Towards Interpretable Models of Rumor Spread in LLM-Driven Agent Societies

Rahul Nandakumar*
McCombs School of Business
The University of Texas at Austin
Austin, Texas, US
rahul.nandakumar@utexas.edu

Pallavi Desai*
Department of Computer Science
The University of Texas at Austin
Austin, Texas, US
pallavi.desai@utexas.edu

Abstract

Rumors and misinformation shape public opinion and collective behavior, yet the mechanisms governing their spread in complex social systems remain opaque. We seek to advance the study of rumor dynamics by combining graph-theoretic analysis with LLM-agent simulations to exhibit richer agent behaviors such as confidence biases, emotional framing, and memory effects. Our results demonstrate how rumors spread, what network features accelerate or hinder them, and practical guidance for mitigating the spread of rumors. Code is available at: `https://github.com/RahulMark42/Networks`

1. Introduction

Rumor and misinformation diffusion have been modeled through epidemic-style models such as SIR [1]. While these models demonstrate how network structure influences information flow, they reduce human communication to abstract probabilistic rules. Therefore, the study of cognition, memory, and interpretation in rumor spread is limited. We formally pose the motivation to our work as follows: *How do different network structures influence the spread, mutation, and containment of rumors in societies?*

It has been shown that LLM based agents can act as a means of simulating social behavior and information exchange [2] [3]. We develop a framework where each agent's decision: whether to share, ignore, or modify a rumor, depends on its persona, the rumor text, and its network position. By combining graph-theoretic analysis with LLM-agent simulations, we aim to move beyond descriptive modeling toward interpretable, explanatory insights into how misinformation propagates and how it can be mitigated in real societies.

2. Prior Work

Traditional models like SIR have been the foundation for information and rumor diffusion within social networks. Mathematical models established the theoretical groundwork when Duan et al. [4] extended the Maki-Thompson model of rumor spreading to show that the proportion of a population that has never heard a given rumor converges to a limiting constant as population size increases. This work generalizes Sudbury's [1] findings, which show that mathematical models can predict final rumor reach, but it does not capture the complexity of human behavior.

A more transformative shift came with the advent of LLM-based agent simulations. Gao et al. [2] introduced S3 (Social-network Simulation System) which uses LLM agents to simulate users in a social network while capturing emotion, attitude, and interaction behaviors. They simulated information propagation, attitude changes, and emotional contagion in both individuals and populations with considerable accuracy. Park et al., [3] building on this framework, had a memory stream to maintain comprehensive experiential records, reflection mechanisms to synthesize memories, and planning systems to translate conclusions into behavioral sequences. These agents were able to diffuse information, form relationships, and coordinate group activities, resulting in more complicated social dynamics being captured.

Our work captures the nuances of human behavior, such as how different personalities can affect rumor propagation, that cannot be represented in a mathematical model. We are building on the work of S3 but we focus on various network topologies and personal characteristics and how they impact the way the rumor propagates, rather than specific agent behaviors.

3. Approach

We model rumor diffusion in social systems using three network topologies. Each node represents an individual agent capable of observing and sharing a rumor, and each edge represents a social connection through which information can be transmitted. We generate networks of comparable size ($N = 1000$) and mean degree ($\bar{d} \approx 20$). Specifically the networks are, Erdős–Rényi (ER) (a random graph

*Equal contribution

where each pair of nodes is connected independently with probability $p = \frac{\bar{d}}{N-1}$), Barabási–Albert (BA) (a preferential attachment model generating a scale-free network with high-degree hubs), and Watts–Strogatz (WS) (a small-world network balancing local clustering and global reach, representative of tightly knit communities connected by occasional long-range ties.)

Our main research questions are: (1) How does the structure of a social network affect the spread of a rumor? (2) How do different agent personas influence the spread? (3) How does using an LLM agent to make decisions change the dynamics compared to a probability? For example, suppose you have a network of 6 people and you seed the rumor at the most connected node (node 0). Node 0 exposes 1 and 3. Each node decides whether to share the rumor based on their persona and LLM output, and the process repeats for several rounds. This allows us to measure how many people eventually hear the rumor, how quickly the rumor reaches half the network, and which personas are most likely to share. We initially assume that the network is static once generated, nodes are homogeneous except for persona, edges are undirected, rumors are spread in discrete rounds, and agent decisions are independent. These assumptions allow us to clearly assess how the three network types impact rumor spread without external confounding variables.

4 Experimental Setup and Results

The networks we generate and the real-world networks have the following properties: Facebook and Twitter statistics are computed on the union of five ego-networks, whereas rumor propagation is simulated on each ego-network separately and averaged. We could possibly extend this experiment to Reddit networks in the future with the appropriate data.

Table 1: Network Properties

Graph	Nodes	Edges	Avg Degree	Avg Path Len	Diameter	Avg CC
Erdős–Rényi	1000	9936	19.87	2.64	4	0.02
Scale-Free	1000	9900	19.80	2.55	4	0.06
Small-World	1000	10000	20.00	3.22	5	0.52
Facebook	1654	33482	40.49	4.38	15	0.54
Twitter	580	4554	15.70	3.69	10	0.36

4.1 SIR Model

In our first set of experiments, we used a classical SIR (Susceptible, Infectious, Recovered) diffusion model to initially study rumor diffusion. We assign each node a persona (*skeptical*, *neutral*, *conformist*, or *sensationalist*) drawn from a weighted distribution. At each discrete round, exposed agents decide whether to `SHARE` or `IGNORE` the rumor using an LLM-based function, which integrates persona bias and rumor semantics. We acknowledge the limitation that LLM behaviors are often non-transparent, and

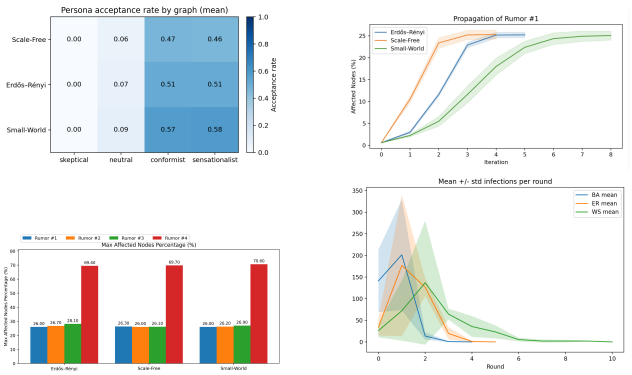

Figure 1: Persona acceptance rates by graph type. (Top left) Cumulative propagation of Rumor #1. (Top right) Maximum affected fraction for four rumors across networks. (Bottom Left) Maximum affected fraction for four rumors across networks. (Bottom right) Mean $\pm$ std of infections per round across graph types.

in further research, an element of explainability could be added, such as attention maps or feature attribution, to mitigate this. The process continues until there are no new nodes to share or maximum rounds are reached.

We test four synthetic rumors with increasing sensational and conspiratorial framing. We see in Figure 1 that (1) Persona effects dominate: Conformist and sensationalist agents exhibited high sharing probabilities ($0.5-0.6$), while skeptical and neutral personas rarely shared (< 0.1). (2) Structural topology influences speed: ER and BA networks reached saturation within 2–3 rounds due to high connectivity and hubs, while WS networks spread more gradually but sustained activity longer, consistent with small-world clustering effects. (3) Context of the rumor amplifies reach: Sensational rumors (e.g., Rumor #4) achieved up to 70% spread across all topologies, compared to $25-30\%$ for neutral content. Overall, the results suggest that while network structure governs the tempo of rumor propagation, persona composition and rumor tone determine its extent.

4.2 SEIR Model

We note that real rumor diffusion does not follow pure SIR dynamics. In real online platforms, seeing a rumor does not immediately imply re-sharing. We propose an SEIR (Susceptible, Exposed, Infectious, Recovered) based model where the latent period (S $\rightarrow$ E $\rightarrow$ I) captures the delay between seeing a post and deciding whether to share and infectious duration (I $\rightarrow$ R) models some finite attention - users share for only a few rounds. We say a user shares the rumor only after accumulating m exposures, and infectious users contact only a fixed number of neighbors per round.

For the next set of experiments, we use a small panel of COVID-era rumors with human-assigned true/false labels

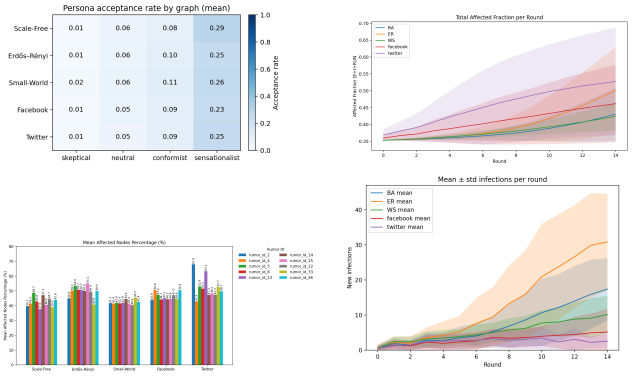

Figure 2: Speed and intensity of node infections and personas that contribute in SEIR.

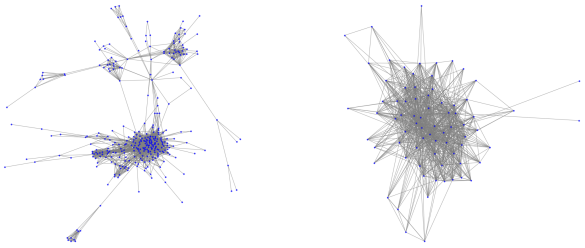

Figure 3: Facebook (left) and Twitter (right) ego networks.

from the COVID-19 Rumor Dataset [5]. We see from Figure 2, in early rounds $t \leq 4$, there are very small infection counts (1–5 new nodes), because the agents must accumulate ≥ 2 exposures, and wait through a latent period before becoming infectious. ER graphs show the steepest increase in new infections. In synthetic small world networks, high clustering and limited shortcuts seem make them naturally rumor-resistant under SEIR. We see similar behaviour in real-world ego networks on Facebook and Twitter. Facebook ego networks typically consist of highly clustered friend groups with a low number of bridging edges, while Twitter ego networks have more cross-community "weak ties". We further reinforce our conclusion for our SIR model, where we see that in an SEIR style rumor spreading model, persona effects conformist and sensationalist agents shared more $(0.1-0.3)$, while skeptical and neutral personas rarely shared (< 0.1).

4.3 SEIR Model with Fact Checking

We now extend our SEIR rumor-spreading model with an explicit fact-checking mechanism. A small fraction of nodes (5%) are designated as fact-checkers. They are chosen by top degree, so in BA graphs these are almost exactly the hubs. After a fixed latency of one round, each fact-checker sends a correction message to its neighbors. On

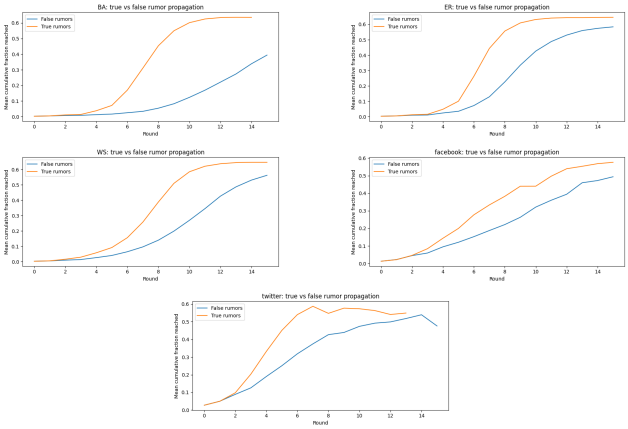

Figure 4: True/false rumor propagation for network types.

all network types, true rumors consistently outrun false rumors. In BA graphs, assigning these fact-checkers to hubs targets the main backbone of the network. When a false rumor tries to use the hubs as amplifiers, those nodes have already had their sharing probability reduced, so the cascade slows and often fails to reach deep into the network, hence the lower False curve. For the Facebook and Twitter ego networks, the ego node and its high-degree friends form a dense local core. Fact-checking the core is enough to significantly slow false rumors, but high clustering means they can still persist inside local communities, hence the smaller but non-zero gap between true and false curves.

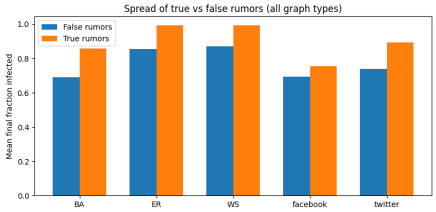

Figure 5: Fraction of infected nodes for true vs. false rumors over all graph types.

Across all graph types, true rumors infected a larger fraction of the network than false rumors. For BA, true rumors finish around 80% vs 70% for false. Compared to Facebook & Twitter ego networks, both labels spread less because of smaller, clustered components, but true rumors still have a $\approx 10\%$ advantage.

We can see that fact-checking seems to be very effective. High-degree hubs dominate information flow, so placing fact-checkers there creates bottlenecks that significantly reduce false-rumor spread. In real ego networks (Facebook, Twitter), strong clustering and low-degree regions create pockets that are harder for fact-checks to reach, yet correct-

ing the ego and its core neighbors still meaningfully dampens false propagation. Overall, we can say that effectiveness of fact-checking depends strongly on topology: the more centralized or uniformly connected the network, the more aggressively false rumors are suppressed.

4.4 Modularity and the Echo Chamber Effect

To isolate the role of community structure, we generate synthetic networks using a stochastic block model (SBM) with $N = 1000$ nodes and $K = 5$ equal-sized communities, as shown in Figure 6. Each node has average degree $\bar{d} = 20$. We control the strength of community structure using a mixing parameter $\eta \in \{0.02, 0.05, 0.10, 0.20, 0.35, 0.50\}$. We acknowledge that a limitation of synthetic data is that it may not be representative of real-world situations; however, we choose to use it here to create a baseline for these networks so the findings can be taken forward into further research on real datasets.

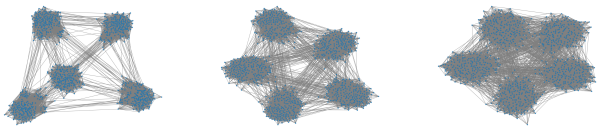

Figure 6: Synthetic networks using a stochastic block model.

Across all modularity values, most runs reach very large cascades: $\approx$ 90–100% of nodes become affected once the rumor manages to escape the seed community.

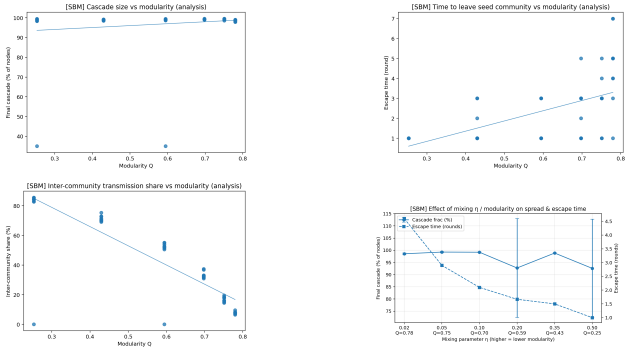

Figure 7: Comparing SBM modularity to various metrics.

We see in Figure 7 that modularity has only a weak effect on final outbreak size. In other words, if the rumor escapes one community, the rest of the network is still structurally dense enough to support a cascade. There are also some points around 35–40% final size corresponding to runs where the rumor fails to leave the seed community (or dies out shortly after escaping). We observe a positive trend between modularity Q and escape time: higher Q networks show longer rounds before the rumor reaches a different community and for low Q networks, escape happens within 1-2 rounds. There is a strong negative correlation between modularity and inter-community transmission share. High Q graphs are dominated by within-community transmissions ($\approx$10–20% inter-community). We can see that these high modularity networks compare to speaking in an *echo chamber*, as the voice keeps echoing within the room, most spreading is seen intra-community, and cross-community edges play a very minor role. Low modularity networks behave more like a well-mixed social space, where a large fraction of transmissions are inter-community.

4.5 SEIR with Rumor Mutation

We move to the next part of this discussion. Unlike traditional SIR/SEIR models where the nodes act as passive relays, we update our framework to treat every agent as a cognitive entity capable of processing and altering information. We introduce a mutation mechanism to track how content changes over time. We keep the agent personas distributed similar to the previous scenarios, and in addition we designate 5% of nodes as *Correctors*. To quantify information distortion, we introduce the following metrics:

1. We utilized Sentence-BERT (all-MiniLM-L6-v2) to generate embeddings of size 384 for both the original ground-truth rumor (v_{orig}) and the mutated variant (v_{gen}). Then we measure the preservation of the original meaning using Cosine Similarity. The similarity score is defined as $S = \cos(\theta) = \frac{v_{orig} \cdot v_{gen}}{\|v_{orig}\| \|v_{gen}\|}$. We also use UMAP (Uniform Manifold Approximation and Projection) to visualize the 384-dimensional SBERT embeddings of all rumor variants by reducing them to a 2-dimensional projection.

2. We quantified how *natural* or *surprising* the mutated text appears to a language model. For a tokenized sequence $X = (x_1, x_2, \ldots, x_t)$ of length t, the probability of the sequence is modeled autoregressively as:

$$P(X) = \prod_{i=1}^{t} P(x_i \mid x_1, \ldots, x_{i-1})$$

The Cross Entropy Loss, is calculated as:

$$\mathcal{L}(X) = -\frac{1}{t} \sum_{i=1}^{t} \ln P(x_i \mid x_1, \ldots, x_{i-1})$$

The Perplexity (PP) is then defined as,

$$PP(X) = \exp(\mathcal{L}(X)) = \exp\left(-\frac{1}{t} \sum_{i=1}^{t} \ln P(x_i \mid x_{<i})\right)$$

Lower perplexity indicates more fluent, naturalistic text. We employed a pre-trained GPT-2 model to calculate this perplexity. In our framework, Llama-3 acts as the *Writer* who generates the rumors, and GPT-2 acts as the *Judge* to evaluate the writing.

3. We assessed the accessibility of the text using the Flesch-Kincaid Grade Level which is used to estimate the U.S. school grade level required to understand the text based on sentence length and syllable count. This can be calculated as

$$\text{Grade Level} = 0.39 \left(\frac{\text{Total Words}}{\text{Total Sentences}} \right)$$

$$+11.8 \left(\frac{\text{Total Syllables}}{\text{Total Words}} \right) - 15.59$$

.

As shown in the readability distributions in Figure 8, evolved rumors maintained a consistent, high grade level (12–14, corresponding to high school/early college) with minimal variation across graph types. Our agents seem to preserve text complexity to align with their persona's voice rather than aggressively distort the content. In particular, BA and WS networks do not show a lower readability score, indicating that hubs and short path lengths accelerate diffusion, but do not inherently simplify the linguistic form of the rumor.

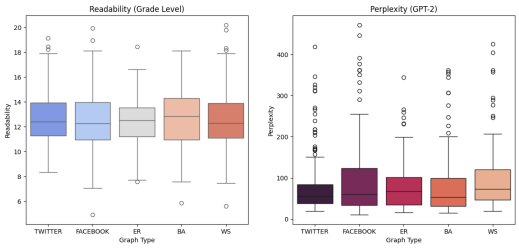

Figure 8: Perplexity and Readability Scores.

We investigate whether social clusters distort or preserve the truth. Our results indicate a counter-intuitive Semantic Anchor effect (Figure 9). In random networks, the semantic similarity of the rumor drops rapidly, losing nearly half its fidelity within 4 generations ($S < 0.7$). In the absence of community structure, the rumor mutates freely, leading to rapid *truth decay*. In contrast, highly clustered ego-networks (Facebook, Twitter) maintained significantly higher semantic similarity ($S > 0.85$) even after 6 generations. Despite having the highest volume of total edits, as shown in Figure 10, as agent activity is high in these networks, the tight-knit community structure appears to reinforce the original context. The local redundancy acts as an

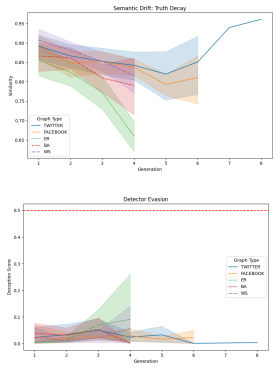

Figure 9: Semantic Drift and Fake News Detector evasion.

anchor, preventing the rumor from drifting too far from the ground truth.

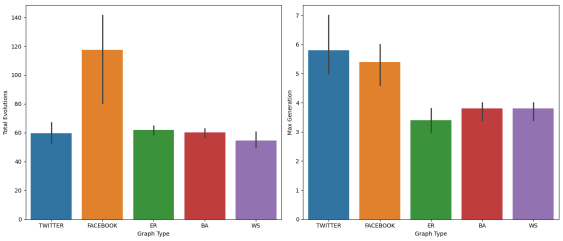

Figure 10: Total Evolutions and Number of Generations per Graph Type.

Finally, the UMAP projection (Figure 11) confirms that these mutations are not random. The rumor variants form tight, well-separated clusters corresponding to distinct generations. This indicates the emergence of *Generational Dialects*, where a *Generation 3* rumor is linguistically distinct from a *Generation 1* rumor. Points from later generations remain clustered near their ancestors rather than wandering arbitrarily, demonstrating that an evolutionary pressure organizes these rumor content into stable semantic groups.

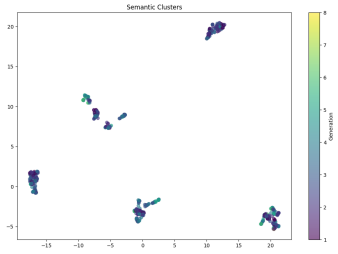

Figure 11: UMAP Semantic Clusters.

We also measure a deception score - the probability of bypassing a RoBERTa based fake-news detector. We set

this threshold at the standard value of 0.5, and check our rumours against it (Figure 9). While average scores remained low, specific evolutionary pathways, particularly in random networks showed a spike in the evasion probability by Generation 4. While our rumors aren't inherently *news* articles and therefore might not be picked up by fake news detectors, this implies a latent *mutation risk*, where a rumor can evolve into a variant that is statistically indistinguishable from real news.

Table 2: Qualitative Evolution: Examples of Persona-Driven Mutation

Original Rumor	**Evolved Variant (Gen 3)**	**Driver**
Staff at Gold Coast Hospital danced for a man in quarantine.	"The ENTIRE staff... ERUPTED into a choreographed dance party... the drama was simply EPIC!"	Sensationalist (Amplifies Drama)
Chinese doctors confirmed African people are resistant to COVID.	"Reports suggest that some African populations may have a genetic advantage... but this requires further evidence."	Skeptic (Sanitizes & Adds Doubt)

5. Conclusion

Our framework demonstrates that both network topology and persona-driven behavior fundamentally shape rumor diffusion dynamics. We observe that *sensational content* amplifies spread, and that *conformist* and *sensationalist* personas are the primary drivers of large cascades. In contrast, neutral and skeptical agents suppress diffusion, resulting in smaller and slower cascades.

The SEIR model captured several qualitative properties of real rumor cascades. Across synthetic networks (ER, BA, WS) and real ego networks (Facebook, Twitter), true rumors consistently spread faster and reach larger final cascades than false ones. This effect is driven by persona-dependent credibility, which collectively assigns higher sharing likelihoods to true content. Introducing fact-checking agents into the SEIR process demonstrated that even having small, strategically placed interventions can meaningfully slow or suppress false cascades. This provides valuable insights that can potentially be extrapolated to human behavior and help mitigate the spread of misinformation in the real world.

We also isolated the effect of network modularity using stochastic block model networks with controlled mixing. High-modularity networks acted as barriers or echo rooms, delaying cross-community diffusion and substantially reducing inter-community transmission. However, once a rumor escaped the seed community, the final cascade size remained large across all mixing levels. Thus, modularity primarily governs the timing and reliability of global cascades, not the eventual reach. Low-modularity graphs, by contrast, made faster, well-mixed propagation with higher inter-community transmission shares.

We learned that true information seems to spread more efficiently overall, misinformation requires more structural vulnerability to succeed, and fact-checking interventions are most effective when deployed on structurally central nodes. Community structure modulates when a rumor becomes global but does not determine not whether it does. This suggests that control strategies must address early-stage cross-community leakage.

Our introduction of LLM-driven mutation taught us that rumors do not degrade into noise; they obey a *Directed Evolution*. Agents optimized rumors for linguistic fluency, driving Perplexity down while maintaining high readability. We discovered that clustered social networks act as *Semantic Anchors*, preserving the truth $2\times$ better than random networks by providing local context reinforcement.

While we have successfully coupled graph theory with LLM based agentic simulations, there are several avenues that remain for future exploration: Our current model assumes static social graphs. Future work could incorporate dynamic edges where agents sever ties with neighbors who repeatedly share low-credibility rumors (trust-based rewiring) or form new connections based on shared beliefs (homophily). (1) We currently model one rumor in isolation. A more realistic simulation would involve multiple rumors competing for agent attention, like a *Battle of Misinformation* (2) Extending the simulation to Reddit-style networks, where we need to model *upvoting/downvoting* behavior and threaded discussions, which fundamentally differ from the *re-share* mechanic of Twitter/Facebook. (3) We could incorporate Vision-Language Models (VLMs) to simulate the mutation of visual memes and their impact on viral spread. For member contribution, Rahul was responsible for the majority of the coding part of the project due to the computational resources available in his lab. Pallavi contributed to the design of the networks, helping write the report, and creating slides for our progress presentations and weekly meetings.

References

[1] A. Sudbury. The proportion of the population never hearing a rumour, 1985.

[2] Gao et al. S^3: Social-network simulation system with large language model-empowered agents, 2025.

[3] Park et al. Generative agents: Interactive simulacra of human behavior, 2023.

[4] Y. Duan and A. Ganesh. The proportion of the population never hearing a rumour. 2021.

[5] C. et al. A covid-19 rumor dataset. *Frontiers in Psychology*, 12:644801, 2021.