Effect-prompting shifts the narrative framing of networked interactions

J. Hunter Priniski
Department of Psychology
UCLA

Bryce Linford
Department of Psychology
UCLA

Darren Cao
Information Sciences Institute
USC

Fred Morstatter
Information Sciences Institute
USC

Jeff Brantingham
Department of Anthropology
UCLA

Hongjing Lu
Departments of Psychology and Statistics
UCLA

Abstract

Narrative interaction plays an important role in shaping people's beliefs and behaviors both online and in the offline world. We present an experiment examining whether a simple intervention of effect prompting—asking participants to list the effects of complex events—impacts the narrative framing of their networked interactions. After reading a text-based narrative about the Fukushima nuclear disaster, participants in a fully connected network interacted with their neighbors and received rewards for submitting hashtags that matched those of their network partners. Half of the groups received an *effect-prompting* intervention, which shifted participants toward producing more effect-oriented hashtags during networked interactions. We found that the effect-prompting instruction influenced the hashtags participants generated during the network interaction. However, the extent of this shift in hashtags depended on how likely the group was to achieve global coherence.We also examined these dynamics with networks of interacting large language model (LLM) agents using Llama-3.1-8B-Instruct. The study highlights how language-based prompting can subtly shift the narrative framing of online communication.

Keywords: Narrative interaction, networked group dynamics, natural language, large language models, hashtags

Introduction

Narrative interaction is central to human activity. The narratives that people discuss and interact with in networked environments have the potential to shape people's beliefs and behaviors both online and in the offline world. What cognitive and social mechanisms support the onset of shared narratives in networked groups remains an open empirical question. Recent advances in natural language modeling and experimental frameworks for studying group behavior make it possible to apply cognitive science principles to the study of narrative interaction in networked environments.

Narrative-based memory in humans

Understanding how narrative dynamics arise in networked groups begins with appreciating how narratives structure people's memories of information and events. When people read a how-to manual or a fictional story, or watch a movie or TikTok, they don't actively encode superficial linguistic and visual features in memory. Rather, language and visual features serve as a set of "instructions" for building coherent mental representations that explain and predict events in the world (Zwaan and Radvansky, 1998).

The cognitive psychology research on narrative processing has centered on discerning what semantic and causal information individuals encode and store when reading text-based narrative materials to form a coherent mental model of the narrative. Experiments have focused on how individuals interpret the situational context encoded by narratives, and how they store representations of narrative information in long-term memory, to enable efficient retrieval and effective higher-level reasoning (Morrow et al., 1989; Zwaan and Radvansky, 1998; Zwaan et al., 1995). While reading a text-based narrative, for instance, people represent time, space, causality, attributes, characters, and intentionality in memory—dimensions that constitute the core semantics of a situation model (Zwaan and Radvansky, 1998). They are sensitive to how narrative complexity (i.e., situation model complexity) affects information retrieval via the "fan effect," suggesting that more complex situation models are required by more intricate narrative structures and have a greater impact on memory and reasoning (Radvansky and Zacks, 1991). Furthermore, individuals actively generate inferences about characters' goals, causal connections, and temporal/spatial contexts, to fill gaps in what is explicitly stated by the text (Graesser et al., 1994). These active and intuitive inferences are crucial for constructing a coherent situation model from generally sparse text, and are reasons why human language is a parsimonious vehicle of rich and coherent forms of causal meaning.

How narrative frames emerge from networked group interactions

Narratives not only ground our representations of the world, they shape collective understanding of people's lives (McAdams, 2001), social groups (Polletta et al., 2011), and political activities (Adams et al., 2022; Papacharissi, 2015, 2016; Yang, 2016). Narratives are living entities in constant flux, changing as the beliefs and mental models of the people who endorse them change (Ochs and Capps, 2009; Priniski et al., 2023). Narrative interaction in networked environments (e.g., tweeting for political change, posting a TikTok of your morning routine) is a more agential process than passively reading a text-based narrative (Dawson, 2020; Papacharissi, 2016; Wong et al., 2022; Yang, 2016). Once passive, serialized receivers of top-down narratives (e.g., a priest dictating religious canon; a single news station covering global affairs), the public now can leverage networked communication channels to exert bottom-up narrative influences that shape group

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

understanding and public discourse (e.g., #BlackLivesMatter, #MeToo) (Nguyen et al., 2021; Priniski, Mokhberian, et al., 2021).

At the group-level, shared or polarized narratives can emerge from local interactions between individual nodes learning to coordinate communication behaviors with network neighbors. While coordinating narrative interactions, individuals must integrate background causal knowledge about narrative content with coordination rewards learned dynamically from social context (i.e., what do neighbors think/say about the narrative) (Priniski, Linford, et al., 2024; Priniski, Solanki, and Horne, 2024). Social learning strategies for optimal integration of background causal knowledge about narrative content with one's social context (i.e., outputs from others in their unique neighborhood) can result in different individuals sampling different narrative entities (e.g., characters, key events, causal relationships) when coordinating interaction behaviors with their network neighbors. How an individual allots attention across a set of narrative entities constitutes their *narrative frame*, which can shift over the course of networked interaction as an individual learns which entities facilitate higher-utility narrative interaction behaviors.

Dozens of experiments demonstrate that mixing interactions between all members of a group allows for shared behaviors to emerge, whereas connecting individuals to only a handful of neighbors results in subgroups aligning on different behaviors (Centola, 2015, 2022; Priniski, Linford, et al., 2024). In cases where groups align on a shared behavior, we assume that the constituent nodes have a shared narrative frame for interpreting evidence in that they have shared beliefs and behaviors. This is in contrast to cases where opposing network sub-clusters have opposing narrative frames, including echochambers which can learn to endorse radical narratives including conspiracy theories (Priniski and Holyoak, 2022; Priniski, McClay, and Holyoak, 2021). Here, we develop a simple intervention designed to shift the narrative frame of fully-connected groups by influencing which narrative variables participants attend to immediately prior to coordinating behaviors with network neighbors.

To measure shifts in narrative framing during networked interaction, we developed and applied a novel embedding-based *narrative alignment* measure that maps communication data to the causal variables in a discussed narrative (see Figure 1). Because network experiment data is costly and group behavior is highly variable, we also began preliminary work developing communication networks of generative agents for the purpose of manipulating group outcomes and narrative framing effects. As discussed in more detail below, advances in generative language models now allow researchers to model language-based changes over the course of communication, with implications for studying how narratives frame the content of generated language data among humans.

Simulating group-based narrative dynamics with Large Language Models

Recent advances in generative language models have led many cognitive scientists to apply them as models of human intelligence. LLMs have also been applied to study human collective behavior and group communication dynamics. Simulations using communication networks of LLMs appear to mirror key behavioral trends commonly observed in human social networks. For example, communicating LLMs form correspondence connectivity networks with degree distributions akin to those resulting from preferential attachment mechanisms (Chang et al., 2024; De Marzo et al., 2023; Papachristou and Yuan, 2024), a central mechanism behind the structuring of numerous online and offline networks (Newman, 2001). Networks of communicating LLM-based agents can also exhibit well-known social phenomena including homophily. For example, He et al. (2024) found that interacting AI chatbots form distinct communities based on shared language and content preferences, although the magnitude and robustness of such phenomena is still uncertain (Chang et al., 2024).

Although simulation-based studies are valuable for understanding the dynamics of community formation and language change, human group behavior is immensely complex, and noise accumulates as interactions unfold over time (Thurner, 2024). Benchmarking simulation studies against matched network experiments with human participants is therefore essential to determine the contexts in which generative language models mirror the collective behaviors observed in human groups.

Systematic prompt-edit comparisons are also essential to test causal hypotheses about LLM behavior. Our motivation for the present simulation study is *not* to determine if LLMs can simulate human communication dynamics, but rather, is to first establish a starting point for testing which prompt features are necessary for LLM groups to reach consensus. Only once group behavior stabilizes with minimal prompting instruction can we then test subsequent prompt-edits that instantiate intervention effects like those described here. To this end, we tested a minimal control prompt to guide LLM interactions without access to the narrative text. Agents exclusively relied on social interaction data to update responses as interactions unfolded, doing so without background knowledge about narrative content crucial for framing communication.

Network experiment on narrative interaction in human groups

Participants

We sampled a total of $N = 420$ participants (16,800 interactions) from the Prolific subject pool, and placed them into a fully-connected online social network (three groups of $N = 20$, 3 groups of $N = 50$). Half of these groups received an *effect-prompting* intervention before network interaction while the remaining groups did not.

Materials

Nuclear Disaster Narrative Across all networks, participants first read a four-paragraph narrative description of the Fukushima nuclear disaster prior to network interaction. Because this narrative frames subsequent network communication for all participants, we define this narrative as the *focal narrative*. The narrative explains how a large earthquake triggered a tsunami that caused damage to a nuclear reactor and resulted in radiation leaks, population displacement, and an energy-saving movement "Setsuden." We selected this narrative because it describes a rich set of causal relations (a generative causal chain producing a branching common cause sequence) and included both negative (e.g., displacement, poisoning) and positive effects (e.g., energy-saving movement). Fig. 1 illustrates the causal structure summarized from the Fukushima disaster narrative. Note that this causal diagram was not presented to participants in the experiment.

Preinteraction *effect-prompting* intervention Participants in the experimental group received an *effect-prompting* intervention that asked them to write the five effects of the nuclear disaster before producing hashtags via online communication. Following findings on the efficacy of prompts to shift online behaviors (Pennycook and Rand, 2022; Pennycook et al., 2020), the goal of effect-prompting instruction is to subtly increase participants' attention to the effects described by the narrative materials before engaging in narrative interactions (hashtags) with network neighbors.

0 The Fukushima Nuclear Disaster was a 2011 nuclear accident at the Daiichi Nuclear Power Plant in Fukushima, Japan.

1 The cause was the Tōhoku earthquake on March 11, 2011, the most powerful earthquake ever recorded in Japan.

2 It triggered a tsunami with waves up to 130 feet tall, with 45-foot tall waves causing direct damage to the nuclear power plant.

3 The damage Japanese authorities quickly implemented a 100-foot exclusion zone.

4 Large quantities of radioactive particles were found shortly after throughout the Pacific Ocean and reached the California coast.

5 The resulting energy shortage

6 Inspired media campaigns to encourage Japanese households and businesses to cut back on electrical usage, which led to the national movement *Setsuden* ("saving electricity").

7 The exclusion zone resulted in the displacement of approximately 156,000 people in years to follow.

8 Independent commissions continue to recognize that affected residents are still struggling and facing grave concerns. A WHO report predicts that infant girls exposed to the radiation are 70% more likely to develop thyroid cancer.

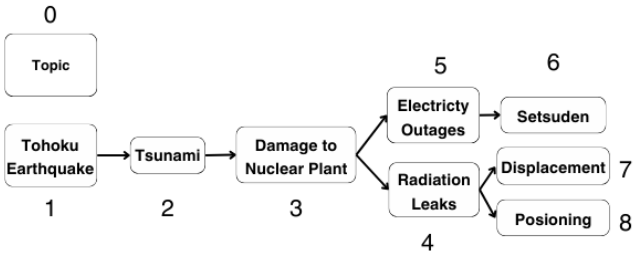

Figure 1: Narrative content and causal model communicated in the nuclear disaster narrative. Causal events are labeled from 1-8, with 0 representing a broad topic description of narrative. This diagram is just for illustration purposes, participants did not see this diagram and narrative. Rather, participants read a four-paragraph narrative.

Experimental Design and Procedure

We used the open-source framework OTree written in Python (Chen et al., 2016) and hosted experiments on a Linux server. Participants joined the experiment through a Qualtrics survey that directed participants to the network experiment.

Procedure As shown in Figure 2, the experiment consisted of a preinteraction block and a networked interaction block. In the **preinteraction block**, all participants read the four-paragraph narrative describing the Fukushima nuclear disaster and then were asked to write a personal narrative (within a 140-character limit) and ten hashtags characterizing the narrative. Participants in the *effect-prompting condition* were asked to list the five effects of the disaster before entering the subsequent network interaction block. Participants in the control condition were not asked to perform any additional task before the network interaction block.

In the **network interaction block**, participants joined a network experiment and engaged in real-time interaction via our online platform. Participants were assigned to one of twelve networks ($N = 20$ or $N = 50$; homogeneously-mixed $\times$ control vs effect-prompting). The networked interaction block consisted of 40 trials, in which participants interacted with their partners. On each trial, participants could be matched with any other participant in the network. They were instructed to write a single hashtag describing the narrative they read in the pre-interaction block. After participants submitted their hashtag response, they were then presented with a new page showing their own hashtag response, their partner's hashtag response, whether they received a point for matching responses with their partner, and their cumulative reward points. Participants were informed of their partner's response *after* submitting their hashtag response (see sample screenshots of an interaction trial in Figure 2).

Experimental Results

We first analyze how the effect-prompting intervention shifted group-level coherence in network communications, as encoded by distribution of hashtag responses. We then apply a narrative alignment measure to analyze the contents in hashtag responses, a novel embedding-based approach to measure which causal events in the disaster narrative each hashtag re-

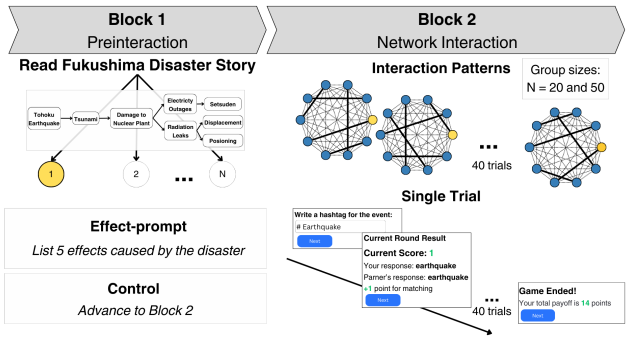

Figure 2: Experiment procedure and networked interaction tasks. We highlight a single node (in yellow) to illustrate a single participant's trajectory through the procedure.

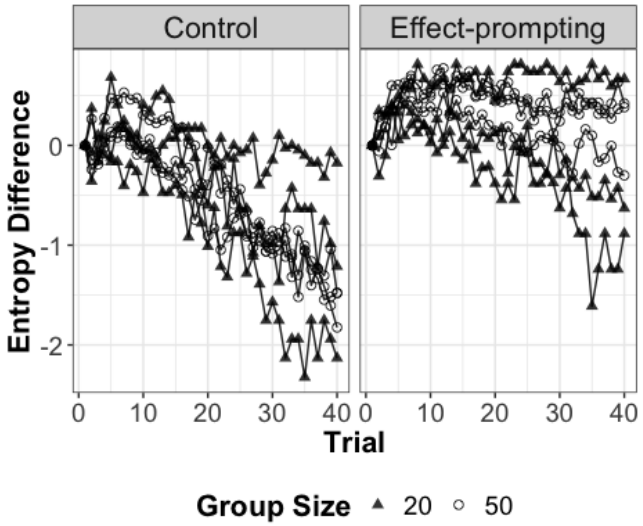

Figure 3: Onset of shared behaviors during communication. Each line shows the outcome of an individual experimental run. Plots show each group's behavioral entropy, represented as the *entropy shift* (Current entropy − First trial entropy) of each group's full-distribution of hashtag responses (lower values = more shared behaviors).

sponse closest aligns with, and assess how the intervention shifted hashtag alignment with causal events in the focal narrative (i.e., the nuclear disaster narrative).

Online communication coheres group behaviors

We first examine response convergence using two metrics: proportion of a group providing a dominant hashtag response, and the entropy of each group's full distribution of hashtag responses. Entropy measures the variability of responses across the entire group: lower entropy indicates more similar or coherent responses in the group, while higher entropy suggests greater diversity or variation in responses (Avolio et al., 2019; Hallett et al., 2016). We fit a Beta-distributed GLM to predict the proportion of a group producing a dominant response as a function of *trial number* interacted with *condition* (control or effect-prompting), while controlling for network Size. Shared behaviors emerged reliably over time in the control networks ($\beta_{Trial} = 0.04$, 95% CI $[0.03, 0.05]$) doing so more slowly, if at all, than in effect-prompted networks ($\beta_{Trial:Condition} = -0.04$, 95% CI $[-0.05, -0.03]$). Next, we fit a Gaussian-distributed GLM to predict the change in the entropy of the full response distribution over the course of online communication. As shown in Figure 3, a group's response entropy steadily decreased as a function of subsequent interactions in control condition ($\beta_{Trial} = -0.04$, 95% CI $[-0.04, -0.04]$), dropping more slowly in effect-prompted condition ($\beta_{Trial:Condition} = 0.03$,95% CI $[0.02, 0.04]$). Hence, the entropy results are consistent with the results from proportion of domination responses, showing more convergence in the control condition than in the effect-prompting condition.

As shown in Figure 3, control groups cohered hashtags more quickly than effect-prompted network. A weakness of this study is not having groups properly matched for group coherence outcomes (which is immensely difficult to do). Our current findings suggest that the effect-prompting intervention effects may interact with how quickly a group integrates background causal knowledge with social reward information to reach consensus. When individuals are in a groups with a clear normative response, social utilities will drown out any attentional affects on the focal narrative's causal structure. We unpack this idea more thoroughly below.

Effect-prompting shifts hashtags to encode effects in the disaster narrative

Next, we applied natural language modeling to examine how the content of the hashtag responses shifted following pre-interaction effect-prompting. This analysis involved computing the cosine similarity of embeddings for user-generated hashtags and *narrative entities*, subsections of the narrative's text encoding discrete events which serve as inputs and outputs of causal relations expressed in the disaster narrative's structure, see Figure 1. First, we initialized a pre-trained SentenceTransformer model ('all-MiniLM-L6-v2') (Wolf, 2020) to compute embeddings for both narrative segments and participant responses. The narrative text was segmented into causal events, and each segment was encoded into a high-dimensional embedding capturing its semantic content. Using cosine similarity, each participant's hashtag response on each trial was matched to the most semantically similar causal event in the narrative.

As shown in Figure 4, we visualize the distribution of hashtags that most closely align with each entity in the focal narrative. Participants in the effect-prompting condition were more likely to express responses that aligned with later causal events – particularly relating to the *Energy outages* and *Setsuden* effects of the disaster – in addition to producing less topic hashtags (e.g., #FukushimaNuclearDisaster) than control groups. It is worth noting the high number of earthquake

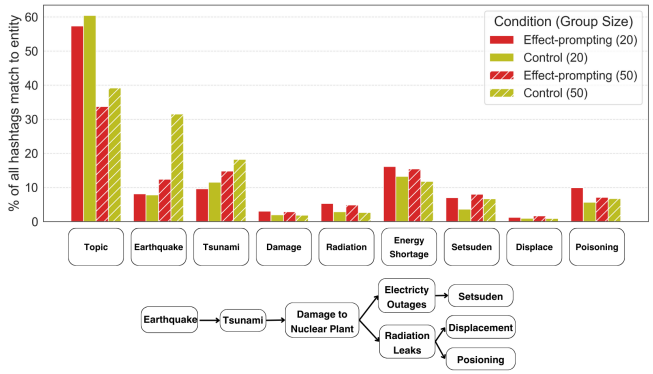

Figure 4: The number of human generated hashtags across all experimental runs that most closely align (i.e., highest cosine similarity) with each narrative entity The y-axis represents the frequency of hashtags most similar the narrative entity, with the x-axis representing each narrative entity. See Figure 1 for diagrams of causal entities.

hashtags in N = 50 control groups is due to a group reaching consensus around that hashtag, which increases the size of the effect (i.e., by pulling control groups towards a causal variable, and away from narrative effects).

Shifting from topic hashtags reduces coordination

We observed that the effect-prompting intervention led to slower convergence rates of consensus and more effect-oriented hashtag responses in experiments. To examine what might result in these outcomes, we fit identical Bernoulli-distributed GLMs to predict when an individual produces a matching hashtag with their neighbor on each trial as a function of the narrative entity a participant's hashtag most closely aligned with, interacting with condition (control vs. effect-prompting), and controlling for cosine similarity between the hashtag and narrative entity. This semantic similarity measure tested whether convergence with focal narrative content facilitates coordination.

Across both experiments, generating topic hashtags (e.g., #Nuclear, #FukushimaNuclearDisaster) showed the highest coordination probability between the network neighbors (see Figure 5). Participants in effect-prompting conditions were significantly less likely to coordinate with neighbors than those in the control condition in $N = 20$ networks $\beta_{Topic:Condition} = -0.80$, 95% CI $[-0.97, -0.62]$ as well as $N = 50$ networks ($\beta_{Topic:Condition} = -1.71$, 95% CI $[-1.99, -1.44]$). In the control condition, coordination on topic hashtags occurred on around 20% of trials for smaller groups and 40% for larger groups; with effect prompting groups about 10% less (refer to the data points on the far left in Figure 5).

When predicting which narrative entities resulted in hashtag coordination, we also included the cosine similarity of each hashtag's embedding with the embedding of the text content of the narrative entity it was matched to. In both

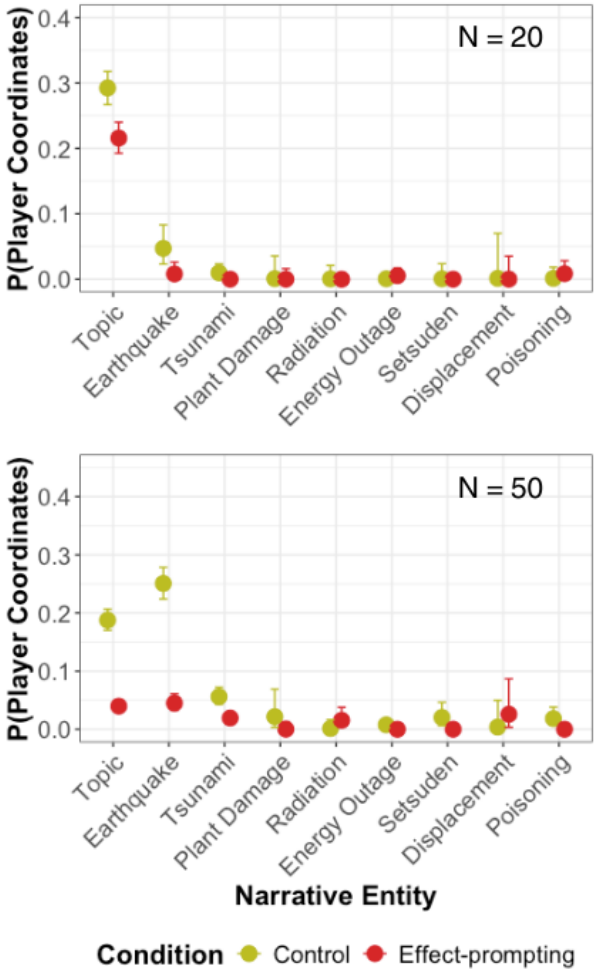

Figure 5: The probability a participant coordinates with their network neighbor when generating a hashtag relating to each narrative entity. Top panel is $N = 20$ groups, bottom is $N = 50$ groups.

groups, we observed an effect of a hashtag's cosine similarity to its matched narrative element. Hashtags more semantically aligned with the content of their matched focal narrative variable were more likely to support coordination ($\beta_{Similarity} = 2.77$, 95% CI $[1.80, 3.76]$ in N = 20; $\beta_{Similarity} = 5.65$, 95% CI $[4.75, 6.59]$ in N = 50). For example, #Tsunami has cosine similarity of 0.6427 to the Tsunami narrative variable while #Waves has a cosine similarity of 0.4247, suggesting that #Tsunami is more likely to produce a match with a neighbor than #Waves.

Model of network communication with LLM-based agents

We compared human performance to a network of interacting LLM models with complete conversation histories (i.e., trial prompts included all interaction history prior to trial) to assess how dynamics in responses account for the variability in hu-

man group-level behavior. However prompts did not include the focal narrative, which suggests that results we present depend on the prompting procedure (see below).

Model Setup: Agent instantiation and interactions

Communication network definition We develop a communication network with N agents A_i connected by an edge-set $E = (e_i, e_j)$. Here, agents are the Llama-3.1-8B-Instruct (Wolf, 2020) language model in a fully-connected, $N = 20$ communication network. Agents interact with randomly paired partners across 40 trials. Networks are fully-connected with uniform edge sampling (no sampling history).On each trial, there is equal probability that any two agents can interact.

Prompting agent's with complete communication histories during network communication On each trial, we present each LLM agent the following prompt, where *current_round* is the current round number, *previous_table* is the complete conversation history from social interactions, and agents in effect prompted networks were also provided the $effect$ narrative variables they generated after reading the focal narrative. Note, that agents are **not** prompted with the focal narrative, so they are not given any situation model/narrative causal structure about the communicated disaster event. The prompt follows:

> In the experiment, you are awarded with **1 point** if you guess the same hashtag as your randomly-assigned neighbor, and **0 points** if you don't guess the same hashtag. Your goal is to earn as many points as possible. You are in **round {`current_round`}** of the experiment. Your guesses, and your neighbor's guesses have been as follows, as represented in the CSV below: {`previous_table`}
>
> Based on this information, the event provided to you in round 1, and the effects observed from round 1: {`effects`}
>
> Please guess a hashtag for this scenario with the goal of matching your randomly-assigned neighbor in this round.

Due to this prompting procedure, each agent's A_i memory of their interaction history was represented by appending trial-level response pairs consisting of the agent's and their partner's response to a growing (ordered) vector of interaction memories M_i (all interaction data for A_i up to trial t). Agents in the effect-prompting condition additionally encoded the five effects they generated at the beginning their interaction memory vector, to replicate the experimental design with humans, who listed five effects of the disaster prior to network communication. When prompted to generate a hashtag they were given M_i as context. Therefore, as interactions progressed agents held a complete and perfect memory store of their interactions, *but with no background causal knowledge encoded in narrative*with distance from narrative effects generated in pre-interaction growing over the course of networked communication.

LLM-based groups don't reach consensus with complete conversation histories

As shown in Figure 6, there was no onset of coherent behavior among communicating LLM agents,revealed by the lack of reduction in entropy of the full response distribution—an opposite trend. This result is clearly opposite from the emergence of group consensus from human responses in both conditions (see Figure 3 for comparison). However, we emphasize this result only applies to the specific prompt instruction used in the simulation experiment. The lack of coherence onset and behavioral similarity across LLMs could be due to complete communication memories resulting in excessive distractions in prompts. LLM agents weren't instructed to pay attention to repeating responses for instance, which could direct agents towards consistency.

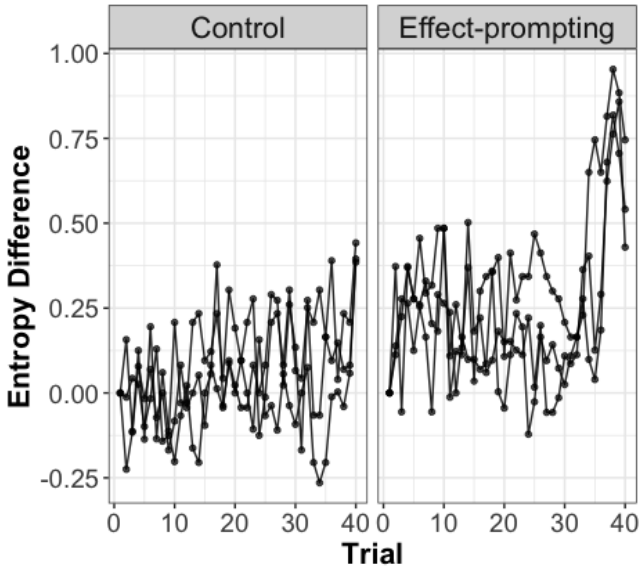

Figure 6: Entropy shift of each LLM group's distribution of hashtags. Systems of LLMs can't reach consensus with only social interaction data (i.e., full interaction histories with partners). Only groups of N = 20 were simulated in this study.

Effect-prompting minimally shifts causal focus of LLM responses

Using the same narrative alignment analysis method applied to human data, Figure 7 shows the distribution of LLM responses across each of the narrative entities (see Figure 1 for more information). For example, effect-prompting drove LLM agents to issue more hashtags close to *Energy/Setsuden* entity, and either no hashtags or dramatically fewer for *Earthquake, Tsunami* (causal factors in the narrative), and *Displacement* entities. We observed a similar trends among human groups, but the effect-prompting shifts with LLMs are

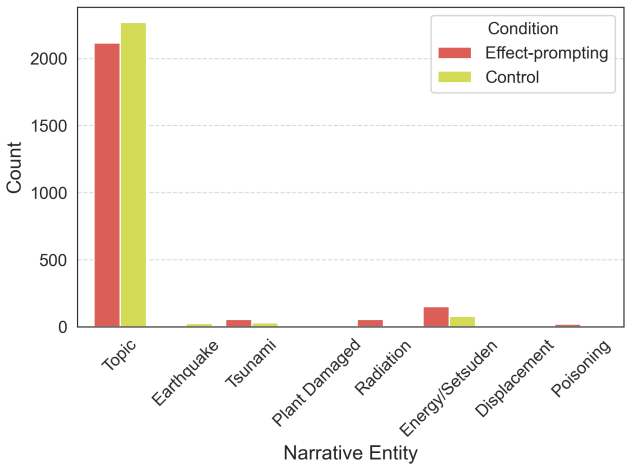

Figure 7: The number of LLM generated hashtags across all simulation runs that most closely align (i.e., highest cosine similarity) with each narrative entity. The y-axis represents the count of hashtags most similar to the narrative entity, with the x-axis representing each narrative entity. See Figure 1 for more information.

much smaller. The results were similar across both conditions for *Plant Damage, Radiation Leaks*, and *Poisoning* entities, with LLM responses hovering around zero for many narrative entity categories.

Discussion

We found that a simple effect-prompting instruction, asking participants to list the effects caused by a disaster, shifted people to produce hashtags in a different way. Specifically, during networked communication, participants generated hashtags that more closely reflected the effect variables within the narrative's causal structure, rather than generic topics or cause variables. However, the magnitude of the shift depends on how quickly groups reach consensus. Groups that quickly coordinate have learned to exploit social rewards for coordinating hashtags; these utility representations will drown out any attentional shift on the prompted narrative's causal structure. At the present, experimental runs are not ideally matched for consensus rates across experimental conditions. Future analysis of the experimental data will better match coherence onset rates across conditions to disentangle the relationship between group coherence onset and prompting effects.

In the current design, pre-interaction shifts in an individual's cognitive framing of the narrative can be easily overtaken by their group's drive towards coherence. To increase their effect size, narrative framing interventions should be designed to leverage social coherence mechanism. For example, rather than simply ask participants to list out effects, the intervention should also target individual reward functions to increase coordination rewards over effect-oriented hashtags. By integrating the current experimental design with the narrative alignment measure we proposed, individuals can be rewarded additional points for coordinating hashtags relating to specific narrative entities (e.g., 1 point for a topic hashtag match; 5 points for a effect entity match). This methodological advance for rewarding narrative alignment will also make network experiments better suited to study the social learning of causal (rather than word level) information from language.

In addition, Further extensions on different narrative materials are necessary to confirm the generalizability of this effect (e.g., to narratives with different causal structures). Future effect-prompts could consider biasing participants to *specific* hashtags (e.g., by decreasing the number of hashtags relating to the narrative's topic), allowing for identical shifts to everybody's priors to better enhance coordination. To increase coordination rates among effect-prompted groups, future interventions should additionally prompt participants to produce semantically precise and *causally parsimonious* hashtags, which maximally encode the causal content of the narrative entity they represent.

Turning now to the network models of LLM agents, in this simulation study we did not provide agents with the text passage narrative during network interaction, which resulted in a large amount of response entropy that exploded across interactions (see effect-prompting panel of Figure 6). Given previous work that reports networks of agents finding agreement, we suspect the inability of our LLM agents to reach agreement is due to the prompt instruction we provided, as agents weren't instructed to pay attention to causal related information in the narrative. Future work will extend prompts following an experimental design approach, by making minimal edits to the prompt and measuring impact on group coherence rates. First edits will encode narrative content into the prompts, which should help refine the space of possible outputs to focus on the narrative frame increasing group-level coherence. Specifically, edits should identify which training procedures, network structures, and prompts are necessary for groups of generative agents to align narrative interaction data.

Once language dynamics stabilize, effect-prompting can be retested on these more complex LLM agents (as defined by the amount of narrative and interaction information contained in the prompt). Structured prompt engineering will likely be necessary to properly mix the LLM agent's background semantic and causal knowledge with information gathered from networked communication. Computational cognitive models in reinforcement learning and human memory can be further applied to increase agent memory sampling efficiency and reasoning during interaction.

Acknowledgments

This work was funded in part by the AFOSR MURI grant No. FA9550-22-1-0380 and the DARPA Army Research Office (ARO), under Contract No. W911NF-21-C-0002.

References

Adams, C., Bozhidarova, M., Chen, J., Gao, A., Liu, Z., Priniski, J. H., Lin, J., Sonthalia, R., Bertozzi, A. L., & Brantingham, P. J. (2022). Knowledge graphs of the qanon twitter network. *2022 IEEE International Conference on Big Data (Big Data)*, 2903–2912.

Avolio, M. L., Carroll, I. T., Collins, S. L., Houseman, G. R., Hallett, L. M., Isbell, F., Koerner, S. E., Komatsu, K. J., Smith, M. D., & Wilcox, K. R. (2019). A comprehensive approach to analyzing community dynamics using rank abundance curves. *Ecosphere*, *10*(10), e02881.

Centola, D. (2015). The social origins of networks and diffusion. *American journal of sociology*, *120*(5), 1295–1338.

Centola, D. (2022). The network science of collective intelligence [Publisher: Elsevier]. *Trends in Cognitive Sciences*, *26*(11), 923–941.

Chang, S., Chaszczewicz, A., Wang, E., Josifovska, M., Pierson, E., & Leskovec, J. (2024). Llms generate structurally realistic social networks but overestimate political homophily. *arXiv preprint arXiv:2408.16629*.

Chen, D. L., Schonger, M., & Wickens, C. (2016). oTree—An open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, *9*, 88–97.

Dawson, P. (2020). Hashtag narrative: Emergent storytelling and affective publics in the digital age. *International Journal of Cultural Studies*, *23*(6), 968–983.

De Marzo, G., Pietronero, L., & Garcia, D. (2023). Emergence of scale-free networks in social interactions among large language models. *arXiv preprint arXiv:2312.06619*.

Graesser, A. C., Singer, M., & Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychological review*, *101*(3), 371.

Hallett, L. M., Jones, S. K., MacDonald, A. A. M., Jones, M. B., Flynn, D. F. B., Ripplinger, J., Slaughter, P., Gries, C., & Collins, S. L. (2016). Codyn: An r package of community dynamics metrics. *Methods in Ecology and Evolution*, *7*(10), 1146–1151.

He, J. K., Wallis, F. P., Gvirtz, A., & Rathje, S. (2024). Artificial intelligence chatbots mimic human collective behaviour. *British Journal of Psychology*.

McAdams, D. P. (2001). The psychology of life stories. *Review of general psychology*, *5*(2), 100–122.

Morrow, D. G., Bower, G. H., & Greenspan, S. L. (1989). Updating situation models during narrative comprehension. *Journal of memory and language*, *28*(3), 292–312.

Newman, M. E. (2001). Clustering and preferential attachment in growing networks. *Physical review E*, *64*(2), 025102.

Nguyen, T. T., Criss, S., Michaels, E. K., Cross, R. I., Michaels, J. S., Dwivedi, P., Huang, D., Hsu, E., Mukhija, K., Nguyen, L. H., Yardi, I., Allen, A. M., Nguyen, Q. C., & Gee, G. C. (2021). Progress and push-back: How the killings of ahmaud arbery, breonna taylor, and george floyd impacted public discourse on race and racism on twitter. *SSM - Population Health*, *15*, 100922.

Ochs, E., & Capps, L. (2009). *Living narrative: Creating lives in everyday storytelling*. Harvard University Press.

Papacharissi, Z. (2015). *Affective publics: Sentiment, technology, and politics*. Oxford University Press.

Papacharissi, Z. (2016). Affective publics and structures of storytelling: Sentiment, events and mediality. *Information, Communication & Society*, *19*(3), 307–324.

Papachristou, M., & Yuan, Y. (2024). Network formation and dynamics among multi-llms. *arXiv preprint arXiv:2402.10659*.

Pennycook, G., McPhetres, J., Zhang, Y., Lu, J. G., & Rand, D. G. (2020). Fighting covid-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention. *Psychological science*, *31*(7), 770–780.

Pennycook, G., & Rand, D. G. (2022). Accuracy prompts are a replicable and generalizable approach for reducing the spread of misinformation. *Nature communications*, *13*(1), 2333.

Polletta, F., Chen, P. C. B., Gardner, B. G., & Motes, A. (2011). The sociology of storytelling. *Annual review of sociology*, *37*(1), 109–130.

Priniski, J. H., & Holyoak, K. J. (2022). A darkening spring: How preexisting distrust shaped COVID-19 skepticism. *PLOS ONE*, *17*(1), e0263191.

Priniski, J. H., Linford, B., Krishna, S., Morstatter, F., Brantingham, J., & Lu, H. (2024). Online network topology shapes personal narratives and hashtag generation. *Proceedings of the Annual Meeting of the Cognitive Science Society*.

Priniski, J. H., McClay, M., & Holyoak, K. J. (2021). Rise of QAnon: A mental model of good and evil stews in an echo chamber. *Proceedings of the Annual Meeting of the Cognitive Science Society*.

Priniski, J. H., Mokhberian, N., Harandizadeh, B., Morstatter, F., Lerman, K., Lu, H., & Brantingham, P. J. (2021). Mapping moral valence of tweets following the killing of George Floyd. *arXiv preprint arXiv:2104.09578*.

Priniski, J. H., Solanki, P., & Horne, Z. (2024). A computational framework for distinguishing motivated reasoning and practical rationality.

Priniski, J. H., Verma, I., & Morstatter, F. (2023). Pipeline for modeling causal beliefs from natural language. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, 436–443.

Radvansky, G. A., & Zacks, R. T. (1991). Mental models and the fan effect. *Journal of experimental psychology: learning, memory, and cognition*, *17*(5), 940.

Thurner, S. (2024). New forms of collaboration between the social and natural sciences could become necessary for understanding rapid collective transitions in social systems. *Perspectives on Psychological Science*, *19*(2), 503–510.

Wolf, T. (2020). Transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Wong, E., Holyoak, K., & Priniski, J. H. (2022). Cognitive and emotional impact of politically-polarized internet memes about climate change. *Proceedings of the Annual Meeting of the Cognitive Science Society*.

Yang, G. (2016). Narrative Agency in Hashtag Activism: The Case of #BlackLivesMatter. *Media and Communication*, *4*(4), 13–17.

Zwaan, R. A., Langston, M. C., & Graesser, A. C. (1995). The construction of situation models in narrative comprehension: An event-indexing model. *Psychological science*, *6*(5), 292–297.

Zwaan, R. A., & Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological bulletin*, *123*(2), 162.