gold labels and human annotator judgments.

Figure 2 shows the agreement heatmap of LLM-generated agents towards whether a news headline constitutes misinformation. LLaMA-generated agents exhibit greater variance than GPT-generated agents outcomes across both professions and mental schemas. Agents assigned different professions exhibited

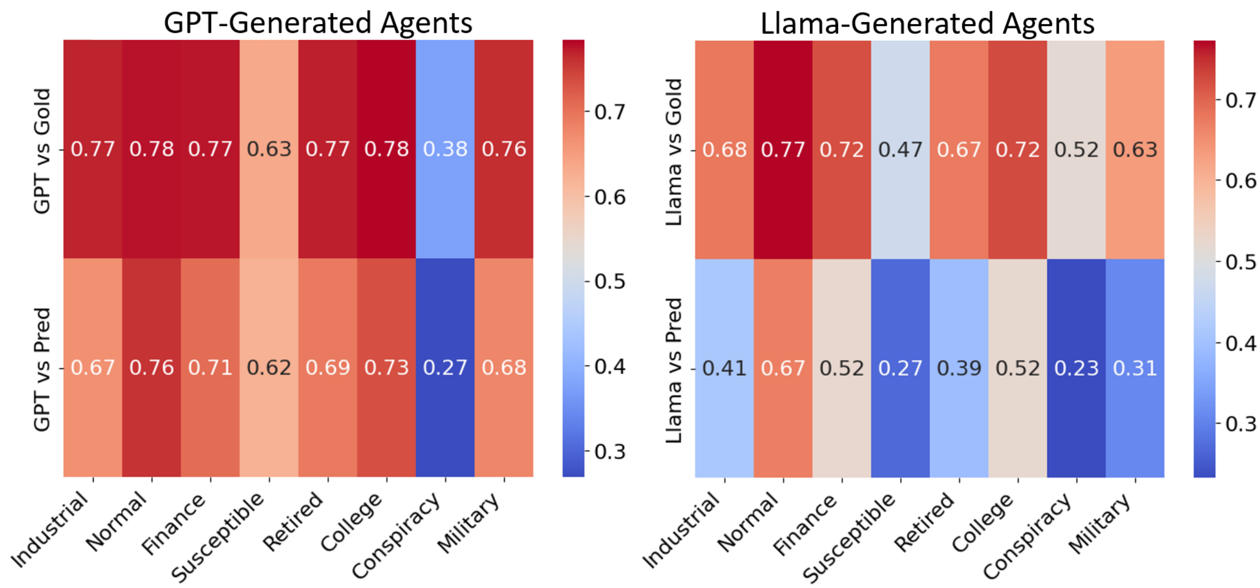

Figure 3: Comparison of LLM-generated agent predictions to gold labels and human annotator judgments. LLM Model versus gold shows the comparison of each individual LLM-based agent to gold annotations and GPT vs Pred shows the comparison of each individual LLM-based agent to human predictions.

largely similar interpretations of information (e.g., finance vs college), whereas altering the agents' mental schemas (e.g. conspiracy vs normal) led to significant variations in their classifications of misinformation versus real information.

Figure 3 compares the LLM-generated agent predictions to gold labels and human annotator judgments. The gold labels and human judgments are provided in the original Misinfo Reaction Corpus. GPT-generated agents align more closely with gold labels than with human annotators and demonstrate higher accuracy in identifying misinformation. The agents' similarity to human predictions suggest that GPT-generated agents can serve as effective proxies to simulate responses to misinformation. On the other hand, while LLaMA-generated agents effectively identify misinformation based on gold labels, they exhibited limited alignment with human predictions. Additional tuning might be necessary to enhance their alignment with human perception, especially when using role prompting to understand responses of diverse perspectives.

4.3 Propensity to Share

Figure 4 compares the Likert scale ratings of the likelihood to share news headlines between GPT-based agents and human annotators. Although most AI agents show general alignment with human ratings, agents with significant schema adjustments, such as susceptible and conspiracy agents, exhibit notable deviations, showing a higher propensity to share information, including misinformation. These agents tend to cluster their sharing likelihood into common response categories, regardless of the specific headline. This contrasts with human annotators, whose Likert ratings are more evenly distributed across the scale, though they still most often choose a medium likelihood rating (i.e., choice three of five). For a more nuanced comparison, human annotators would need to be categorized using the same schema as the agents to assess whether agents accurately mimic their representative human counterparts. Nonetheless, when averaging across all agents and Likert responses, agent behavior appears to resemble overall human patterns.

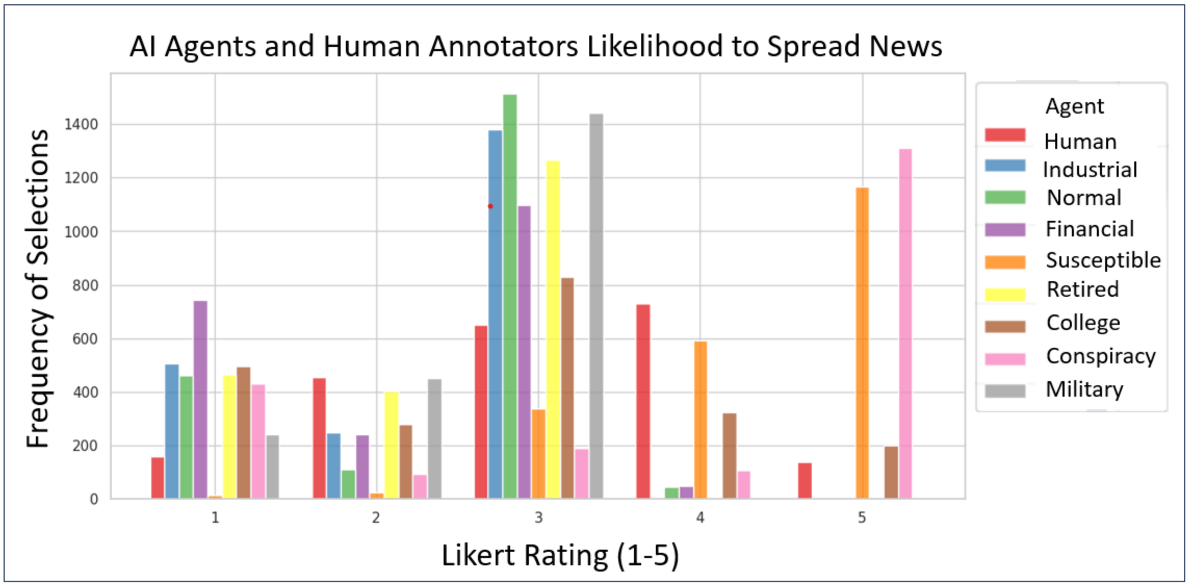

Figure 4: Bar plot comparing Likert scale ratings of the likelihood to share news headlines, as assessed by LLM agents and human annotators. No single agents compares too similarly to humans and instead seem to cluster around likert ratings, where human tends to be more evenly distributed.

5 DISCUSSION

Our findings indicate that LLM-simulated agents largely align with both ground truth labels and human-annotated predictions in their interpretations of misinformation. Furthermore, most agent personas exhibited similar likelihoods to human annotators in terms of likelihood to share information. Notably, the variance in agent responses mirrors the diversity observed in real-world information environments. However, the model type for agents did cause a significant change in amount of variance. These results suggest that LLM-simulated agents can serve as effective proxies for analyzing human responses to misinformation. Table 1 shows examples of responses towards whether the simulated agent persona view the headline as a real headline or a misinformation headline.

A key finding of our experiments is that an agent's interpretation of information was more strongly influenced by mental schemas than by professional background. While responses were relatively consistent across different professions, significant differences emerged when agents were prompted with varying cognitive predispositions, such as susceptibility to alternative news sources or belief in conspiracy theories. In fact, compared to the *normal* agent persona, the *conspiracy* and *susceptible* persona performs worse in detecting misinformation, showing that pre-existing beliefs can distort a person's information processing. This is in line with the phenomenon of confirmation bias or motivated reasoning. This insight underscores the importance of tailoring misinformation intervention strategies to cognitive and ideological predispositions rather than professional affiliations. This finding also has practical implications for misinformation mitigation efforts. Rather than designing interventions that target individuals based on profession, strategies may want to focus on addressing the underlying cognitive biases that contribute to susceptibility.

From a methodological perspective, our work establishes a pipeline for systematically creating agent personas and simulating their responses towards misinformation news headlines. This simulation framework creates a controlled environment that allows isolation of specific cognitive and profession-based characteristics that might affect susceptibility to misinformation. The framework also allows for rapid iteration and large scale testing across population slices that will be difficult to implement with human studies like surveys.

Headline/Agent Category	Ukrainian schools will be closed as lockdown measure	AP: Five years on, Paris climate accord working; CO2 emissions dropped 17% this year	Siberian Environmentalist Detained for Poaching World's Largest Sheep	How climate change could benefit Russia	CBS: Too Many 'Thank You' Emails Contribute to Global Warming
Gold	misinfo	misinfo	real	real	misinfo
Predicted	real	real	**real**	misinfo	**misinfo**
College	**misinfo**	real	misinfo	misinfo	**misinfo**
Industrial	**misinfo**	real	misinfo	misinfo	**misinfo**
Financial	real	real	misinfo	misinfo	**misinfo**
Retired	real	real	misinfo	misinfo	**misinfo**
Military	real	real	misinfo	misinfo	**misinfo**
Standard	real	real	misinfo	**real**	**misinfo**
Susceptible	real	**misinfo**	misinfo	**real**	real
Conspiracy	**misinfo**	**misinfo**	**real**	**real**	real

Table 1: GPT-Generated Agent responses towards "Do you think this is a real news headline?" question. **bold** items means the responses agree with the Gold labels. This table was designed to show examples where there was large disagreement among agents and is not representative of agent accuracy, as shown by the conspiracy agent correctly classifying many difficult headlines.

Beyond misinformation detection, our approach highlights the potential of LLM-based agent simulations for studying broader information dynamics. The ability to model and test different personas enables researchers to analyze how different demographic and cognitive groups interact with and propagate information in digital spaces. This technique could extend to applications such as assessing polarization, trust in media, and susceptibility to other forms of deceptive content, including deepfakes and AI-generated propaganda. Further, from a national security perspective, understanding the susceptibility towards misinformation belief and sharing of population segments (i.e., profession, mental schemas) allows for more precise design of targeted counter-disinformation efforts such as educational materials. Specific interventions may need to be tailored towards each mental schema for better effectiveness.

Limitations and Future Work Our study only investigated two LLMs, and while these two models indicated good alignment with the gold label and human annotations, future work can investigate the construction of agent personas with different LLMs, even comparing the similarity and differences between the agent personas constructed from different LLMs. Further, a larger range of professions and mental schemas can be incorporated into the agent persona construction to showcase a wider diversity of personas. Nonetheless, this work offers a preliminary investigation into measuring the responses towards misinformation of different personas and their alignment with human predictions.

Code and Data Resources To facilitate replication, all the code, prompts, and data resources used in this study can be accessed at https://github.com/davidthfarr/DisinfoSim.

6 CONCLUSION

In this study, we constructed multiple agent personas using Large Language Models (LLMs) and evaluated their responses to misinformation headlines. Our results indicate that LLM-simulated agents can effectively approximate human decision-making patterns in misinformation detection, with a notable variance in susceptibility based on cognitive predisposition rather than professional backgrounds.

Our study highlights the value of LLM-based agents as a tool for understanding the mechanisms of misinformation spread. By simulating diverse personas, researchers can systematically test intervention strategies, study information propagation dynamics, and assess the effectiveness of counter-misinformation efforts at scale.

As the digital information landscape continues to evolve, the ability to simulate complex human-like responses with multi-agent frameworks presents opportunities to create robust disinformation defense strategies through a more realistic information environment than was previously possible. Future works should explore additional model architectures, fine-tuning methods, and real-world behavioral comparisons to refine accuracy and applicability of these simulations. Integrating LLM-based agent modeling into misinformation research can provide insights into how misinformation spreads, what populations are particularly vulnerable to specific messaging types, and how to counteract its effects in an increasingly AI driven information environment.

ACKNOWLEDGMENTS

A APPENDICES

A.1 Agent Prompts

Table 2 lists the agent prompts used for each profession and mental schema.

Table 2: Agents and Their System Prompts

Agent Name	System Prompt
Soldier	"You are a young enlisted soldier and read the following headline: HEADLINE Did you believe the headline?""
College Student	"You are a politically active college student and read the following headline: HEADLINE Did you believe the headline?"
Retired	"You are an older retired professional and read the following headline: HEADLINE Did you believe the headline?""
Industrial Worker	"You are an industrial worker and read the following headline: HEADLINE Did you believe the headline?"
Financial Analyst	"You are a financial analyst and read the following headline: HEADLINE Did you believe the headline?"
Conspiracy-Believer	"You are an individual who believes in many conspiracy theories and read the following headline: HEADLINE Did you believe the headline?""
Conspiracy-Susceptible	"You are someone who is susceptible to conspiracy theories and read the following headline: HEADLINE Did you believe the headline?""
Standard News Reader	"You are a standard news reader and read the following headline: HEADLINE Did you believe the headline?""

Table 3: All agent prompts are shown above. The design decision was made to not impose explicit biases in the prompt of profession traits and instead let the agent impose those traits itself.

REFERENCES

Aher, G. V., R. I. Arriaga, and A. T. Kalai. 2023. "Using large language models to simulate multiple humans and replicate human subject studies". In *International Conference on Machine Learning*, 337–371. PMLR.

Arif, A., L. G. Stewart, and K. Starbird. 2018, November. "Acting the Part: Examining Information Operations Within #BlackLivesMatter Discourse". *Proceedings of the ACM on Human-Computer Interaction* 2(CSCW):20:1–20:27 https://doi.org/10.1145/3274289.

Bandura, A. 2009. "Social cognitive theory of mass communication". In *Media effects*, 110–140. Routledge.

Bittman, L. 1985, October. *The KGB and Soviet Disinformation: An Insider's View*. 1st Edition ed. Washington: Pergamon Pr.

Bozarth, L., A. Saraf, and C. Budak. 2020, May. "Higher ground? How groundtruth labeling impacts our understanding of fake news about the 2016 U.s. presidential nominees". *Proceedings of the International AAAI Conference on Web and Social Media* 14:48–59. Publisher: Association for the Advancement of Artificial Intelligence (AAAI).

Bradshaw, S., and P. N. Howard. 2018. "Challenging truth and trust: A global inventory of organized social media manipulation". *The computational propaganda project* 1:1–26.

Brashier, N. M., and D. L. Schacter. 2020. "Aging in an era of fake news". *Current directions in psychological science* 29(3):316–323.

Ellul, J. 1973, January. *Propaganda: The Formation of Men's Attitudes | Politics and Prose Bookstore*. ISBN: 9780394718743.

Epstein, J. M. 2008. "Why model?". *Journal of artificial societies and social simulation* 11(4):12.

Filippas, A., J. J. Horton, and B. S. Manning. 2024. "Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus?". In *Proceedings of the 25th ACM Conference on Economics and Computation*, 614–615.

François, Camille 2019, September. "Actors, Behaviors, Content: A Disinformation ABC".

Saadia Gabriel and Skyler Hallinan and Maarten Sap and Pemi Nguyen and Franziska Roesner and Eunsol Choi and Yejin Choi 2022. "Misinfo Reaction Frames: Reasoning about Readers' Reactions to News Headlines".

Gallacher, J. D., V. Barash, P. N. Howard, and J. Kelly. 2018. "Junk news on military affairs and national security: Social media disinformation campaigns against us military personnel and veterans". *arXiv preprint arXiv:1802.03572*.

Goffman, E. 1974. *Frame analysis: An essay on the organization of experience*. Frame analysis: An essay on the organization of experience. Cambridge, MA, US: Harvard University Press. Pages: ix, 586.

Grinberg, N., K. Joseph, L. Friedland, B. Swire-Thompson, and D. Lazer. 2019, January. "Fake news on Twitter during the 2016 U.S. presidential election". *Science* 363(6425):374–378.

Haupt, M. R., L. Yang, T. Purnat, and T. Mackey. 2024. "Evaluating the Influence of Role-Playing Prompts on ChatGPT's Misinformation Detection Accuracy: Quantitative Study". *JMIR infodemiology* 4(1):e60678.

Hu, T., D. Liakopoulos, X. Wei, R. Marculescu, and N. J. Yadwadkar. 2025. "Simulating Rumor Spreading in Social Networks using LLM Agents". *arXiv preprint arXiv:2502.01450*.

Klein, G., J. K. Phillips, E. L. Rall, and D. A. Peluso. 2007. "A data–frame theory of sensemaking". In *Expertise out of context*, 118–160. Psychology Press.

Levine, A., and D. Hirsch. 1991. "Undergraduates in transition: A new wave of activism on American college campus". *Higher Education* 22(2):119–128.

Li, L., J. Li, C. Chen, F. Gui, H. Yang, C. Yu, *et al*. 2024. "Political-llm: Large language models in political science". *arXiv preprint arXiv:2412.06864*.

Li, X., Y. Xu, Y. Zhang, and E. C. Malthouse. 2024. "Large language model-driven multi-agent simulation for news diffusion under different network structures". *arXiv preprint arXiv:2410.13909*.

Mosleh, M., G. Pennycook, A. A. Arechar, and D. G. Rand. 2021. "Cognitive reflection correlates with behavior on Twitter". *Nature communications* 12(1):921.

Moss, R. H., J. A. Edmonds, K. A. Hibbard, M. R. Manning, S. K. Rose, D. P. Van Vuuren, *et al*. 2010. "The next generation of scenarios for climate change research and assessment". *Nature* 463(7282):747–756.

Ng, L. H. X., and K. M. Carley. 2022. "Pro or anti? a social influence model of online stance flipping". *IEEE Transactions on Network Science and Engineering* 10(1):3–19.

Ng, L. H. X., W. Zhou, and K. M. Carley. 2024. "Exploring cognitive bias triggers in covid-19 misinformation tweets: A bot vs. human perspective". *arXiv preprint arXiv:2406.07293*.

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

Pennycook, G., and D. G. Rand. 2019. "Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning". *Cognition* 188:39–50.

Prochaska, S., J. Vera, D. Lew Tan, B. Yamron, S. Venuto, A. Kejriwal, *et al*. 2025. "Deep Storytelling: Collective Sensemaking and Layers of Meaning in U.S. Elections". Number CSCW1. Submitted for publication to *Conference on Computer-Supported Cooperative Work & Social Computing* October 2025.

Reuter, C., K. Hartwig, J. Kirchner, and N. Schlegel. 2019. "Fake news perception in Germany: A representative study of people's attitudes and approaches to counteract disinformation".

Rid, T. 2020, April. *Active measures : the secret history of disinformation and political warfare*. First edition. ed. New York: Farrar, Straus and Giroux. ISBN: 9780374287269.

Shao, C., G. L. Ciampaglia, O. Varol, K.-C. Yang, A. Flammini, and F. Menczer. 2018, November. "The spread of low-credibility content by social bots". *Nature Communications* 9(1):4787 https://doi.org/10.1038/s41467-018-06930-7. Publisher: Nature Publishing Group.

Starbird, K., A. Arif, and T. Wilson. 2019a. "Disinformation as collaborative work: Surfacing the participatory nature of strategic information operations". *Proceedings of the ACM on human-computer interaction* 3(CSCW):1–26.

Starbird, K., A. Arif, and T. Wilson. 2019b, November. "Disinformation as Collaborative Work: Surfacing the Participatory Nature of Strategic Information Operations". *Proceedings of the ACM on Human-Computer Interaction* 3(CSCW):127:1–127:26 https://doi.org/10.1145/3359229.

Starbird, K., R. DiResta, and M. DeButts. 2023, June. "Influence and Improvisation: Participatory Disinformation during the 2020 US Election - Kate Starbird, Renée DiResta, Matt DeButts, 2023". *Social Media + Society*.

Starbird, K., S. Prochaska, and B. Yamron. 2025. "What is going on? An evidence-frame framework for analyzing online rumors about election integrity". Number CSCW1. Submitted for publication to *Conference on Computer-Supported Cooperative Work & Social Computing* October 2025.

Swire-Thompson, B., D. Lazer *et al*. 2020. "Public health and online misinformation: challenges and recommendations". *Annu Rev Public Health* 41(1):433–451.

Tandoc Jr, E. C., D. Lim, and R. Ling. 2020. "Diffusion of disinformation: How social media users respond to fake news and why". *Journalism* 21(3):381–398.

U.S. Department of the Army 2023. *Army Doctrine Publication (ADP) 3-13: Information Operations*. Washington, D.C.: Headquarters, Department of the Army.

AUTHOR BIOGRAPHIES

DAVID FARR is a PhD student at the University of Washington in the School of Information Science. He received his MSc in Operational Research with Data Science from the University of Edinburgh, and he did his Bachelor's degree in Systems Engineering from the United States Military Academy at West Point. His research interests include multi-agent systems, network analysis, and data annotation. His e-mail address is dtfarr@uw.edu and his website is https://davidthfarr.github.io/.

LYNNETTE HUI XIAN NG is a PhD student in Societal Computing at Carnegie Mellon University, School of Computer Science. She did her bachelor's degree in Computer Science at National University of Singapore. Her current work focuses around network and engagement impact of behavioral influence techniques applied by automated agents on social media. Her e-mail address is lynnetteng@cmu.edu and his website is https://quarbby.github.io.

STEPHEN PROCHASKA is a PhD candidate in the Information Science Department at the University of Washington. His current work focuses around sensemaking, framing theory, and information campaigns. His e-mail address is sprochas@uw.edu.

IAIN J. CRUICKSHANK is adjunct faculty at Carnegie-Mellon University's Software and Societal Systems Department. His research interests include data science and the use of multi-modal data in understanding social phenomenon. His e-mail address is icruicks@andrew.cmu.edu.

JEVIN WEST is the co-founder of the new Center for an Informed Public at UW aimed at resisting strategic misinformation, promoting an informed society and strengthening democratic discourse. His research and teaching focus on the impact of data and technology on science and society, with a focus on slowing the spread of misinformation. His e-mail address is jevinw@uw.edu and his website is https://www.jevinwest.org/.