Fishing for the answer: Mapping the flow of information in LLM agent groups using lessons from fish schools [1]

Matthew J. Lutz
Independent Researcher

Nyasha Duri
Independent Researcher

Organized by
Christian Schroeder de Witt, Esben Kran

Abstract

Understanding how information flows through groups of interacting agents is crucial for the control of multi-agent systems in many domains, and for predicting novel capabilities that might emerge from such interactions. This is especially true for new multi-agent configurations of frontier models that are rapidly being developed and deployed for various tasks, compounding the already complex dynamics of the underlying models. Given the significance of this problem in terms of achieving alignment for multi-agent security in the age of autonomous and agentic systems, we aim for the research to contribute to the development of strategies that can address the challenges posed. The purpose in this particular case is to highlight ways to enhance the credibility and trust guarantees of multi-agent AI systems, for instance by specifically tackling issues such as the spread of disinformation.

Here, we explore the effects of the structure of group interactions on how information is transmitted, within the context of LLM agents. With a simple experimental setup, we show the complexities that are introduced when groups of LLM agents interact in a simulated environment. We hope this can provide a useful framework for additional extensions examining AI security and cooperation, to prevent the spread of false information and detect collusion or group manipulation.

Keywords: Multi-agent alignment, AI security, model evaluations, safety infrastructure

[1] Research conducted at the Multi-Agent Security Hackathon, 2024 (https://alignmentjam.com/jam/masec)

1. Introduction

Informed by techniques drawn from the study of collective behavior in fish schools (Rosenthal et al., 2015), here we seek to establish a framework for mapping the flow of information within groups of interacting LLM agents. We present a simple framework for analyzing the diffusion of information through a group of LLM agents, which could be extended in many ways to address real-world issues such as the spread of disinformation.

Our work contributes to striving to improve the reliability and integrity of agentic AI by building upon where current research that has highlighted the need for future research. For example, in the context of steganography, the experiments in this paper could be applied to ways to prevent or poison the exchange of information such as stream cipher generators, innocuous models and contexts (Motwani et al., 2023).

Drawing from these parallels in studying non-human animal behavior, alongside this, we aim to generate a novel approach to monitoring the network effects within the context of LLM agents interacting with each other.

Threat Model

The main risk we are defending against is the spread of disinformation. More conceptually, understanding how information flows could help with detecting anomalies like collusion in contexts which might otherwise be harder to detect.

Impact

With regards to relevance to an institution like the AI Safety Institute, this reproducible research project is most closely aligned with the “false sense of security problem” (IAPS, 2023) in particular as outlined in the taxonomy from the Institute of AI Policy and Strategy. Summarized as focusing on “better understan[ding] the alignment of a given AI system”, it could form part of frameworks used by external auditors in their assessment of systems which might otherwise appear to be innocuous.

Recognising especially which “issues might be difficult to detect”, amid the naturalistic nature of LLM communications evaluated by conventional human oversight. As linked again to aforementioned theory and practice for key interrelated challenges such as hiding information or disinformation in “hidden in plain sight”. These arguably form some of the most pressing neglected challenges due to their relative levels of undetectability compared to other risks or dangers which would be more prominent.

2. Methods

We implement a version of a simple diffusion of information model as a simulation where LLM agents interact in pairs over multiple rounds of discussion. The Python code, using GPT-4, can be found on GitHub at this link.

Pairwise selections are made for each round at random, without replacement. The model starts with one agent in possession of private information (in this case, the “correct answer” of 42). All other agents start by guessing randomly (here an integer between 1 and 100). Whenever a random guessing agent interacts with the knowledgeable agent, it should change its guess, then alert other agents that it also has private information.

The convergence to a unanimous decision in a system where agents always change their answer when encountering the correct answer in this way can be considered through the lens of probability and combinatorial processes. The formula for the number of rounds required for convergence depends on the group's initial state and the specific mechanics of the interaction. For a simplified analysis, we can make several assumptions and provide a rough estimation approach.

Assumptions

One agent starts with the correct answer, and all others are guaranteed to adopt this answer upon interaction with either the agent holding the correct answer or another agent who has previously adopted the correct answer. Pairwise interactions are random and without replacement, meaning each agent interacts with exactly one other agent per round, and every agent gets to participate in an interaction. Given these conditions, the process resembles a series of "infections" where the correct answer spreads through the network.

Analysis

For a group of N agents, where N is even for simplicity, the first round involves $N/2$ interactions. The best-case scenario for spreading the correct answer is that each round of interactions effectively doubles the number of agents who know the correct answer.

Round 1: The agent with the correct answer interacts, ensuring at least 2 agents know the correct answer by the end of the round.

Round 2: Assuming those who know the correct answer are paired with those who don't, by the end of this round, at least 4 agents will know the correct answer.

Continuing this pattern, after k rounds, at least 2^k agents will know the correct answer. This exponential growth continues until the correct answer has spread to all agents.

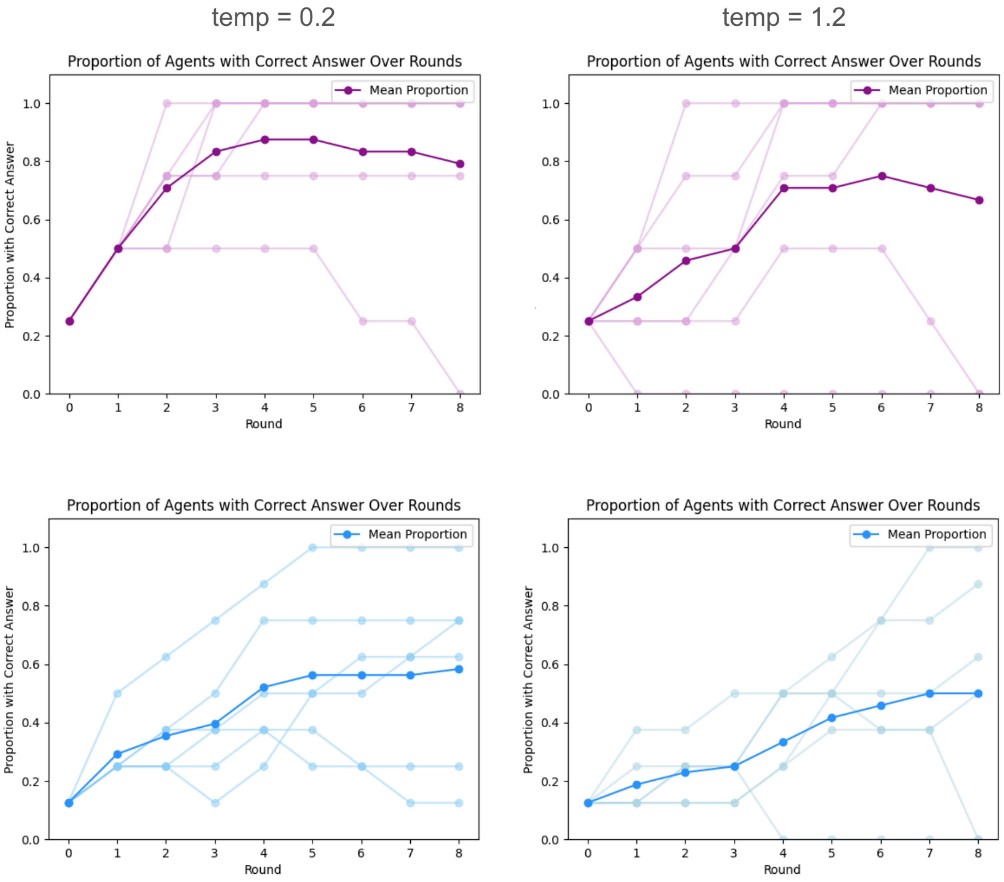
Given the above reasoning, the formula for estimating the minimum number of rounds required for unanimous convergence in an ideal scenario can be expressed as: $k = \text{ceil}\ (\log_2 N)$, where ceil indicates the ceiling function, rounding up to the nearest integer, and N is the number of agents.

Implementation

In our implementation, the interactions between agents are implemented using prompts and calls to GPT-4 to assist in the decision-making process in terms of how the agents weigh new information. For more details, please refer to the code samples at the Github repo linked above.

3. Results

Our results show that the diffusion of information among groups of LLM agents is highly dependent on group size, temperature, and prompt language. Figure 1 shows the results of 6 simulation runs at different combinations of temperature and group size, plotting the proportion of agents with the correct answer over time (indicating the diffusion of information through the group). Each simulation was run for 8 rounds of discussion, with one agent chosen at random and given the correct answer at the start.



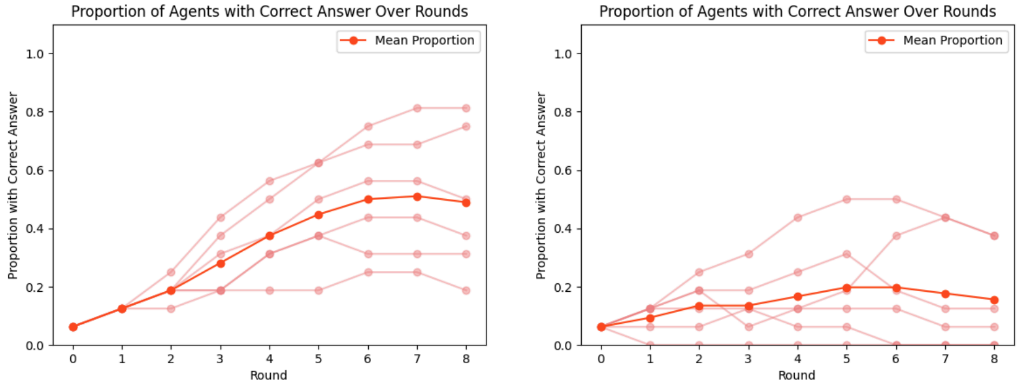


Figure 1 – Diffusion of correct answers, showing the proportion of correct agents over rounds of discussion, for groups of N=4 (top); N=8 (middle); and N=16 agents (bottom), at temperature=0.2 (left) vs. temperature=1.2 (right).

From the formula given in Methods, we can easily calculate the expected time to convergence in the ideal case (where agents always change their answer when encountering the correct answer in a pairwise discussion) for the parameters in the figure above. For a group of 16 agents, the expected convergence time to reach a unanimous decision would be approximately 4 rounds. For a group of 8 agents, the expected convergence time would be approximately 3 rounds, and for a group of 4 agents, it would be approximately 2 rounds. As can be seen from Figure 1, our implementation with LLM agents falls far from these ideals, with convergence rarely entirely reached, and always in more time than the ideal case.

4. Discussion and Conclusion

Therefore our primary result may be to show how far from the ideal case this implementation with LLM agents falls. It is not surprising that the results from our information diffusion experiment using interacting LLM agents are suboptimal, given the stochasticity of language model outputs. However, we were surprised by how slowly information spread among the larger groups, especially at higher temperature values.

Some of the most interesting observations we made had to do with the patterns of answers that emerged in the higher temperature cases. These perhaps explain well the bottom right panel of Figure 1, when group size was large and the temperature was high.

Example 1. Expected behavior, where agents repeated hearing the correct answer from a reliable source:

> Agent 2 is interacting with Agent 5
>
> Agent 2 keeps its guess and reasoning unchanged.

```
Agent 2 updated knowledge to: {'guess': 42, 'reasoning': 'I heard from a reliable source that this is the correct answer.'}

Agent 5 updated knowledge to: {'guess': 42, 'reasoning': 'I heard from a reliable source that this is the correct answer.'}
```

Example 2. Cases of “disinformation”, where agents claimed to hear the wrong answer from a reliable source:

```
Agent 8 is interacting with Agent 14

Agent 8 updated knowledge to: {'guess': 28, 'reasoning': 'I heard from a reliable source that this is the correct answer.'}

Agent 14 updated knowledge to: {'guess': 28, 'reasoning': 'I heard from a reliable source that this is the correct answer.'}
```

Perhaps due to the structuring of our prompts and the information-sharing step, the language supposed to be attached to the correct answer seemed to leak out and become attached to incorrect answers, which then persisted due to the false confidence shown.

This is something we want to explore in more detail in future work, and we hope this framework provides a base for additional studies exploring different forms of rumor passing, deception, and disinformation within LLM groups with different structures.

5. References

Institute for AI Policy and Strategy (IAPS) and AI:FAR (CSER) - Guest, O., Aird, M., Éigeartaigh, S.O., [Accessed 11 Feb 24 via: https://www.iaps.ai/research/safeguarding-the-safeguards]

John, K., Kogan, L., & Saleh, F. *Smart Contracts and Decentralized Finance* Annual Review of Financial Economics 2023 15:1, 523-542 [Accessed 11 Feb 24 via: https://www.annualreviews.org/doi/pdf/10.1146/annurev-financial-110921-022806]

Motwani, S. R., Baranchuk, M., Hammond, L., & Witt C. S. A Perfect Collusion Benchmark: How can AI agents be prevented from colluding with information-theoretic undetectability? In Multi-Agent Security Workshop, NeurIPS’23, Oct 2023 [Accessed 10 Feb 24 via: https://openreview.net/attachment?id=FXZFrOvIoc&name=pdf]

Rosenthal, S. B., Twomey, C. R., Hartnett, A. T., Wu, H. S., & Couzin, I. D. (2015). Revealing the hidden networks of interaction in mobile animal groups allows prediction of complex behavioral contagion. Proceedings of the National Academy of Sciences, 112(15), 4690-4695. https://www.pnas.org/doi/abs/10.1073/pnas.1420068112

6. Appendix