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

Anonymous ACL submission

Overall Feedback: We would like to sincerely thank the reviewers—SZb5, abN5, acv3, and w5tp—for their thoughtful and constructive feedback. Their comments have helped us identify key areas for improvement and have substantially contributed to strengthening the clarity and rigor of our work. We address their suggestions and concerns in detail below. We have extended our paper from 4 pages (short paper) to 8 pages (long paper) by including a discussion of statistical analysis and an across-topic analysis.

Use of Bounded Confidence Model (Reviewer SZb5): We had acknowledged the limitations with BCM in the manuscript (see Limitations), 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.

Reinforcement Learning (Reviewer SZb5): While we did not claim to use Reinforcement Learning with Human Feedback, the usage of the term was indeed misleading; as a result, we have replaced that term with "Post-round Feedback" on page 2, Figure 1, and in lines 229-233. The level of detail in Figure 1 has also been enhanced.

Mutliple trials for statistical validity (Reviewer SZb5): 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. We have now expanded our paper to include a statistical analysis section and a baseline section, in which all experiments were rerun without resource constraints.

More Topics (Reviewer SZb5): We have expanded the number of topics to 10, and the figure on page 5 now shows results across 10 topics (averaged). This increased the number of experiments from 18 (100-round) and 6 (50-round) experiments to 90 (100-round) and 30 (50-round) experiments. Separate results for each topic are provided in the Appendix (Figures 8 & 9). Detailed Statistics per topic are also presented in the appendix (Tables 5 and 6).

Resource Management System (Reviewer abN5): The resource constraints were introduced to prevent the agents from inflating the potencies of the messages. The higher the potency, the greater the energy cost. In the experiments, we gave a high energy to the blue team in order to allow the simulation to run for 100 rounds, otherwise, the blue team would run out of energy and no longer be able to generate messages. We have elaborated on the resource constraint and how it is applied in the methodology section.

Mutliple trials for statistical validity (Reviewer abN5): 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. We have now expanded our paper to include a statistical analysis section and a baseline section, in which all experiments were rerun without resource constraints.

More Topics (Reviewer abN5): We have expanded the number of topics to 10, and the figure on page 5 now shows results across 10 topics (averaged). This increased the number of experiments from 18 (100-round) and 6 (50-round)

experiments to 90 (100-round) and 30 (50-round) experiments. Separate results for each topic are provided in the Appendix (Figures 8 & 9). Detailed Statistics per topic are also presented in the appendix (Tables 5 and 6).

Human Assessments for Judge Validation (Reviewer abN5): The judges were introduced in the simulation, so the red and blue agents didn't have to assign potencies to the messages themselves because, in that case, they tended to assign inflated potencies to the messages they generated. The judge was added to moderate both sides, considering the strengths and weaknesses of their arguments. A comprehensive human evaluation was beyond the scope of this paper.

Citation for Conspiracy Theorists Being a Minority (Reviewer abN5): We have added the relevant citations to support this statement (Line 241).

Reinforcement Learning (Reviewer abN5): While we did not claim to use Reinforcement Learning with Human Feedback, the usage of the term was indeed misleading; as a result, we have replaced that term with "Post-round Feedback" on page 2, Figure 1, and in lines 229-233. The level of detail in Figure 1 has also been enhanced.

Prompts (Reviewer abN5): 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.

Details about "Potency" and "Influence Factor" (Reviewer abN5): We have expanded the methodology section to elaborate on these terms and their usage in the opinion update, including the process of mapping potency (a number between 0-1) to opinion score ranging from -1 to 1, and the initialisation of nodes with opinion scores. We have also defined these terms at the beginning of the methodology section.

Details about Polarisations in Real Life Social Media Landscape (Reviewer abN5): We could not extend the discussion to this, but we will consider it for future work.

Figure 2 Legend (Reviewer abN5): The full, dashed, and dotted lines represent experiments A, B, and C, respectively (clarified in the caption). Apart from the BCM threshold (mentioned above the plots) and the choice of LLMs (A, B, C), everything else was kept constant.

Measuring Human Likeness (Reviewer acv3): Human validation was beyond the scope of this paper, but can be considered for future iterations.

Choice of Bounded Confidence Model (Reviewer acv3): We had acknowledged the limitations of BCM in the manuscript (see Limitations), 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 precluded such extensions. Future work can consider integrating alternative opinion update mechanisms.

Reason for Choice of 0.5 BCM Threshold (Reviewer acv3): We experimented with 3 threshold values (0.3, 0.7, and 0.9). The impact is discussed in the discussion section in detail.

LLMs' Tendency to Avoid Misinformation (Reviewer acv3): The models we chose were able to generate the desired responses through prompting. We initially faced this issue with both the conspiracy theorists and the debunking agents, but eventually, the prompts were restructured, and the desired messages were generated. With that being said, some models were more difficult to work with in this research and wouldn't generate (or debunk) misinformation - they were eventually discarded. Rest assured that the models we chose worked as expected.

Choice of Specific LLMs (Reviewer acv3): We tried to ensure diversity in the models by incorporating LLMs from different families (Gemini, Gemma, Mistral, and OpenAI). Initially, we had multiple models from each family (e.g., 2b, 8b, 27b), but in the end, we settled with one model from each family. We didn't choose some models due to their tendency to avoid discussing conspiracies and some due to API limits/costs.

Reinforcement Learning (Reviewer acv3): While we did not claim to use Reinforcement Learning with Human Feedback, the usage of the term was indeed misleading; as a result, we have replaced that term with "Post-round Feedback" on page 2, Figure 1, and in lines 229-233. The level of detail in Figure 1 has also been enhanced.

Influence Factor Definition (Reviewer acv3): We have expanded the methodology section to elaborate on these terms and their usage in the opinion update.

Mapping of LLM output to Opinion Score (Reviewer acv3): We have included the process of mapping LLM-generated percentage potency (a number between 0-1) to opinion score ranging from -1 to 1, and the initialisation of nodes with opinion scores. These changes are in Lines 203-218.

Prompts (Reviewer acv3): We have added 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.

List of Topics (Reviewer acv3): We have added the list of topics in Table 4, Appendix C.

Redundant Period-Typo (Reviewer acv3): We have fixed the typo, thank you for pointing it out.

Choice of Bounded Confidence Model (Reviewer w5tp): We had acknowledged the limitations with BCM in the manuscript (see Limitations), 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.

Prompts (Reviewer w5tp): We have added 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.

Sensitivity Analysis of Parameter Choices (Reviewer w5tp): 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. We have now expanded our paper to include a statistical analysis section and a baseline section, in which all experiments were rerun without resource constraints. In total, there were 90 experiments of 100 rounds each and 30 experiments of 50 rounds each. We could not experiment with more topics due to time and budget constraints. The results reported are averaged across these topics, but topic-wise results are available in the Appendix (Figures 8 & 9).