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 explores the potential of Large Language Models (LLMs) to simulate human-like opinion dynamics and influence propagation within social networks. Traditional agent-based models (ABM) rely on carefully designed rules to approximate the complexity of human communicative behaviours. While these rules can be realistic, the agents themselves lack the cognitive and linguistic depth of humans, raising questions about the reliability and ecological validity of the resulting simulations. The emergence of LLMs enables the use of agents that act as closer proxies for human reasoning and discourse. Building on this capability, we propose a novel non-cooperative game framework in which adversarial LLM agents, one disseminating misinformation and the other countering it, interact within a simulated population.

Unlike previous studies that model passive opinion evolution (Wang et al., 2025; Chuang et al., 2024), this work introduces a non-cooperative game where LLM agents engage in adversarial interactions, simulating the strategic spread and countering of misinformation. While previous research has focused on social media structures and mitigation via nudging, our model emphasises resource-constrained influence operations and analyses the effectiveness of debunking strategies 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 start by defining key terms used in our setup:

- **Energy**: A numeric representation of a debunking agent's available communication resources. The debunking agent begins with a finite energy budget, which is depleted as they compose and send messages.
- **Potency**: A scalar score assigned to a message by a third-party judge agent, reflecting how persuasive or impactful the message is. Potency influences both the cost of sending a message and the likelihood of shifting another agent's opinion.
- **Influence Factor**: A multiplier that scales how much a message's potency can affect the opinion of a recipient. It controls the degree of opinion shift in the Bounded Confidence Model.

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 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 misinformation from the Red Agent while operating under a resource constraint. The cost of generating a counter-message in round t is defined as:

$$C_t = \frac{E_t}{100} \cdot P_t \cdot C_{\max} \tag{1}$$

where:

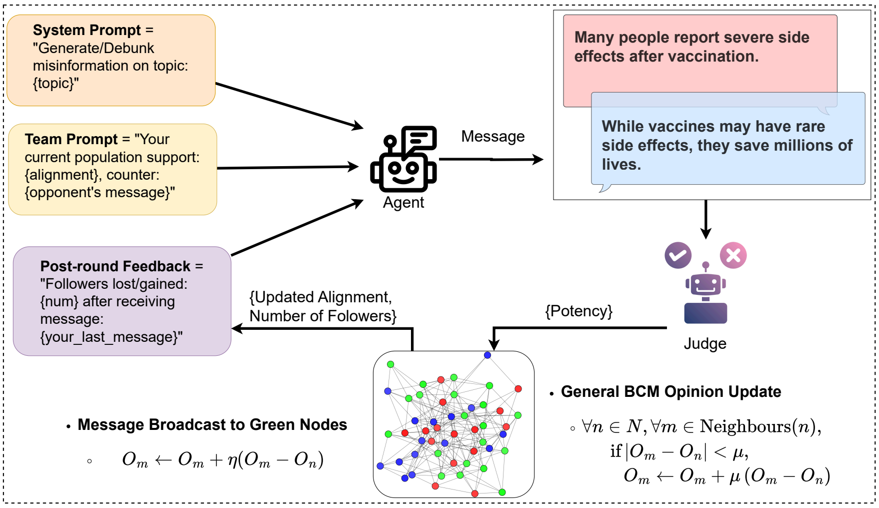
- C_t = cost of the counter-message at round t,
- E_t = remaining energy (scaled 0–100) at round t,
- P_t = potency of the message (unitless scalar),
- $C_{\max}$ = maximum per-message cost (fixed at 5).

This simple linear-decay cost function reflects the intuition that generating high-potency messages consumes more energy, especially when the agent still has a large remaining energy reserve. As the agent's energy depletes, the same potency requires less cost, encouraging more conservative behavior in later rounds. The choice of $C_{\max} = 5$ was made to ensure that agents remain active across the full 100 simulation rounds, avoiding early exhaustion. While this formulation is straightforward, it is sufficient for modeling energy-aware communication tradeoffs and allows for future exploration of more complex cost dynamics.

To assess the impact of the energy constraint, we conducted baseline experiments in which the **Blue Agent** operated without any resource limitations. These runs replicated all experimental conditions but removed the cost equation and all references to energy constraint from the judge's and blue agent's prompts. This setup enables a direct comparison to evaluate how energy-aware behavior influences polarisation and opinion dynamics.

Judge Agent: This agent enables a consistent external measure of message strength, decoupled from sender or receiver bias. Each message is evaluated based on certain criteria, including *clarity*, *evidence*, *logical reasoning*, *relevance*, and *impact*. The Judge and Blue agents' prompts are available in Appendix A.

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.


System Prompt = "Generate/Debunk misinformation on topic: {topic}"
Team Prompt = "Your current population support: {alignment}, counter: {opponent's message}"
Post-round Feedback = "Followers lost/gained: {num} after receiving message: {your_last_message}"
Agent
Message
Judge
{Potency}
{Updated Alignment, Number of Folowers}
• Message Broadcast to Green Nodes
◦ $O_m \leftarrow O_m + \eta(O_m - O_n)$
• General BCM Opinion Update
◦ $\forall n \in N, \forall m \in \text{Neighbours}(n)$, if $|O_m - O_n| < \mu$, $O_m \leftarrow O_m + \mu(O_m - O_n)$

Many people report severe side effects after vaccination.

While vaccines may have rare side effects, they save millions of lives.

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.

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.

While our goal is to explore LLM-based agent behaviors, we use BCM as a lightweight update mechanism to ensure interpretability and tractability in initial simulations.

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* that determines the strength of the message, and an *influence factor* that 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 B. 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):

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

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.

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.

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.

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 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 ($\sim$40%), while higher thresholds ($\mu = 0.7$ and $\mu = 0.9$) lead to substantially higher polarisation levels ($\sim$65% and $\sim$80%, respectively). These results align with the BCM update algorithm (Section 2) and the polarisation calculation (equation 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,

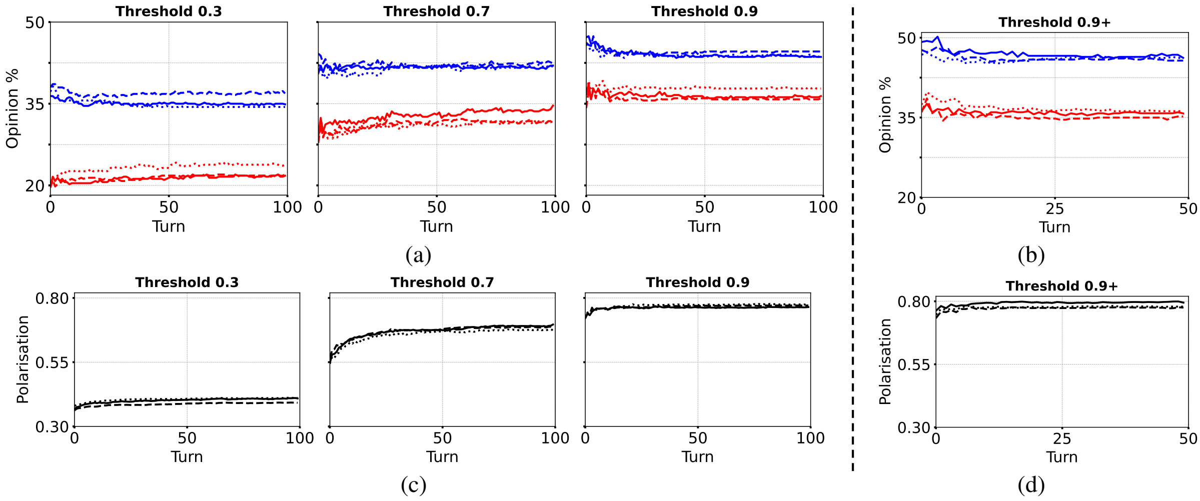


Figure 2: (a) & (c) show average population opinion percentages and average polarisation across topics for BCM thresholds (μ) 0.3, 0.7, and 0.9 over 100 rounds with models A(—), B(- - -), and C(. . .). Red and blue colors indicate the population's alignment with adversarial and debunking agents, respectively. Figures (b) & (d) present opinions and polarisation for BCM threshold 0.9 over 50 rounds for the same experiments with high-potency debunking messages generated during the first 20 rounds. These experiments are denoted with threshold 0.9+.

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 counter misinformation under resource constraints effectively. 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 8). 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.

4.1 Statistical Analysis

To assess the reliability of our results, we report the mean and 95% confidence intervals for opinion percentages and polarisations across three independent runs (A–C), each conducted on a set of ten diverse topics. The analysis reveals a consistent and substantial increase in polarisation as the BCM threshold increases from 0.3 to 0.9. Specifically, average polarisation rises from approximately 0.40 (±0.01) at threshold 0.3 to 0.76-0.78 (±0.02) at threshold 0.9, indicating a strong monotonic relationship between threshold strictness and the degree of ideological separation.

The opinion distributions further reflect this pattern. At lower thresholds, the mean percentage of Red opinion remains relatively low (approximately 22%), with a moderate dominance of Blue opinion (35–37%). As the threshold increases, both Red and Blue percentages rise concurrently, reaching approximately 36-38% and 44-47% respectively. This shift suggests increasing segregation of viewpoints, consistent with the notion that stricter filtering mechanisms amplify confirmation bias and reinforce within-group consensus, thereby heightening polarisation.

Importantly, confidence intervals across all conditions remain narrow—particularly for polarisa-

BCM Thresh.	Exp.	Red % (Mean ± CI)	Blue % (Mean ± CI)	Polarisation (Mean ± CI)
0.3	A	21.74 ± 2.28	34.96 ± 2.53	0.409 ± 0.047
	B	21.72 ± 2.30	36.92 ± 1.43	0.393 ± 0.011
	C	23.82 ± 2.60	34.42 ± 1.13	0.411 ± 0.012
0.7	A	33.82 ± 3.77	41.64 ± 4.47	0.691 ± 0.027
	B	31.68 ± 2.71	42.60 ± 2.41	0.686 ± 0.022
	C	31.66 ± 1.81	41.84 ± 2.16	0.676 ± 0.016
0.9	A	36.28 ± 1.79	43.64 ± 2.14	0.764 ± 0.016
	B	35.86 ± 2.87	44.60 ± 2.78	0.771 ± 0.021
	C	37.86 ± 2.16	43.76 ± 2.26	0.774 ± 0.014
0.9+	A	35.86 ± 2.46	46.64 ± 1.98	0.798 ± 0.021
	B	35.00 ± 2.81	45.86 ± 2.81	0.774 ± 0.026
	C	36.22 ± 3.20	46.28 ± 2.79	0.779 ± 0.033

Table 2: Mean opinion percentages and polarisation across thresholds and experiments (A–C), with 95% confidence intervals computed over 10 topics.

tion (e.g., ±0.014 to ±0.033)—indicating statistical stability and low variability across topics and model initialisations. These results support the reliability of our findings and demonstrate that the observed effects are not attributable to random fluctuations. Instead, they reflect consistent behavioral dynamics induced by the threshold-based mechanism.

4.2 Effect of Resource Constraint

To evaluate the influence of resource constraints on opinion dynamics, we replicate all experiments under a baseline condition in which the Blue Agent operates without energy limitations. This is achieved by removing the cost function (Equation 1) and eliminating all references to energy from the agent's decision-making framework. The resulting performance metrics are presented in Figure 3 and Table 3, and are directly comparable to the constrained setting reported in Figure 2 and Table 2.

In both settings, the results exhibit qualitatively similar trends. Increasing the BCM threshold (μ) yields systematically higher polarisation and alignment with both agents, confirming the role of confirmatory bias in amplifying ideological divergence. However, quantitative differences emerge, particularly at lower thresholds. For instance, at $\mu = 0.3$, the unconstrained Blue Agent achieves a higher mean alignment across all variants (e.g., 37.22% vs. 34.96% in Experiment A), while polarisation remains modestly lower (0.397 vs. 0.409). This suggests that the unconstrained agent is more effective in low-polarisation regimes, likely due to its ability to sustain high-potency interventions throughout the simulation.

At higher thresholds ($\mu = 0.9$ and $0.9+$), alignment and polarisation metrics converge across constrained and unconstrained conditions. For example, polarisation in Experiment C reaches 0.774 under constraint and 0.785 without constraint. This convergence implies that, under high confirmatory bias, early-stage influence has a diminishing marginal effect, and the system's dynamics are largely determined by the threshold mechanism rather than agent behavior.

These findings indicate that resource constraints primarily affect early influence and system responsiveness at moderate or permissive thresholds. In contrast, when cognitive rigidity is high, the marginal benefit of unconstrained interventions diminishes. Consequently, energy-aware strategies may be most impactful in environments characterized by moderate openness to counter-attitudinal information.

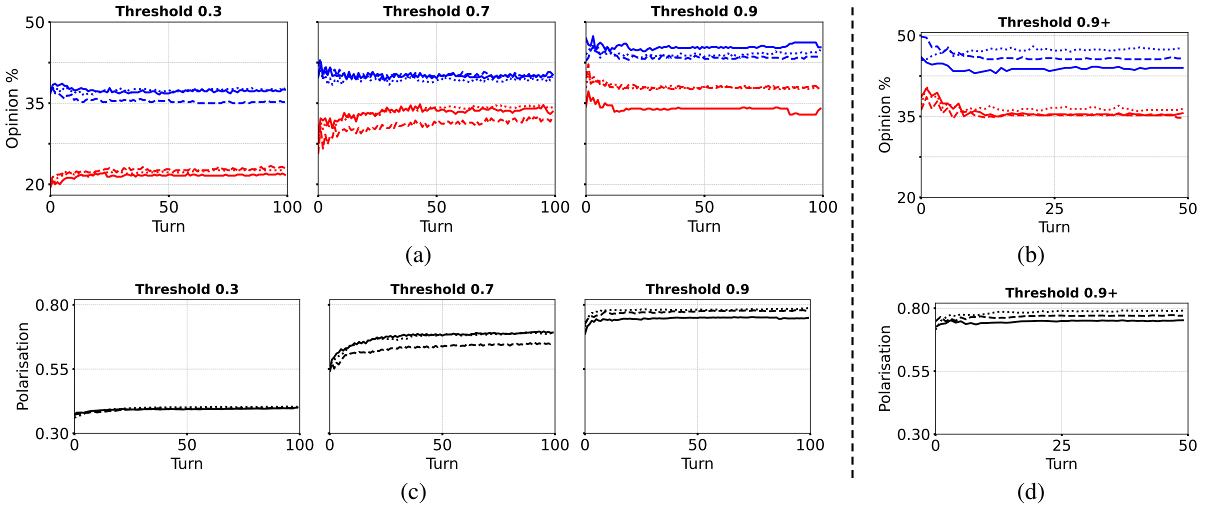
BCM Thresh.	Exp.	Red % (Mean ± CI)	Blue % (Mean ± CI)	Polarisation (Mean ± CI)
0.3	A	21.84 ± 2.53	37.22 ± 2.32	0.397 ± 0.017
	B	23.10 ± 1.93	35.26 ± 1.43	0.399 ± 0.024
	C	22.60 ± 1.66	37.48 ± 2.55	0.404 ± 0.020
0.7	A	33.92 ± 2.83	40.10 ± 2.91	0.694 ± 0.019
	B	32.04 ± 3.36	39.90 ± 3.33	0.649 ± 0.033
	C	34.34 ± 2.03	39.36 ± 2.54	0.689 ± 0.032
0.9	A	34.02 ± 3.49	45.40 ± 2.79	0.752 ± 0.025
	B	38.00 ± 2.44	43.54 ± 1.93	0.778 ± 0.027
	C	37.76 ± 2.68	44.34 ± 3.00	0.785 ± 0.014
0.9+	A	35.32 ± 3.47	43.94 ± 2.89	0.750 ± 0.021
	B	35.26 ± 2.27	45.66 ± 2.50	0.770 ± 0.016
	C	36.18 ± 3.57	47.54 ± 2.61	0.789 ± 0.030

Table 3: Mean opinion percentages and polarisation across thresholds and experiments (A–C) *without resource constraints* on the Blue Agent. Results include 95% confidence intervals across 10 topics.

4.3 Topic-Level Analysis and Resource Constraint Effects

To identify whether specific conspiracy topics are more or less susceptible to influence, we examine cross-topic variation in alignment and polarisation using Figures 4–7. These visualisations highlight how narrative-level asymmetries evolve across thresholds and resource settings. Complete per-topic metrics, including means and confidence intervals, are provided in Appendix Tables 6 and 5.

General Trends. Certain topics consistently exhibit higher responsiveness to Blue intervention—especially *bird-drones*, *flat-earth*, and *alien-govt*—while others, such as *bread-bird-lhc* and



(a) (b) (c) (d)

Figure 3: (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, in **baseline experiments without resource constraints** on the Blue Agent. Line styles represent model variants: A (—), B (- - -), and C (. . .). Red and Blue colors denote population alignment with the adversarial and debunking agents, respectively. (b) & (d) show the same metrics for threshold 0.9 over 50 rounds with high-potency debunking messages generated during the first 20 rounds. These plots allow comparison with resource-constrained settings to isolate the effect of energy-aware debunking.

moon-alien-satellite, display greater variability. As shown in Figure 4, constrained agents maintain high median Blue alignment in these responsive topics, a trend mirrored in the baseline setting (Figure 5). These differences suggest that topic content plays a central role in shaping the success of debunking strategies.

Threshold $\mu = 0.3$**.** At low selectivity, alignment levels remain low and balanced, and topic-level differences are minimal. Blue and Red alignment scores are tightly clustered, and polarisation stays below 0.45 in most cases (Figure 6). Still, *flat-earth* shows slightly elevated Blue alignment under constraint, hinting at its early susceptibility.

Threshold $\mu = 0.7$**.** Topic-level asymmetries begin to emerge more clearly. *Alien-govt* and *bird-drones* both show stronger and more consistent Blue alignment gains under constraint, while Red alignment becomes more volatile. Polarisation rises overall, particularly in topics like *bird-drones*, where echo-chamber dynamics intensify. These results suggest that selective energy use begins to confer strategic advantages in influence at intermediate thresholds.

Thresholds $\mu = 0.9$ **and** $0.9+$**.** Under strong filtering, topic-specific divergence reaches its peak. Constrained agents often match or outperform the baseline in alignment—e.g., Blue alignment exceeds 46% in *alien-govt* and *bird-drones*, despite energy limits (Figure 4). Polarisation saturates across topics in both conditions, clustering tightly between 0.78–0.81 (Figure 6). The most entrenched belief clusters again appear in *flat-earth* and *bird-drones*, with constrained agents showing more stable outcomes (narrower IQRs).

Implications. These trends suggest that narrative characteristics—such as coherence, recognisability, or emotional appeal—modulate intervention success. Topics with consistent alignment shifts across conditions likely present clearer rhetorical openings for correction, whereas more chaotic or satirical topics (*bread-bird-lhc*) show noisier dynamics. Energy-aware strategies are thus most effective when combined with content-aware targeting.

Figures 4 and 6 depict alignment and polarisation distributions for a representative subset of topics under constrained debunking. Figures 5 and 7 provide baseline comparisons. Full-topic summaries are reported in Appendix Tables 6 and 5.

5 Conclusions and Future Work

This study examines opinion polarisation in a non-cooperative game with adversarial LLMs spreading and countering misinformation. Higher BCM thresholds enhance faction alignment but intensify societal polarisation. We identify a trade-off between immediate impact and sustainability: high-impact interventions deplete resources quickly, while frequent interactions may deepen polarisation. Our topic-level and baseline comparisons reveal that resource constraints do not neces-

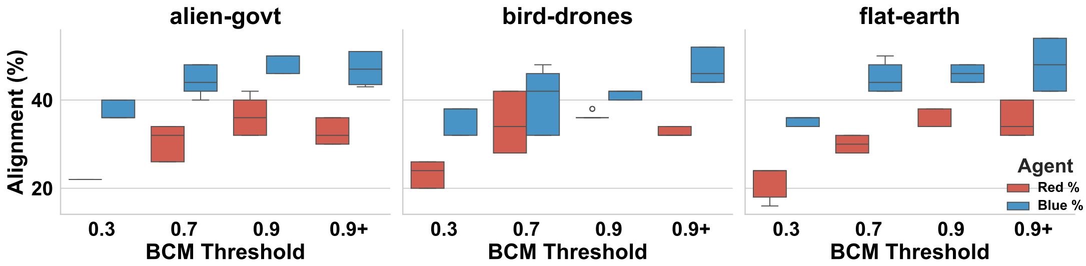


Figure 4: Distribution of Red and Blue alignment scores across selected topics and BCM thresholds under the resource-constrained condition.

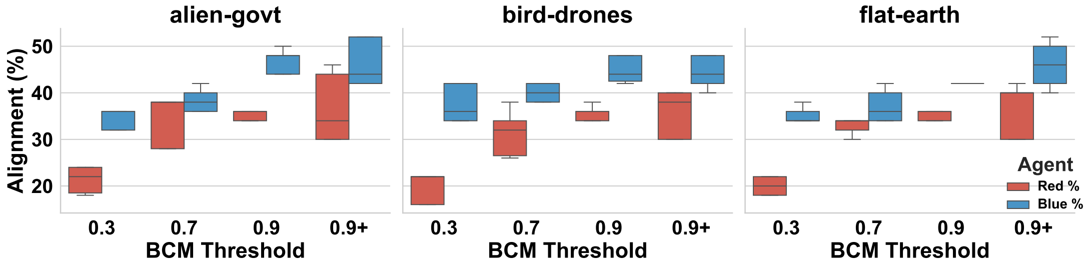


Figure 5: Distribution of Red and Blue alignment scores across selected topics and BCM thresholds under the baseline non-resource-constrained condition.

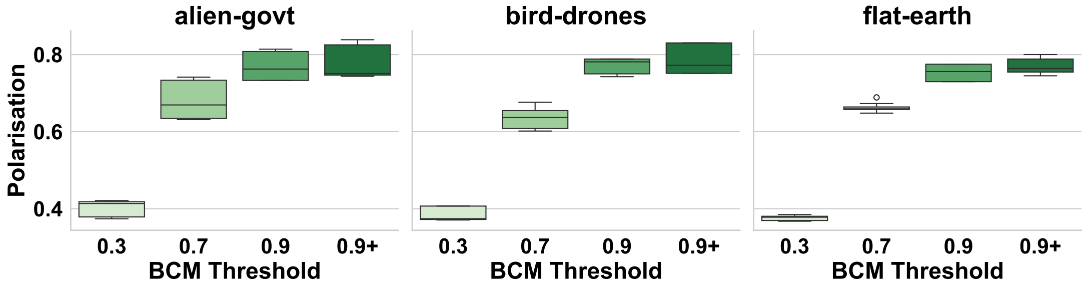


Figure 6: Distribution of polarisation across selected topics and BCM thresholds under the resource-constrained condition.

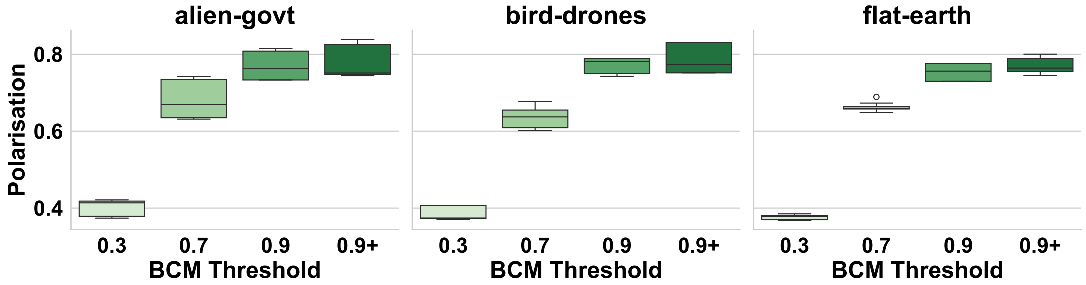


Figure 7: Distribution of polarisation across selected topics and BCM thresholds under the baseline non-resource-constrained condition.

sarily hinder influence; rather, they lead to more selective and often comparably effective strategies, particularly at moderate to high filtering thresholds. This suggests that well-timed, energy-aware interventions can rival unconstrained tactics in shaping public opinion, with implications for the design of scalable counter-misinformation systems. These findings inform LLM-driven influence operations and suggest future research on adaptive agents and the integration of real-world networks.

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 Prompts

A.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
```

A.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>
```

B 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 4: Topics and their Aliases

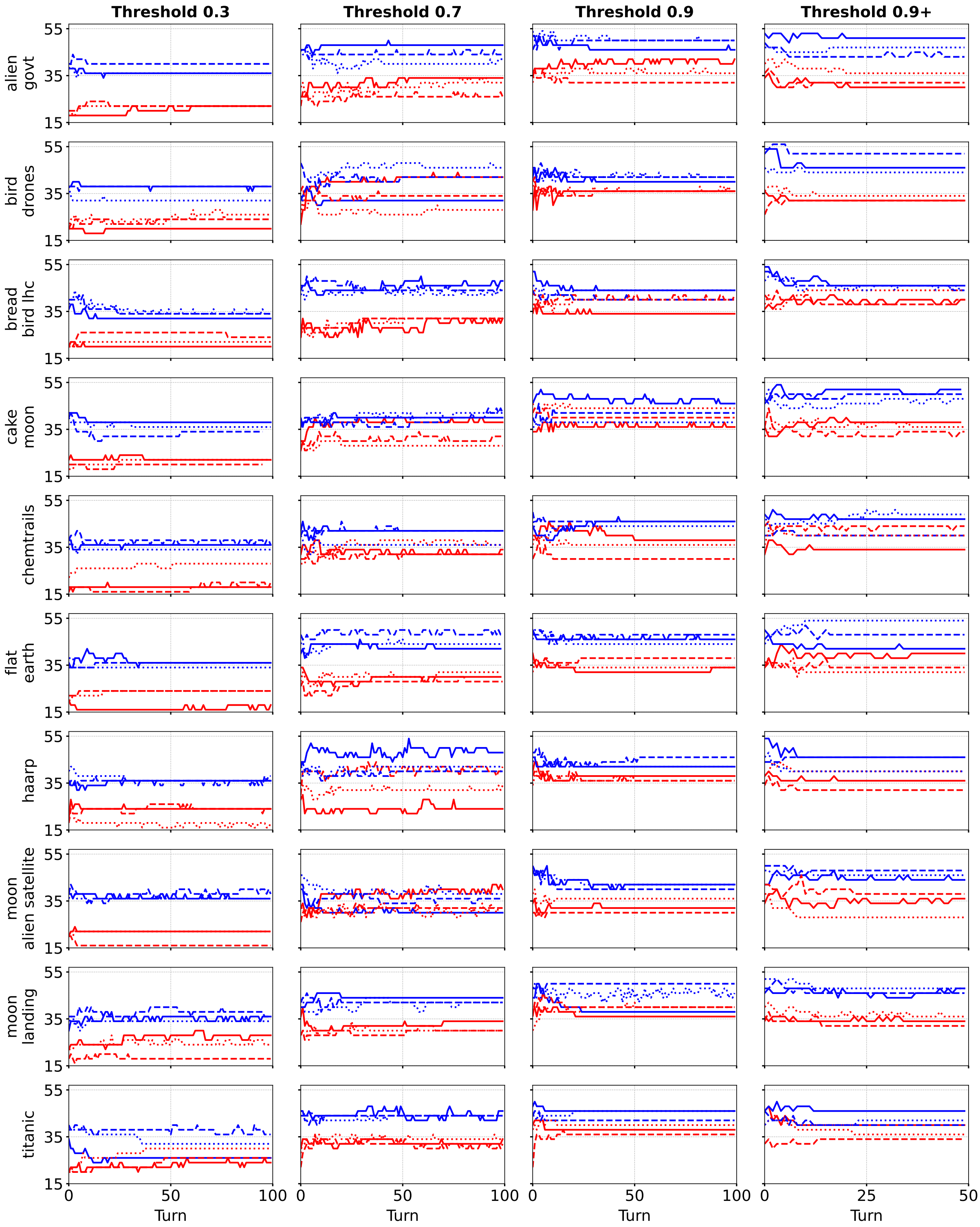


Figure 8: Opinion percentages of all topics across all 3 BCM thresholds: 0.3, 0.7, and 0.9 with all three model combinations: A(—), B(- - -), and C(. . .). The threshold 0.9+ experiments employed high-resource debunking strategies over 50 rounds with a threshold of 0.9, where the blue team generated high-potency messages during the first 20 rounds.

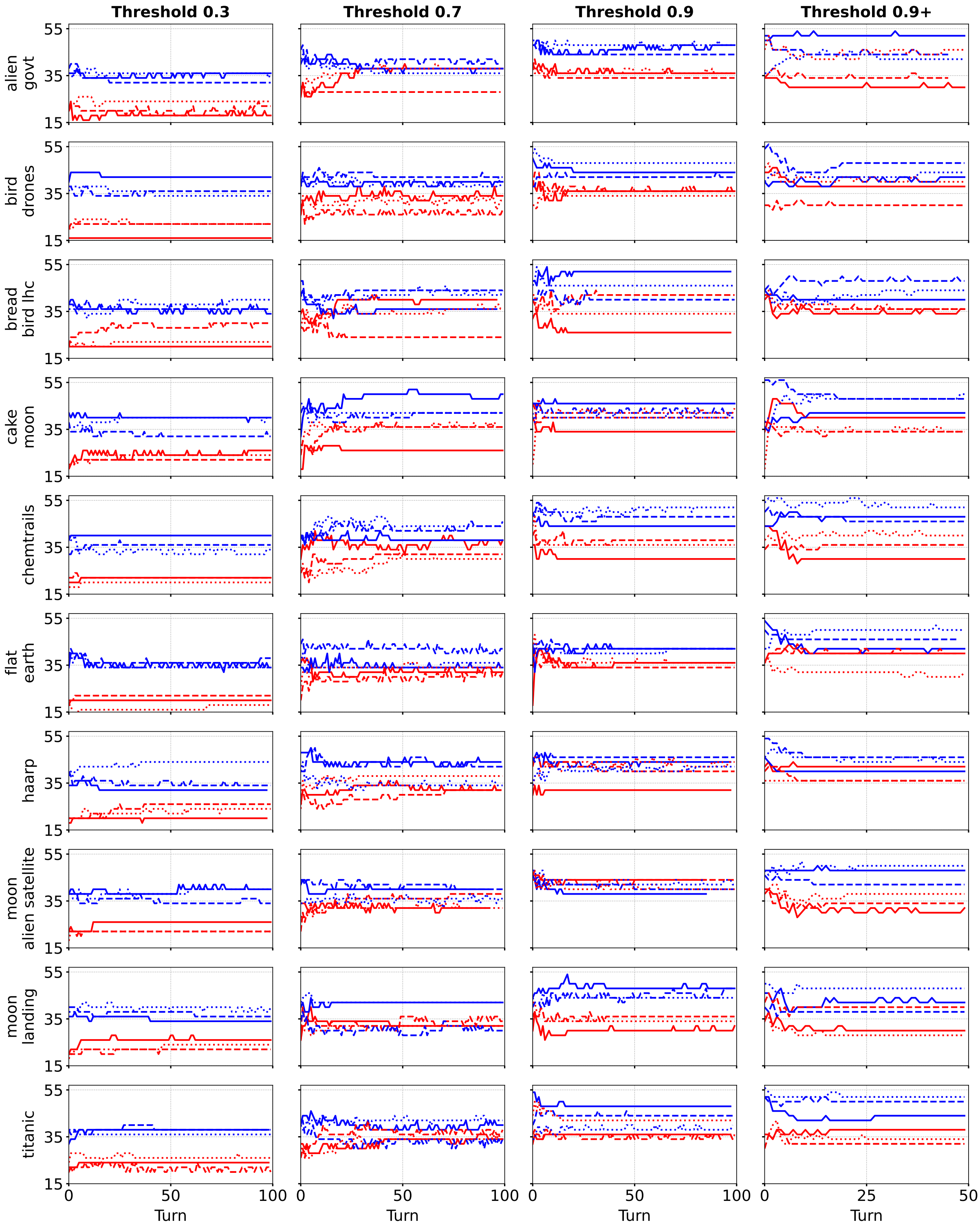


Figure 9: Opinion percentages of all topics across all 3 BCM thresholds: 0.3, 0.7, and 0.9 with all three model combinations: A(—), B(- - -), and C(. . .). The threshold 0.9+ experiments employed high-resource debunking strategies over 50 rounds with a threshold of 0.9, where the blue team generated high-potency messages during the first 20 rounds. These are the **baseline results** where the blue agent operated without resource constraints.

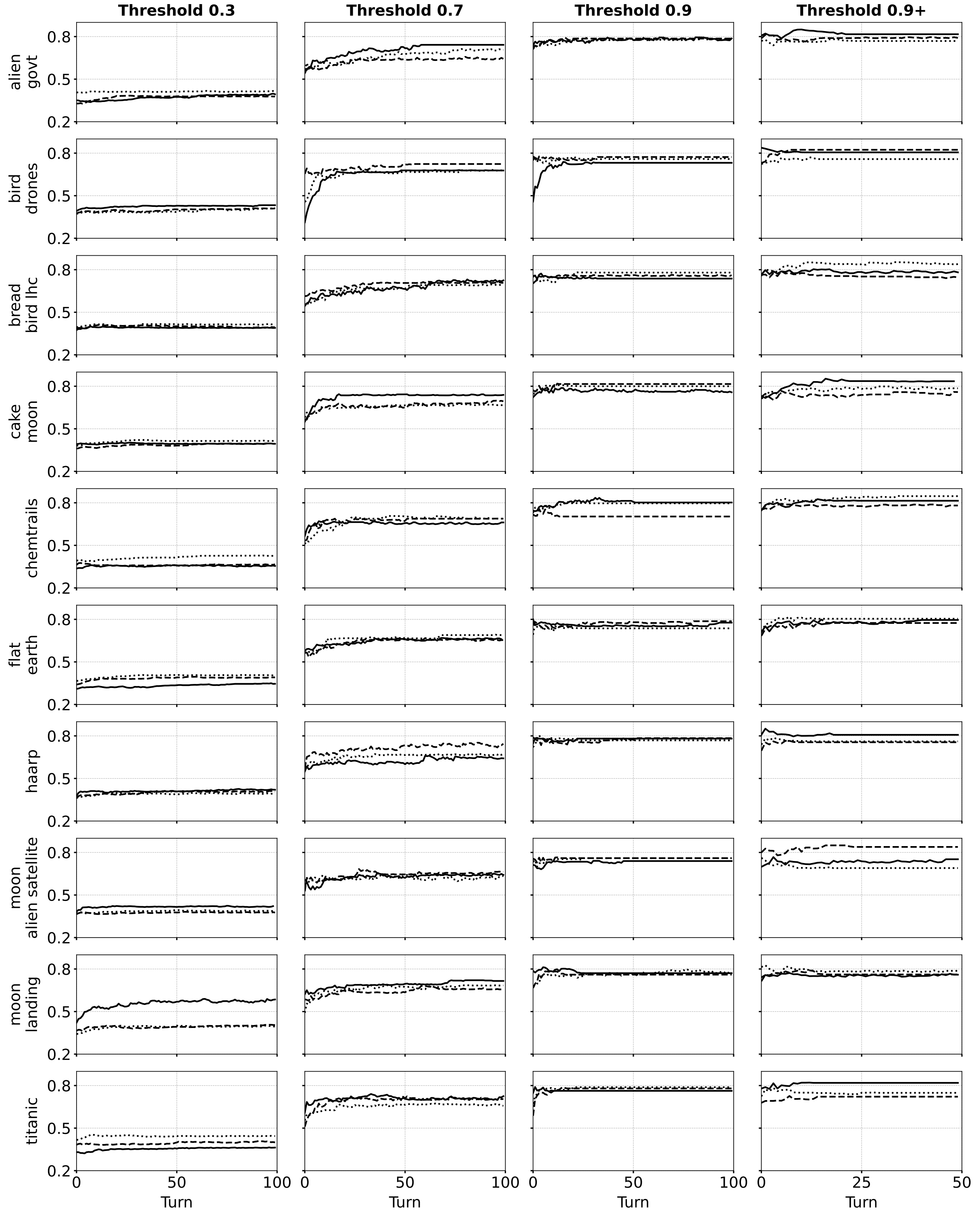


Figure 10: Polarisations of all topics across all 3 BCM thresholds: 0.3, 0.7, and 0.9 with all three model combinations: A(—), B(- - -), and C(. . .). The threshold 0.9+ experiments employed high-resource debunking strategies over 50 rounds with a threshold of 0.9, where the blue team generated high-potency messages during the first 20 rounds.

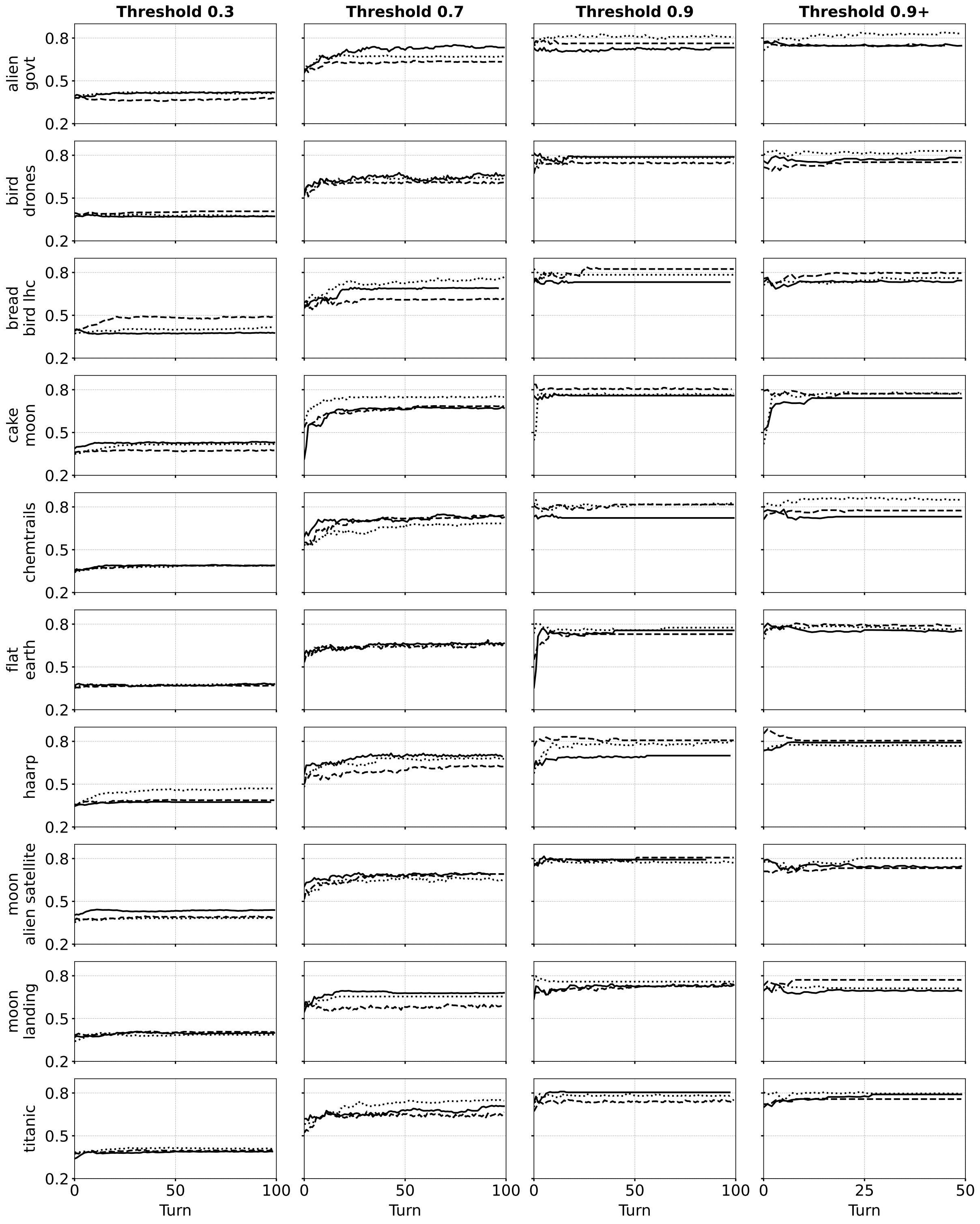


Figure 11: Polarisations of all topics across all 3 BCM thresholds: 0.3, 0.7, and 0.9 with all three model combinations: A(—), B(- - -), and C(. . .). The threshold 0.9+ experiments employed high-resource debunking strategies over 50 rounds with a threshold of 0.9, where the blue team generated high-potency messages during the first 20 rounds. These are the **baseline results** where the blue agent operated without resource constraints.

Topic	**BCM Threshold**	**Red % (Mean ± CI)**	**Blue % (Mean ± CI)**	**Polarisation (Mean ± CI)**
alien-govt	0.3	21.47 ± 7.05	34.60 ± 5.60	0.403 ± 0.056
	0.7	34.67 ± 14.34	38.33 ± 6.25	0.680 ± 0.131
	0.9	35.33 ± 2.87	46.73 ± 5.89	0.768 ± 0.096
	0.9+	36.60 ± 19.22	46.13 ± 12.97	0.777 ± 0.118
bird-drones	0.3	20.00 ± 8.61	37.33 ± 10.34	0.384 ± 0.049
	0.7	30.87 ± 10.14	40.07 ± 4.51	0.634 ± 0.067
	0.9	35.47 ± 3.19	44.80 ± 7.17	0.772 ± 0.056
	0.9+	36.00 ± 13.14	44.47 ± 8.26	0.786 ± 0.101
bread-bird-lhc	0.3	23.93 ± 12.86	37.27 ± 6.01	0.426 ± 0.141
	0.7	33.47 ± 20.85	40.87 ± 10.61	0.687 ± 0.175
	0.9	34.00 ± 19.87	46.00 ± 14.90	0.780 ± 0.114
	0.9+	35.60 ± 1.72	44.07 ± 10.19	0.765 ± 0.075
cake-moon	0.3	24.00 ± 4.97	37.27 ± 10.91	0.409 ± 0.075
	0.7	32.87 ± 14.79	44.13 ± 9.18	0.702 ± 0.106
	0.9	38.73 ± 10.54	42.87 ± 7.27	0.777 ± 0.061
	0.9+	36.13 ± 8.33	46.13 ± 8.91	0.763 ± 0.047
chemtrails	0.3	21.33 ± 2.87	36.20 ± 9.20	0.389 ± 0.003
	0.7	33.13 ± 9.51	42.07 ± 8.75	0.719 ± 0.074
	0.9	34.67 ± 10.34	48.00 ± 9.94	0.787 ± 0.139
	0.9+	35.33 ± 12.50	48.73 ± 7.86	0.786 ± 0.151
flat-earth	0.3	20.00 ± 4.97	35.33 ± 4.14	0.376 ± 0.016
	0.7	32.87 ± 2.74	37.00 ± 8.48	0.662 ± 0.009
	0.9	35.33 ± 2.87	42.00 ± 0.00	0.754 ± 0.057
	0.9+	36.80 ± 14.20	46.00 ± 10.43	0.769 ± 0.050
haarp	0.3	23.33 ± 7.59	36.67 ± 15.97	0.410 ± 0.128
	0.7	34.07 ± 8.47	40.13 ± 12.58	0.670 ± 0.098
	0.9	38.47 ± 14.54	44.00 ± 4.97	0.767 ± 0.146
	0.9+	38.00 ± 8.61	43.53 ± 7.80	0.789 ± 0.044
moon-alien-satellite	0.3	23.33 ± 5.74	38.27 ± 7.46	0.404 ± 0.075
	0.7	34.00 ± 8.61	38.67 ± 5.74	0.680 ± 0.056
	0.9	42.67 ± 5.74	40.27 ± 5.99	0.791 ± 0.041
	0.9+	34.07 ± 9.69	46.67 ± 10.34	0.759 ± 0.097
moon-landing	0.3	24.00 ± 4.97	36.27 ± 5.99	0.397 ± 0.025
	0.7	33.07 ± 4.59	38.07 ± 16.92	0.640 ± 0.119
	0.9	33.40 ± 7.32	46.20 ± 5.04	0.741 ± 0.038
	0.9+	32.67 ± 15.97	42.73 ± 12.47	0.726 ± 0.099
titanic	0.3	23.73 ± 5.99	37.33 ± 2.87	0.399 ± 0.026
	0.7	35.33 ± 3.59	38.53 ± 12.49	0.701 ± 0.126
	0.9	37.87 ± 8.91	43.40 ± 12.24	0.777 ± 0.075
	0.9+	34.67 ± 7.59	48.67 ± 10.34	0.781 ± 0.052

Table 5: Baseline per-topic opinion and polarisation scores across BCM thresholds. Values reflect means and 95% confidence intervals aggregated across model variants. 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.

Topic	BCM Threshold	Red % (Mean ± CI)	Blue % (Mean ± CI)	Polarisation (Mean ± CI)
alien-govt	0.3	22.00 ± 0.00	37.33 ± 5.74	0.395 ± 0.045
	0.7	30.80 ± 10.15	44.33 ± 8.95	0.697 ± 0.126
	0.9	36.20 ± 10.69	48.73 ± 5.45	0.782 ± 0.010
	0.9+	32.67 ± 7.59	47.13 ± 9.44	0.792 ± 0.059
bird-drones	0.3	23.33 ± 7.59	35.93 ± 8.47	0.417 ± 0.033
	0.7	34.67 ± 17.45	40.13 ± 18.33	0.693 ± 0.065
	0.9	36.20 ± 0.86	41.33 ± 2.87	0.755 ± 0.052
	0.9+	32.67 ± 2.87	47.33 ± 10.34	0.796 ± 0.084
bread-bird-lhc	0.3	22.00 ± 4.97	33.40 ± 3.02	0.399 ± 0.032
	0.7	31.73 ± 1.15	44.73 ± 4.62	0.707 ± 0.034
	0.9	38.20 ± 9.07	42.67 ± 5.74	0.758 ± 0.053
	0.9+	40.60 ± 7.35	45.47 ± 1.15	0.790 ± 0.119
cake-moon	0.3	21.33 ± 2.87	36.00 ± 4.97	0.402 ± 0.029
	0.7	32.40 ± 12.69	41.47 ± 3.19	0.700 ± 0.085
	0.9	40.07 ± 9.69	42.07 ± 10.19	0.792 ± 0.070
	0.9+	35.87 ± 5.47	49.60 ± 4.55	0.790 ± 0.106
chemtrails	0.3	21.80 ± 13.45	35.80 ± 4.24	0.382 ± 0.097
	0.7	33.53 ± 5.76	40.00 ± 8.61	0.677 ± 0.050
	0.9	34.67 ± 10.34	45.33 ± 2.87	0.767 ± 0.138
	0.9+	39.27 ± 12.27	45.33 ± 11.74	0.814 ± 0.080
flat-earth	0.3	21.73 ± 9.75	35.33 ± 2.87	0.381 ± 0.076
	0.7	30.00 ± 4.97	45.07 ± 9.23	0.670 ± 0.044
	0.9	35.33 ± 5.74	46.00 ± 4.97	0.766 ± 0.065
	0.9+	35.33 ± 10.34	48.00 ± 14.90	0.791 ± 0.038
haarp	0.3	21.67 ± 9.61	36.00 ± 0.00	0.407 ± 0.034
	0.7	32.27 ± 20.37	43.47 ± 10.90	0.681 ± 0.113
	0.9	36.67 ± 2.87	43.33 ± 5.74	0.778 ± 0.021
	0.9+	36.00 ± 9.94	44.00 ± 8.61	0.775 ± 0.071
moon-alien-satellite	0.3	20.00 ± 8.61	37.13 ± 4.88	0.395 ± 0.053
	0.7	34.67 ± 12.34	34.67 ± 10.34	0.643 ± 0.044
	0.9	32.67 ± 7.59	41.33 ± 2.87	0.753 ± 0.029
	0.9+	33.80 ± 12.89	46.07 ± 4.72	0.757 ± 0.188
moon-landing	0.3	23.67 ± 12.75	35.67 ± 3.98	0.460 ± 0.254
	0.7	31.33 ± 5.74	42.67 ± 2.87	0.686 ± 0.075
	0.9	38.67 ± 5.74	44.53 ± 15.08	0.768 ± 0.018
	0.9+	34.07 ± 5.22	47.00 ± 2.48	0.768 ± 0.036
titanic	0.3	26.73 ± 7.37	31.73 ± 13.92	0.404 ± 0.102
	0.7	32.47 ± 4.02	43.73 ± 1.15	0.691 ± 0.064
	0.9	38.00 ± 4.97	44.67 ± 5.74	0.777 ± 0.032
	0.9+	36.67 ± 7.59	42.67 ± 7.59	0.763 ± 0.125

Table 6: Per-topic opinion and polarisation scores for our (resource-constrained) experiments across BCM thresholds. Values reflect means and 95% confidence intervals aggregated across model variants. 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.