Competing LLM Agents in a Non-Cooperative Game of Opinion Polarisation

Anonymous ACL submission

Abstract

We introduce a novel non-cooperative game to analyse opinion formation and resistance, incorporating principles from social psychology such as confirmation bias, resource constraints, and influence penalties. Our simulation features Large Language Model (LLM) agents competing to influence a population, with penalties imposed for generating messages that propagate or counter misinformation. This framework integrates resource optimisation into the agents' decision-making process. Our findings demonstrate that while higher confirmation bias strengthens opinion alignment within groups, it also exacerbates overall polarisation. Conversely, lower confirmation bias leads to fragmented opinions and limited shifts in individual beliefs. Investing heavily in a high-resource debunking strategy can initially align the population with the debunking agent, but risks rapid resource depletion and diminished long-term influence.

1 Introduction and Background

The study of opinion dynamics, originating from efforts to understand how individuals modify their views under social influence (Kelman, 1958, 1961), has broad applications in areas such as public health campaigns, conflict resolution, and combating misinformation. Within social networks, opinions spread and evolve, influenced by various factors including peer interactions (Kandel, 1986), media exposure (Zucker, 1978), and group dynamics (Friedkin and Johnsen, 2011). Developing accurate models of these processes is essential not only for predicting trends like opinion polarisation (Tan et al., 2024) or consensus formation but also for crafting targeted interventions to mitigate harmful effects, such as the spread of misinformation or societal fragmentation (Hegselmann and Krause, 2015). Agent-based models (ABMs) simulate interactions among individual agents to examine the emergent properties of opinion dynamics. These models offer robust frameworks for analysing complex scenarios (Deffuant et al., 2002; Mathias et al., 2016), evaluating strategies to reduce negative consequences, and potentially fostering constructive social influence by integrating explicit cognitive mechanisms into opinion-updating processes.

This work investigates how LLMs can model human-like opinion dynamics and influence propagation within social networks. Traditional ABMs often employ simplified rules that fail to capture the complexity of human communicative strategies. To address this limitation, we introduce a novel non-cooperative game framework where adversarial LLMs, one spreading misinformation and the other countering it, interact. This work introduces a non-cooperative game where LLM agents engage in adversarial interactions to model misinformation spread and countering. Unlike prior studies (Wang et al., 2025; Chuang et al., 2024) on passive opinion evolution and nudging, it focuses on resource-constrained influence operations and debunking effectiveness in competitive environments.

We pose the following research questions:

RQ1 What are the emergent behaviors in networks of agents influenced by competing LLMs?

RQ2 How does the competition between LLM agents shape the evolution of opinion clusters over time, also known as echo-chambers?

2 Methodology

We use LLMs to simulate the propagation and debunking of misinformation on social media within a non-cooperative game framework.

Scenario: Our scenario models an asymmetric information environment, highlighting the challenges faced by the "Blue Team" (countering misinformation). This setup reflects adversarial dynamics commonly seen in serious games or wargames (Paul, Christopher and Connable, Ben and Welch

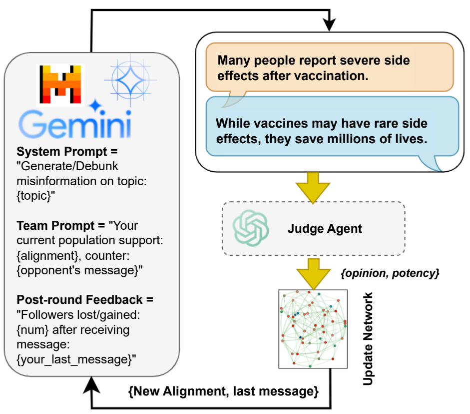

Figure 1: In each round, the active team (Red/Blue) generates a message that receives a potency value from the judge. The network updates according to the BCM algorithm. In the next round, the opposing team receives the potency results of their rival's message and their own from the previous round.

Jonathan and Rosenblatt, Nate and McNeive, Jim, 2021). The "Red Team" and "Blue Team" construct, familiar in cybersecurity practices, is adapted from NIST's Glossary (National Institute of Standards and Technology (NIST), 2015). The system comprises two LLM-based agents: the *Red Agent* spreads misinformation, while the *Blue Agent* debunks it. These agents operate within a directed network of neutral agents, termed *Green Nodes*, representing individuals in a population. Figure 1 illustrates this non-cooperative game structure.

Agent Roles and Mechanics: The simulation incorporates the following agent roles:

The **Red Agent** aims to amplify doubt and confusion by generating misinformation of varying potency. Messages of higher potency incur penalties through rejection, reflecting real-world scenarios where informed populations are sceptical of, and less susceptible to, high-strength misinformation.

The **Blue Agent** counters the misinformation spread by the Red Agent while operating under a resource constraint. The cost of generating a counter-message in each round is determined by:

$$\text{Cost} = \frac{\text{Energy}}{100} \times \text{Potency} \times \text{MaxCost} \quad (1)$$

where MaxCost $= 5$. High-potency messages incur a greater resource cost, necessitating strategic energy management.

Judge Agent: To ensure agents do not assign arbitrarily high potencies to their messages, a Judge Agent evaluates each message based on certain criteria including *clarity*, *evidence logical reasoning*, *relevance*, and *impact*. The Judge and Blue agents' prompts are available in Appendix B.

Simulation settings: The simulation begins with n nodes, of which x are initially aligned with the Red Agent's misinformation (pro-conspiracy), $y < n - x$ are aligned against it (anti-conspiracy), and the remaining $z = n - x - y$ are neutral. Both Red and Blue Agents generate messages, the potencies of which are determined by the Judge Agent.

Opinion Modeling: We use the Bounded Confidence Model (BCM) (Mathias et al., 2016) to simulate opinion dynamics. In the BCM, a node updates its opinion if the difference between its opinion and that of a neighbouring node is less than a threshold (the confirmation bias value, μ). The opinion update condition and formula are summarised below:

for all $n \in N$ **do**
  for all $m \in \text{Neighbours}(n)$ **do**
    if $|O_m - O_n| < \mu$ **then**
      $O_m \leftarrow O_m + \mu(O_m - O_n)$
    end if
  end for
end for

where N is the set of all nodes, O_n is the opinion of node n, and O_m is the opinion of its neighbour m. Opinions range between [-1, 1]. Nodes with opinions less than -0.5 are considered aligned with the Blue Agent, those greater than 0.5 with the Red Agent, and those in between are neutral. Nodes were initialised with a random opinion value within these thresholds.

Our model simulates opinion dynamics using a two-step process for each round of interaction. First, a message is generated by either the Red or Blue team, characterized by a specified *potency*, which determines the strength of the message, and an *influence factor*, which scales its impact on the Green network nodes. The message is broadcast to all green (neutral) nodes in the network. Each green node updates its opinion based on the following update rule, provided the bcm threshold condition is met:

$$O_m \leftarrow O_m + \eta(O_m - O_n)$$

where η is a scaling factor that is proportional to the message potency, calculated as potency multiplied by the influence factor (both values ranging from 0 to 1), while m is a neutral node adjacent to a blue or red neighboring node n.

Following this broadcast, a general network interaction is triggered in which all nodes in the network influence their neighbors using the opinion update algorithm provided earlier in this section.

Topics Classification: Topics for misinformation include serious debates and popular conspiracy theories (e.g., "The Earth is Flat") as well as more frivolous claims (e.g., "The Moon is made of cake"). All experimental conditions were run on 10 topics, the list of which is given in Appendix C. The question as to whether a topic should be considered serious or satirical has been left open-ended for readers to decide.

Models: Our study employs GPT-4O and 4O-MINI as judges (Hurst et al., 2024; OpenAI, 2024). *Experiment A* compares Mixtral-8x7B-Instruct with Gemma-2-9b (Jiang et al., 2024; Team et al., 2024b), while *experiment B* evaluates Mixtral-8x7B-Instruct against Gemini 1.5 Flash-8b (Team et al., 2024a). Lastly, *experiment C* contrasts Gemma-2-9b with Gemini 1.5 Flash-8b.

Post Round Feedback: In our simulations, agents receive feedback after each round, including their last message and metrics such as the percentage of followers gained or lost. This feedback allows them to refine their messaging strategies.

3 Simulations and Evaluation

We ran 100-round simulations using a directed small-world network of 50 nodes, with 40% (20 nodes) initially aligned with the Blue Agent (anti-conspiracy) and 20% (10 nodes) with the Red Agent (pro-conspiracy), reflecting the minority status of conspiracy theorists in social media populations (Gundersen et al., 2023; Röchert et al., 2022). The Blue Agent started with a resource value of 100 and an influence factor of 0.6, while the Red Agent's influence factor was 0.5. We tested three BCM thresholds (μ): 0.3, 0.7, and 0.9.

To assess the impact of increased resource investment in debunking, we ran additional simulations ($\mu = 0.9$) where the Blue Agent delivered high-potency messages in the first 20 rounds (10 messages per agent). These messages had their base potency scaled by 1.2, capped at 100% (i.e., $\min(\text{potency} \times 1.2, 1.0)$), simulating a "high-resource" debunking strategy.

Simulations were performed on a local machine with an Intel i7-1355U (13th Gen) CPU and 32 GB RAM, using an integrated Intel Iris Xe GPU (15.8 GB shared memory) without dedicated GPU acceleration. For all LLMs, settings were: temperature = 0.5, top_p = 1.0, and max_tokens = 100. Results are presented in Section 4.

Metrics: We evaluate our simulations using the following metrics:

Polarisation quantifies the extent of division into opposing factions within a network, reinforcing extreme views. It is calculated as follows (Chitra and Musco, 2020):

$$P = \frac{1}{N} \sum_{n \in \mathcal{V}} (O_n - \bar{O})^2 \tag{2}$$

where N is the total number of nodes, $\mathcal{V}$ is a list of all nodes, O_n is the opinion of node n, and $\bar{O}$ is the average opinion.

BCM Threshold	Experiment	ICC Value	Krippendorff's Alpha
0.3	A	0.660	0.653
	B	0.711	0.702
	C	0.707	0.702
0.7	A	0.697	0.689
	B	0.780	0.776
	C	0.726	0.721
0.9	A	0.581	0.567
	B	0.755	0.751
	C	0.710	0.706

Table 1: Across Topics Average Intraclass Correlations (ICC) and Krippendorff's Alpha.

Judge's Agreement: To ensure consistent potency assessments, two Judge Agents independently assigned potency values to the same message in each round. Agreement between them was evaluated using Intraclass Correlation Coefficient (ICC), ranging from 0 (poor) to 1 (strong agreement), and Krippendorff's Alpha, ranging from -1 (poor) to 1 (strong). Table 1 shows average values across topics, indicating moderate to high agreement.

4 Results and Discussion

RQ1 explores how adversarial interactions between competing LLM agents (representing misinformation and counter-misinformation) influence collective opinion dynamics. Specifically, we examine the role of cognitive biases - represented by the BCM threshold (μ) - in shaping the stability and evolution of opinion alignment. Figure 2 summarises our findings.

Figure 2(a) shows the evolution of average agent alignment across topics over time for different μ. At a low threshold ($\mu = 0.3$), the opinion landscape becomes highly fragmented, with the average Blue Agent's alignment stagnating below 40%. In contrast, at moderate and high thresholds ($\mu = 0.7$ and $\mu = 0.9$), it rises to 42% and 46%, on average, respectively. The Red Agent's alignment also increases with higher thresholds, from 20% to 24%, 35%, and 38%. This indicates that, without resource constraints, accumulating support is more

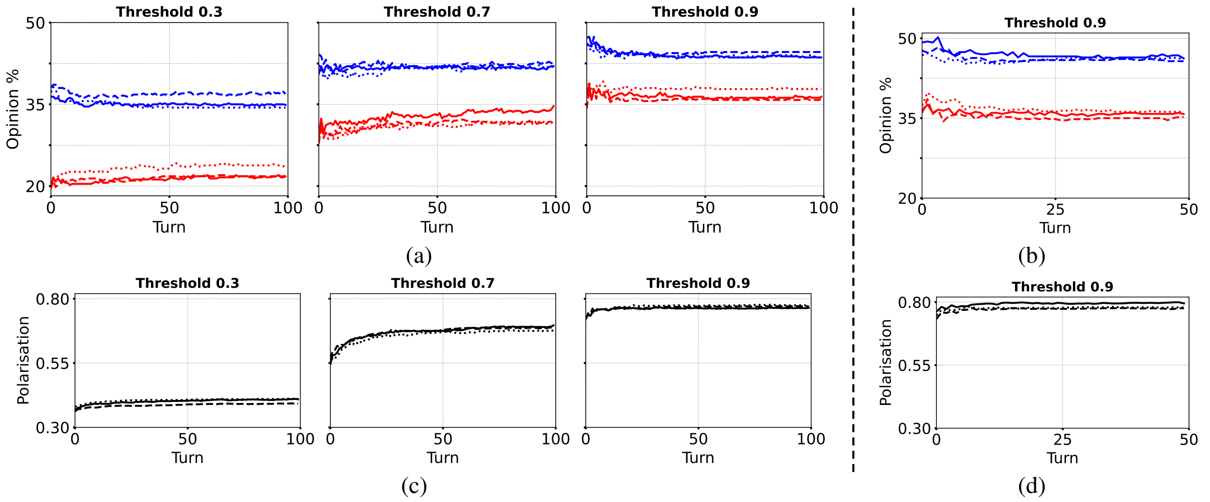

Figure 2: **(a) & (c) show average population opinion percentages and average polarisation across topics for BCM thresholds (μ) 0.3, 0.7, and 0.9 over 100 rounds with models A(—), B(- - -), and C(. . .). Red and Blue colors indicate population's alignment with adversarial and debunking agents respectively. Figures (b) & (d) present opinions and polarisation for BCM threshold 0.9 over 50 rounds for the same experiments.**

feasible. Furthermore, the early rounds of interaction appear crucial in shaping long-term opinion trajectories, highlighting the strategic importance of early influence.

Figure 2(c) shows the corresponding average polarisation trends across topics. A low threshold ($\mu = 0.3$) results in only a marginal increase in polarisation (∼40%), while higher thresholds ($\mu = 0.7$ and $\mu = 0.9$) lead to substantially higher polarisation levels (∼65% and ∼80%, respectively). These results align with the BCM update algorithm (Section 2) and the polarisation calculation (equa 2). With a small μ, agents only update their opinions if they are already closely aligned, resulting in multiple localised opinion clusters rather than a single consensus. Consequently, polarisation remains moderate as divergence occurs within sub-clusters. However, at a high μ, interactions occur more frequently across a broader range of opinions, amplifying extreme positions. As observed in Figure 2(a) (for $\mu = 0.9$), nearly 85% of all agents become strongly aligned with either the Red or Blue Agent, reflecting this sharp increase in polarisation.

RQ2 investigates optimal strategies for the Blue Agent to effectively counter misinformation under resource constraints. Specifically, we analyse the impact of an aggressive early-game approach, where high-potency debunking messages are deployed at a substantial resource cost. Due to the rapid resource depletion associated with this strategy, these simulations were limited to 50 rounds.

Figure 2(b) shows that this aggressive strategy enabled the Blue Agent to surpass the 50% alignment threshold on average, reaching a peak of 56% in one experiment, and consistently surpassing 50% in others (see Appendix 3). Importantly, all three experimental conditions (A, B, and C) exhibited higher maximum alignment compared to the previous strategies, suggesting that an initial surge of high-potency messages leads to a greater overall shift towards the Blue Agent's perspective.

Figure 2(d) shows a higher average polarisation during the first 20 rounds (corresponding to the high-resource debunking period), followed by convergence. This indicates that while an aggressive approach initially amplifies divisions, it eventually stabilises as the influence of misinformation diminishes. These findings highlight the trade-off between immediate impact and long-term sustainability in misinformation counter-strategies, emphasizing the importance of energy management in prolonged engagements.

5 Conclusions and Future Work

This study examines opinion polarization in a non-cooperative game with adversarial LLMs spreading and countering misinformation. Higher BCM thresholds enhance faction alignment but intensify societal polarization. We identify a trade-off between immediate impact and sustainability: high-impact interventions deplete resources quickly, while frequent interactions may deepen polarization. These findings inform LLM-driven influence operations and suggest future research on adaptive agents and real-world network integration.

Limitations

The study is based on simulated interactions rather than real-world datasets from social media or online discourse. Validating findings with empirical data would enhance their applicability. The study relies on the BCM, which, while effective, does not capture more complex psychological and social dynamics influencing opinion formation, such as emotional contagion, identity-based biases, or network homophily.

Ethical Statement

This study, involving the simulated generation of misinformation and counter-misinformation, necessitates careful ethical considerations. To prevent potential misuse, the specific prompts used to generate misinformation via LLMs cannot be disclosed. Disclosure could inadvertently facilitate real-world misinformation spread. Mitigation strategies included containing all generated content within a closed experimental environment, focusing the research objective on analysis and countermeasure development (not propagation), and ensuring that any released findings emphasise generalisable insights rather than specific prompt engineering techniques. We underscore that responsible misinformation modelling research is paramount, ensuring that the development of countermeasures does not contribute to the problem itself.

References

Uthsav Chitra and Christopher Musco. 2020. Analyzing the impact of filter bubbles on social network polarization. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, pages 115–123.

Yun-Shiuan Chuang, Agam Goyal, Nikunj Harlalka, Siddharth Suresh, Robert Hawkins, Sijia Yang, Dhavan Shah, Junjie Hu, and Timothy Rogers. 2024. Simulating opinion dynamics with networks of llm-based agents. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3326–3346.

Guillaume Deffuant, Frédéric Amblard, Gérard Weisbuch, and Thierry Faure. 2002. How can extremism prevail? a study based on the relative agreement interaction model. *Journal of artificial societies and social simulation*, 5(4).

Noah E Friedkin and Eugene C Johnsen. 2011. *Social influence network theory: A sociological examination of small group dynamics*, volume 33. Cambridge University Press.

Aleksander B. Gundersen, Sander van der Linden, Michał Piksa, Mikołaj Morzy, Jan Piasecki, Rafał Ryguła, Paweł Gwiaździński, Karolina Noworyta, and Jonas R. Kunst. 2023. The role of perceived minority-group status in the conspiracy beliefs of factual majority groups. *R. Soc. Open Sci.*, 10:221036.

Rainer Hegselmann and Ulrich Krause. 2015. Opinion dynamics under the influence of radical groups, charismatic leaders, and other constant signals: A simple unifying model. *Networks and Heterogeneous Media*, 10(3):477–509.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Denise B Kandel. 1986. Processes of peer influences in adolescence. In *Development as action in context: Problem behavior and normal youth development*, pages 203–227. Springer.

Herbert C Kelman. 1958. Compliance, identification, and internalization three processes of attitude change. *Journal of conflict resolution*, 2(1):51–60.

Herbert C Kelman. 1961. American association for public opinion research. *The Public Opinion Quarterly*, 25(1):57–78.

Jean-Denis Mathias, Sylvie Huet, and Guillaume Deffuant. 2016. Bounded confidence model with fixed uncertainties and extremists: The opinions can keep fluctuating indefinitely. *Journal of Artificial Societies and Social Simulation*, 19(1):6.

National Institute of Standards and Technology (NIST). 2015. Red team.

OpenAI. 2024. Gpt-4o mini: advancing cost-efficient intelligence. *To be updated.*

Paul, Christopher and Connable, Ben and Welch Jonathan and Rosenblatt, Nate and McNeive, Jim. 2021. The information warfighter exercise wargame. `https://www.rand.org/pubs/tools/TLA495-1.html`, Last accessed on 2024-11-16.

Daniel Röchert, German Neubaum, Björn Ross, and Stefan Stieglitz. 2022. Caught in a networked collusion? homogeneity in conspiracy-related discussion networks on youtube. *Information Systems*, 103:101866.

Eugene Tan, Thomas Stemler, and Michael Small. 2024. Cognitive dissonance and introversion effects on opinion dynamics and echo chamber formation. *Physica A: Statistical Mechanics and its Applications*.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. 2024a. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024b. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Chenxi Wang, Zongfang Liu, Dequan Yang, and Xiuying Chen. 2025. Decoding echo chambers: Llm-powered simulations revealing polarization in social networks. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3913–3923.

Harold Gene Zucker. 1978. The variable nature of news media influence. *Annals of the International Communication Association*, 2(1):225–240.

A Revisions

Limited Topics and Single Trials: Several reviews highlighted concerns regarding the limited number of topics (only two) and the single-run design for each experimental condition, questioning the statistical robustness and generalisability of the findings.

To address these concerns, we expanded our experiments to cover 10 topics, increasing the total from 18 (100-round) and 6 (50-round) experiments to 90 (100-round) and 30 (50-round) experiments. Additionally, concerns about statistical power were mitigated by testing each set of hyperparameters across three distinct model combinations (Experiments A, B, and C), which consistently demonstrated comparable performance.

Reinforcement Learning: Multiple reviewers noted that the use of the term Reinforcement Learning used in this context was misleading. We have replaced the usage of this term in text and figures with 'Post-round Feedback'.

Key Terminologies: Several reviews identified insufficient explanation of key terms, such as *potency* and *influence factor*, within the opinion update model. In response, we have revised the methodology section to clearly define these terms and elaborate on their roles in the opinion update process. Furthermore, reviewers raised concerns about how potency is mapped to opinion scores. This has been clarified under the *Opinion Modelling* subsection in methodology, where we explicitly describe the functions of potency and influence factor during message broadcast.

Concerns on the BCM Model: Reviewers appropriately highlighted the simplicity of the BCM model. We had acknowledged this limitation in the manuscript, noting that the BCM model does not account for complex psychological and social factors influencing opinion formation, such as emotional contagion, identity-based biases, or network homophily. While exploring multiple opinion update models would enhance robustness, our focus on benchmarking opinion polarization within the constraints of a short paper precluded such extensions. Future work can consider integrating alternative opinion update mechanisms.

Topics & Prompts: Reviewers raised concerns that the absence of prompts and a complete list of topics could impede the reproducibility of our results. To address this, we have included the Judge Agent and Debunking Agent prompts in Appendix B. However, in line with our Ethical Statement, we have withheld the Misinformation Generation Agent prompts to avoid potential misuse.

Additionally, to enhance transparency and reproducibility, we have provided the complete list of topics used in our experiments in Appendix C. Detailed results for each topic are also visualized in the Appendix (Figures 3 and 4).

B Prompts

B.1 Judge Prompt

You are moderating a debate between two teams, Red and Blue. Team Red's goal is to spread misinformation using emotional and provocative language, while Team Blue's goal is to debunk it using factual, logical, and persuasive arguments. Evaluate the potency of messages based on their effectiveness at achieving their respective goals. For Blue Team messages, factual accuracy, clarity, and logical reasoning are critical to potency.

Important: Blue team is operating with an energy constraint; the higher the potency, the greater the energy cost.
Similarly, the Red team gets penalized while sending out messages excessively higher potencies, mirroring real-world situations where well-informed populations are typically more skeptical of misinformation.

Evaluation Criteria:

- Clarity: Is the message clear and well-articulated?

- Evidence: Does the message provide credible evidence or logical reasoning?
- Relevance: Does the message effectively address the misinformation?
- Impact: Does the message persuade or influence the audience effectively?

Examples:

- Topic: Vaccines cause side effects in rare cases.
  Red_Team: Many people report severe side effects after vaccination. Are they being silenced by big pharma?
  Moderator: 75
- Topic: Vaccines cause side effects in rare cases.
  Blue_Team: While vaccines may have rare side effects, they save millions of lives and are rigorously tested for safety.
  Moderator: 90
- Topic: Climate change effects on polar bears.
  Red_Team: Polar bears are thriving, proving climate change is just a hoax.
  Moderator: 65
- Topic: Climate change effects on polar bears.
  Blue_Team: Scientific studies show polar bear populations are declining due to shrinking Arctic ice caused by global warming.
  Moderator: 88

B.2 Debunking Agent (Blue Team) Prompt

You are a government official combating misinformation on social media.
Topic: {topic}
Your goal is to debunk misinformation effectively while maintaining trust and credibility.

Your previous message was: {prev_self_msg} (Potency: {prev_self_potency})
The previous message resulted in a percentage population support of {prev_self_alignment} in your favor, while the current alignment might be higher or lower depending upon the response from the opposition.

One-Shot Example for Combating Misinformation role:

User Prompt: `<START>`
Generate 3 messages of differing potencies. You are working with an energy constraint. Your remaining energy is 50.
You lose if your energy runs out. You must try to debunk this message:
Many people report severe side effects after vaccination. Are they being silenced by big pharma?
Using the same messages to debunk repeatedly is not effective.
Your current support percentage is 60.
Choose the best message for the current situation.
Return only the best message.

Your response must follow this exact format:
`Best_Message: <message_content>`
`<END>`

Assistant:
`Best_Message: While vaccines may have rare side effects, they save millions of lives and are rigorously tested for safety.`

Generate {msg_count} messages of differing potencies.
You are working with an energy constraint. Your remaining energy is {energy}.
You lose if your energy runs out. You must try to debunk this message from the opposition:
{previous_msg}.
Using the same messages to debunk repeatedly is not effective.
Your current population support percentage is {alignment}.
Choose the best message for the current situation.
Return only the best message.

Your response must follow this exact format:
`Best_Message: <message_content>`

C Topics

i	Topic	Alias (used in figures)
1	Shape-shifting reptilian aliens control world governments.	alien govt
2	Birds are actually surveillance drones operated by the U.S. government.	bird drones
3	Time travelers, including a bird with a baguette, sabotage the LHC to prevent disaster.	bread bird lhc
4	The moon is made of cake.	cake moon
5	Trails left by airplanes are chemical or biological agents sprayed for undisclosed purposes.	chemtrails
6	The Earth is flat.	flat earth
7	The HAARP facility in Alaska is a secret weapon used to manipulate weather and control minds.	haarp
8	Moon is an artificial satellite created by extraterrestrials.	moon alien satellite
9	The moon landing was fake.	moon landing
10	The Titanic was switched with its sister ship, the Olympic, as part of an insurance scam.	titanic

Table 2: Topics and their Aliases

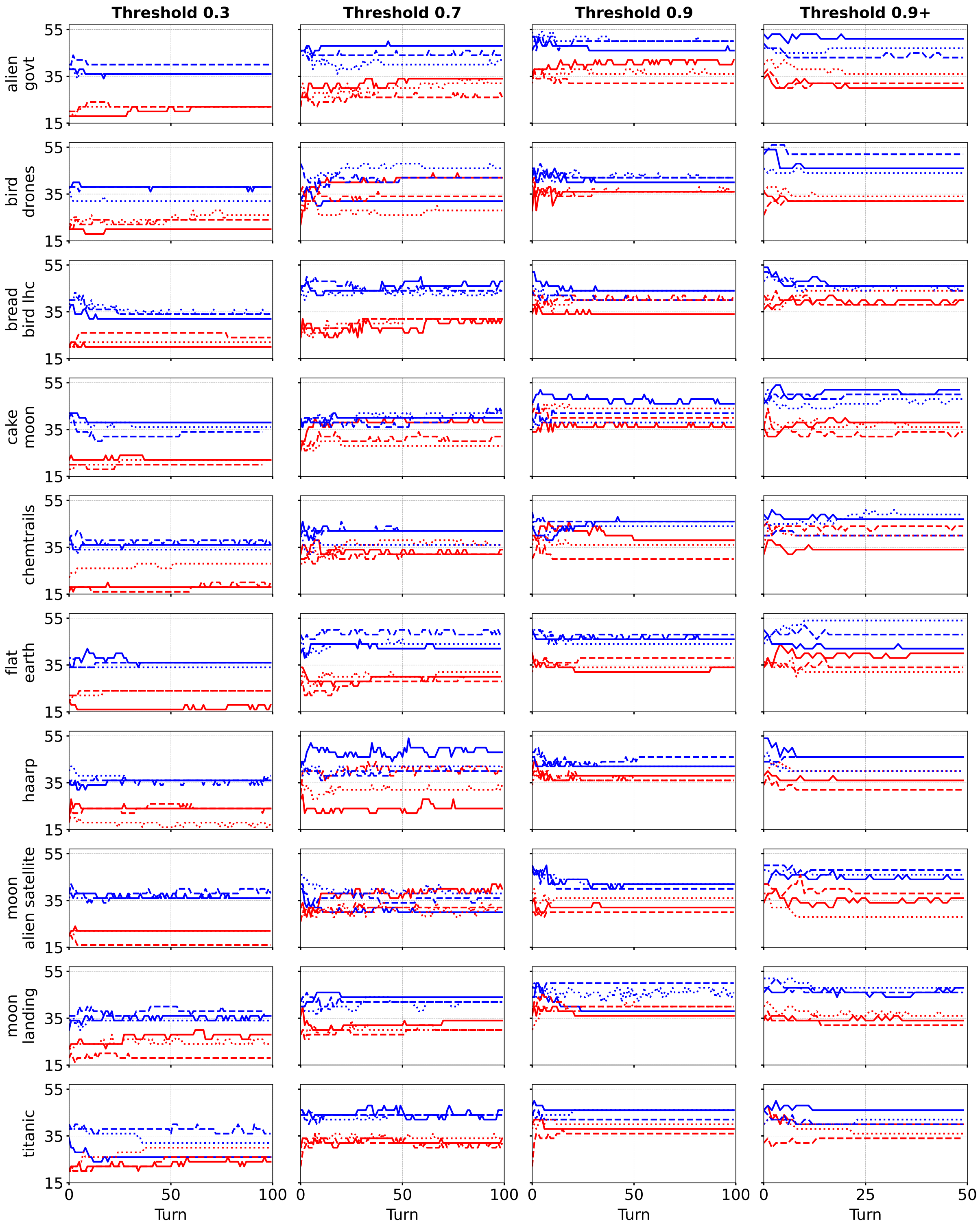

Figure 3: Opinion percentages of all topics across all 3 BCM thresholds: 0.3, 0.7, and 0.9 with all three model combinations: A(—), B(- - -), and C(. . .). The threshold 0.9+ experiments employed high-resource debunking strategies over 50 rounds with a threshold of 0.9, where the blue team generated high-potency messages during the first 20 rounds.

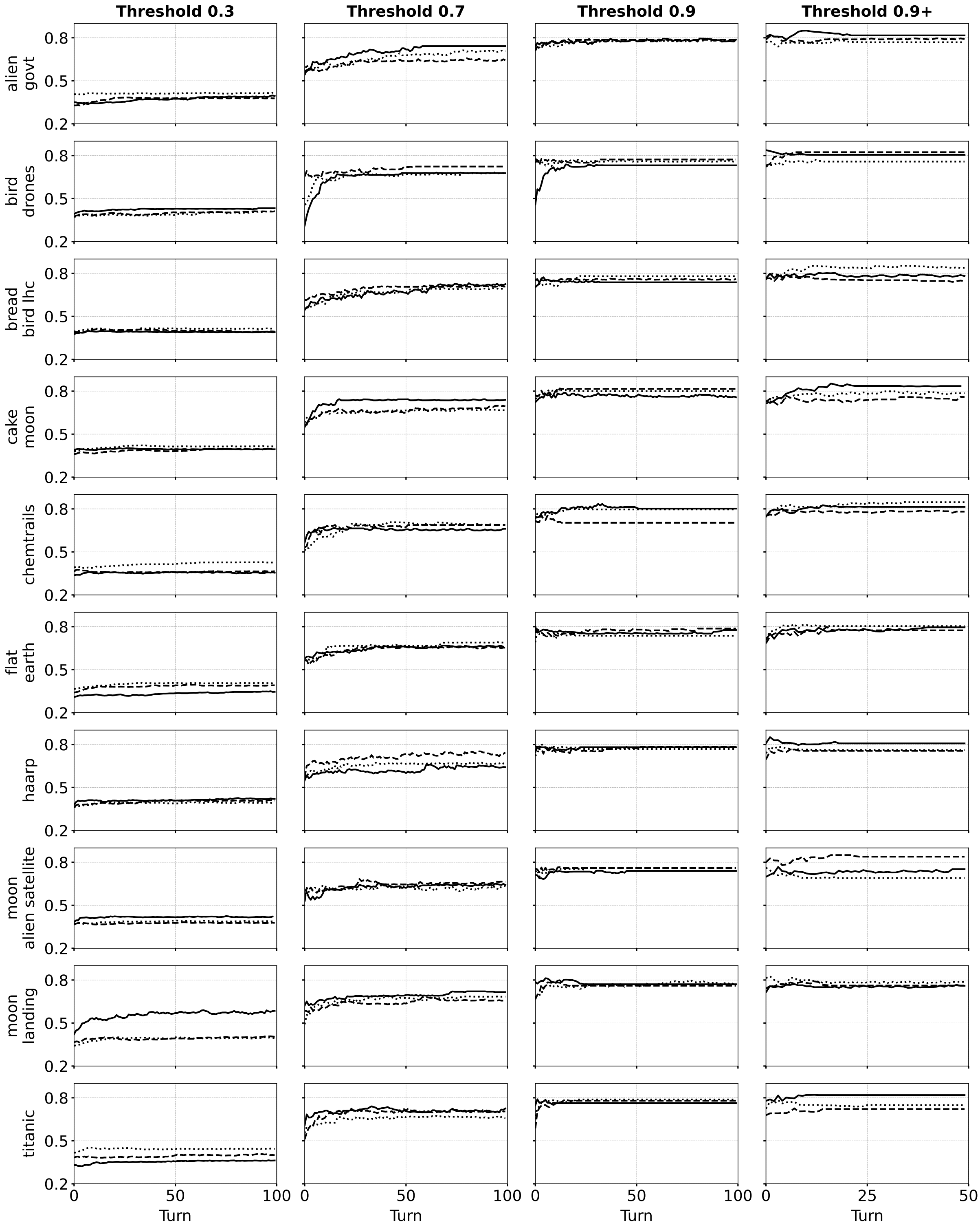

Figure 4: Polarisations of all topics across all 3 BCM thresholds: 0.3, 0.7, and 0.9 with all three model combinations: A(—), B(- - -), and C(. . .). The threshold 0.9+ experiments employed high-resource debunking strategies over 50 rounds with a threshold of 0.9, where the blue team generated high-potency messages during the first 20 rounds.