A Prompt Set

Here we provide an in-depth description of the prompts used in our simulation to model the dynamics of fake news propagation and recognition.

1. Short-Term Memory Function Prompt (f_m^s)

The short-term memory function prompt simulates how individuals process the opinion information they receive in the immediate term. The form of this function is as follows:

> The discussed topic is [topic].
> Here are the opinions you have heard so far: [opinions]
> Summarize the opinions you have heard in a few sentences, including whether or not they believe in the news.

2. Long-Term Memory Function Prompt (f_m^l)

The long-term memory function prompt reflects how individuals integrate information into their long-term memory, influencing their enduring beliefs and behaviors. Its form is:

> Recap of Previous Long-Term Memory:[long memory].
> Today's Short-Term Summary: [short memory].
> Please update long-term memory by integrating today's summary with the existing long-term memory, ensuring to maintain continuity and adding any new insights or important information from today's interactions. Only return long-term memory.

3. Updated Prompt (f_o)

The updated prompt is based on the outputs from both the short-term and long-term memory prompts, determining the individual's final state of action and belief. The general form is:

> Based on the following inputs, update your opinion on the [topic]:
> 1. Previous personal Opinion: [opinion]
> 2. Long Memory Summary of Others' Opinions: [long memory]
> 3. Name: [agent name]
> 4. Trait: [agent persona]
> 5. Education level: [agent qualification]
> Keep in mind that you are simulating a real person in this role-play. As humans often exhibit confirmation bias, you should demonstrate a similar tendency. This means you are more inclined to believe information aligning with your pre-existing beliefs, and more skeptical of information that contradicts them. Your responses will be formatted in JSON. Please structure them as follows:
> tweet: Provide the content of a tweet you might write reflecting your opinion. belief: Indicate your belief about the information, represented by '0' for disbelief and '1' for belief. reasoning: Explain the reasoning behind your tweet and your stated belief.
> For example: {"tweet": "Trump was shot dead at the White House!", "belief": 0, "reasoning": "Trump is very likely to be killed by an assassin, so I believe this news"}

4. Official Clarification Prompt Issued by Official Agents

To combat the propagation of fake news, official agents issue official refutations to all other agents on designated days. The form of this prompt is:

> As the official spokesperson, I hereby issue a formal statement regarding the recent news report circulated on various media platforms concerning topic. After thorough investigation, we have determined that this report is unequivocally false. The sources of this information have been found to be unreliable, and the content significantly misrepresents the facts. We urge the public to refrain from spreading or trusting this misinformation and to rely on our official channels for accurate and timely information. We remain committed to upholding the truth and safeguarding the interests of the public.

B Analysis of Experimental Costs

In this section, we discuss the costs associated with our experiments using the GPT-3.5 API. The OpenAI pricing at the time of our experiment indicated that the cost per thousand tokens was 0.5 USD for every 1M input tokens and 1.5 USD for every 1M output tokens. Therefore, each experiment, which involved approximately 1 to 2M input tokens and 1 to 2.5M output tokens, incurred an estimated cost of approximately 2 to 5 USD.

Conducting real-world research often entails significant expenses, including compensating participants, constructing infrastructure, purchasing data, and the lengthy duration required to obtain results. Such studies can range from thousands to hundreds of thousands of dollars, depending on their size and scope. In contrast, by utilizing GPT-3.5, we can simulate complex social interactions and the propagation of information without the need for extensive surveys or real-world experiments. This comparison highlights the cost-effectiveness of our simulation approach compared to empirical studies conducted in real-world settings.

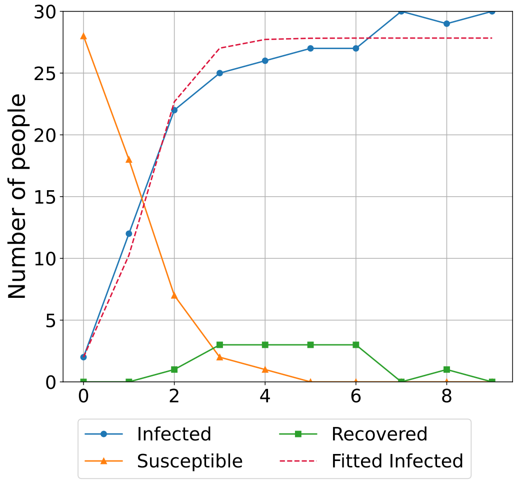

Figure 1: Dynamic group population number changes with GPT-4 as the backbone under the political topic.

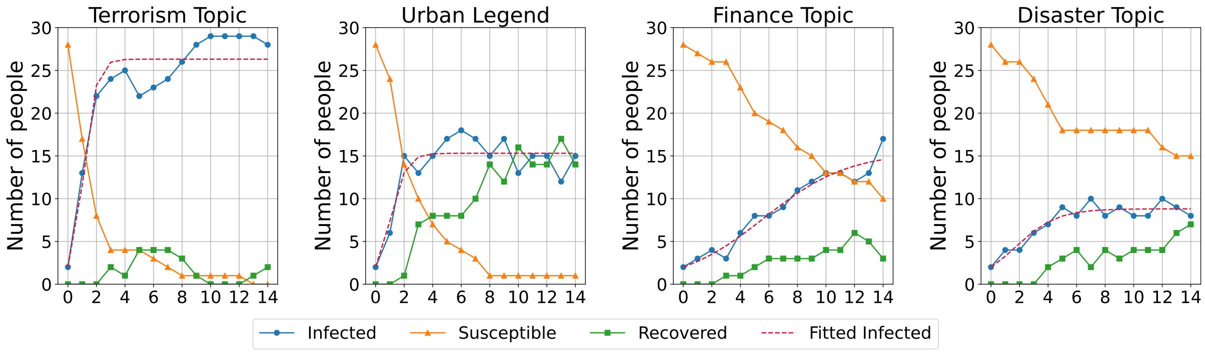

Figure 2: Dynamic group population number changes in terms of different topics, with an accompanying fitting curve based on the SIS model. The red dashed line represents the results of the SIS model fitting.

C Simulation on Different Backbones

Figure 1 illustrates the dynamic changes in group population numbers with GPT-4 as the backbone under the political topic. We observe that the number of individuals in the infected group grows rapidly, leading to a majority believing in fake news. This trend is similar to the performance observed with GPT-3.5 as the backbone. The performance under other topics is also consistent. Therefore, we can conclude that our simulation effectively captures the core behaviors of misinformation propagation and its impact on public opinion formation, regardless of the underlying technological backbone.

D Diverse Topics and Simulation Results

Figure 2 shows the simulation results concerning topics of terrorism, urban legends, finance, and nature disasters. In the context of terrorism, the number of infected individuals rapidly increases, reaching a peak in a short period before declining, yet the overall trend exhibits significant fluctuations. These fluctuations may indicate that public opinions on terrorism-related fake news are divided, and the verification of information could be dynamically changing. For the topic of urban legends, the fluctuating number of infected individuals demonstrates that public interest or trust in such information may vary over time, likely reflecting ongoing skepticism about the authenticity of these news stories. The curve for the finance topic shows an initial rise followed by a decline in the number of infected individuals, suggesting that there is a wave of attention and trust in fake news at the beginning, but most people are eventually able to identify the falsehoods and return to a susceptible state. In the case of nature disasters, the number of susceptible individuals continuously decreases while the number of recovered individuals gradually increases. This may indicate that people can quickly recognize and discard false information regarding disasters, thereby returning to a cautious stance toward such information.

These experimental results showcase the patterns of fake news propagation across different topics, revealing the behavioral patterns of the public in identifying and reacting to fake news. Combining the results in main text, our findings show that political fake news propagates notably faster than topics such as terrorism, natural disasters, science, urban legends, or financial information, consistent with previous studies [Vosoughi *et al.*, 2018].

Below, we detail the news items used in our experiments on fake news propagation across different topics.

- **Political** - *The 14th Amendment in the U.S. Constitution expressly specifies that a criminal conviction for insurrection is needed to disqualify a candidate from state or federal office.*
- **Science** - *Hurricane Has Been Known To Cross the Equator since 2003.*
- **Terrorism** - *World Trade Center leaseholder Larry Silverstein "fortuitously" took out terrorism insurance just months before the 9/11 attacks, implying foreknowledge which feeds into conspiracy theories.*
- **Urban Legends** - *The story of why a loon appears on Canada's one-dollar coin is because the original dies, featuring a different design, were lost in transit.*
- **Natural Disasters** - *A Dutch researcher nonetheless claimed to have predicted an earthquake in Turkey in early 2023, despite the inherent unpredictability of earthquakes.*
- **Financial Information** - *Hillary Clinton's 'Sudden Move' of $1.8 Billion to Qatar Central Bank Stuns Financial World.*

E Ablation Study

We also compare the diversity of information in simulations of fake news propagation between a full framework and one lacking short-term memory capabilities. We measure diversity using Distinct-1 and Distinct-2 metrics, which indicate the proportion of unique single words and two-word combinations, respectively. Results in Table 1 show the full framework achieved higher diversity scores (0.062 for Distinct-1 and 0.155 for Distinct-2) than the framework without short-term memory (0.045 for Distinct-1 and 0.121 for Distinct-2), demonstrating the full framework's ability to maintain greater information diversity in fake news propagation.

	Distinct-1	Distinct-2
Full Architecture	**0.062**	**0.155**
-w/o Short Memory	0.045	0.121

Table 1: Ablation Study on Short-Term Memory

References

[Vosoughi *et al.*, 2018] Soroush Vosoughi, Deb Roy, and Sinan Aral. The spread of true and false news online. *science*, 359(6380):1146–1151, 2018.