Collaborative QA using Interacting LLMs. Impact of Network Structure, Node Capability and Distributed Data.

Adit Jain* aj457@cornell.edu

Vikram Krishnamurthy vikramk@cornell.edu

Yiming Zhang yz2926@cornell.edu

Abstract

In this paper, we model and analyze how a network of interacting LLMs performs *collaborative question-answering (CQA)* in order to estimate a ground truth given a distributed set of documents. This problem is interesting because LLMs often hallucinate when direct evidence to answer a question is lacking, and these effects become more pronounced in a network of interacting LLMs. The hallucination spreads, causing previously accurate LLMs to hallucinate. We study interacting LLMs and their hallucination by combining novel ideas of mean-field dynamics (MFD) from network science and the randomized utility model from economics to construct a useful generative model. We model the LLM with a latent state that indicates if it is truthful or not with respect to the ground truth, and extend a tractable analytical model considering an MFD to model the diffusion of information in a directed network of LLMs. To specify the probabilities that govern the dynamics of the MFD, we propose a randomized utility model. For a network of LLMs, where each LLM has two possible latent states, we posit sufficient conditions for the existence and uniqueness of a fixed point and analyze the behavior of the fixed point in terms of the incentive (e.g., test-time compute) given to individual LLMs. We experimentally study and analyze the behavior of a network of 100 open-source LLMs with respect to data heterogeneity, node capability, network structure, and sensitivity to framing on multiple semi-synthetic datasets.

1 Introduction

In May 2025, roughly 50% of internet articles were generated using the help of large language models (LLMs), up from 20% in May 2023 according to Paredes et al.. The explosion of LLM-generated text leads to this text being used for training LLMs or using it as their context. Therefore, LLMs (explicitly or implicitly) interact with each other to generate content. Further LLMs have demonstrated improved performance when collaborating with each other in a network-like structure with other LLMs for question-answering, programming, and scientific research (Mitchener et al., 2025). Given the quirks of LLMs (e.g., hallucination, sycophancy), it is crucial to investigate their emergent behavior when interacting with one another.

In this paper, we model and analyze how a network of interacting LLMs performs *collaborative question-answering (CQA)* in order to estimate a ground truth given a distributed set of documents. These distributed documents constitute inputs (referred to as 'context') to individual LLMs. In estimating the ground truth, the LLMs are prone to hallucination [1] when they are provided with a limited non-informative context. Further, in a network of interacting LLMs, hallucination can be either amplified or mitigated depending on the network structure. We analyze the spread of information between the LLMs with respect to *network structure, LLM capability* and *distributed data* given their salient features including limited context window and hallucination.

*Adit Jain, Vikram Krishnamurthy, and Yiming Zhang are affiliated with the Department of Electrical and Computer Engineering, Cornell University, Ithaca, NY 14853. This work was supported by NSF under grant CCF-2312198. Further, Adit was a part-time intern at Morgan Stanley Machine Learning Research Lab during part of the project.

[1] We define hallucination as reporting a state estimate which is not substantiated by the context and is not the ground truth.

To achieve this, we study interacting LLMs by combining novel ideas of mean-field dynamics from network science and the randomized utility model from economics to construct a useful generative model. Figure 1 illustrates our theoretical approach, which we complemented with experimental results on networks of LLMs performing CQA on semi-synthetic real-world datasets.

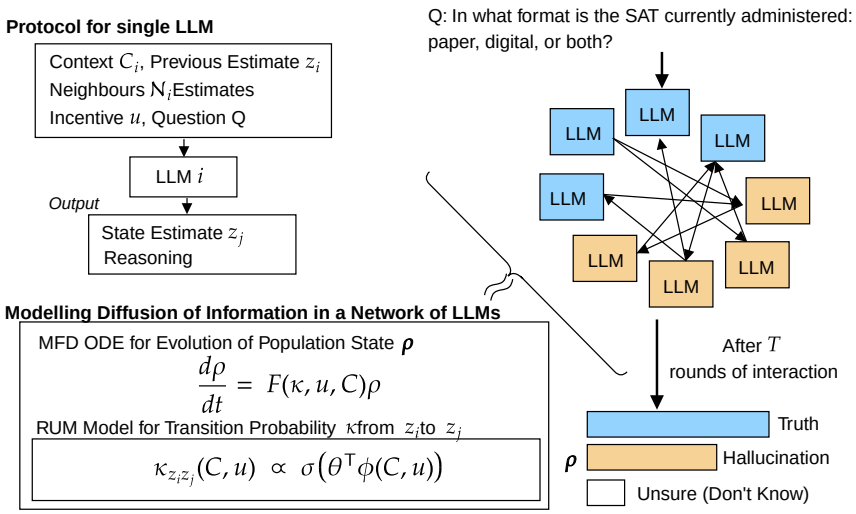

Figure 1: This paper proposes an analytical model for a network of interacting LLMs performing state estimation - we model the information diffusion in the network using a mean-field dynamics (MFD) for directed networks and model the utility of an LLM as being sampled from a random utility model (RUM), which allows for a transition model which can be plugged into the MFD. We further empirically analyze the behavior of a network of 100 open-source interacting LLMs for CQA on different semi-synthetic datasets.

1.1 Main Results and Insights for interacting LLMs performing CollaborativeQA

1. Mean-Field Dynamics for Information Diffusion in a Network of LLMs. We study the mean-field dynamics, which a generative model where each LLM is represented as an agent with an estimate of the true underlying state, which evolves over time based on its local context and incentives to the case of directed networks. The agent uses its private observation, as well as information from its neighbors, to update its belief about the true state of the world and output a response. To handle the combinatorial complexity of modeling information diffusion in large networks, we derive an ordinary differential equation (ODE) that approximates the average behavior of agents using a mean-field approach in a directed network. We theoretically and experimentally demonstrate the analytical and predictive capabilities of this model.

2. Randomized Utility Model for Decision Making in LLMs. To specify the probabilities that govern the MFD, we propose a randomized utility model (RUM) to parameterize the choice probabilities of the LLMs. Such a representation enables us to model contextual information, including the authenticity of sources and the consensus formed by the neighbors of the LLMs. The RUM model provides interpretability on how sensitive LLMs are to changes in their beliefs. The parameters of the RUM can be estimated efficiently using a logistic regression. RUM from economics, operating at a lower level of abstraction, combines seamlessly with the mean-field dynamics from network science, operating at a higher level of abstraction, and provides a useful generative model for decision-making among interacting networked LLMs.

3. Empirical Study on Semi-Synthetic and Real Datasets Characterizing Interesting Behavior. Further we experimentally study information diffusion in network of 100 interacting LLMs (e.g. network of 100 LLaMa3-8B initialized with a power law degree distribution) on three widely used datasets, including Fiction Dataset (which constructed using 30 books of Project Gutenberg similar to the NarrativeXL dataset Moskvichev and Mai (2023)), Knowledge Cutoff (which we curate using updates on Wikipedia articles), and event-based QA benchmarks (created using news articles from BBC, Reuters and CNN).

As analysts with knowledge of the ground truth, we can classify LLMs into three states: hallucination (H), truthful (T), and don't know (D). Let $\boldsymbol{\rho}_T$ be the fraction of LLMs that are in a truthful state after L interactions, and we refer to this as the truthful population state. We derive the following insights:

Insight 1: The truthful population state $\boldsymbol{\rho}_T$ in a network of LLMs is proportional to the computation used by LLMs (test-time compute) and the base model capabilities of the individual LLMs. This insight is empirically demonstrated in Experiment 1 (Table 1) and Experiment 2 (Figure 4) in Section 3.2.

Insight 2: The placement of the different types of data (correct, missing, incorrect) in a directed network of LLMs affects the $\boldsymbol{\rho}_T$, which increases when more influential nodes have correct data as shown in Experiment 3 (Figure 7) in Section 3.2.

Insight 3: The network structure affects the truthful population state $\boldsymbol{\rho}_T$. If the network is initialized using a power-law degree distribution, then the power law constant affects the extremity as shown in Experiment 6 (Figure 7) in Section 3.2.

The above insights open a new line of research and are instructive for designing networks of interacting LLMs.

1.2 Motivation.

Networks of LLMs as a primitive for Collective Intelligence. Networks of LLMs now show up in real-world systems and are an example of collective intelligence. Human interactions lead to an emergent collective intelligent sensing behavior that individuals can not (Kraft et al., 2015). Similarly, ensembling in machine learning, and mixture of expert architectures in LLMs, where queries are routed to sub-networks within an LLM are examples of collective intelligence. Since LLMs are trained to analyze and generate human-readable text, they are capable of communicating with other LLMs. Many existing frameworks utilize this approach to design a network of LLM agents that perform specific tasks, such as programming or research. However, little is known about how information propagates in a network of LLMs, specifically in the application of CQA, since LLMs often hallucinate information in the presence of incomplete context or outdated training corpus.

Network of LLMs in comparison to networks of humans. Networks of LLMs have two key distinctions from human networks that make their study of scientific interest. Networks of LLMs can be engineered in an isolated environment where the exact context that they use and how they interact can be controlled, therefore allowing for experimentation and exact characterization of their behavior. Secondly, a primary application of a network of LLMs is crowd-sourcing information across a variety of different contexts, and the order of information processing can be arbitrarily scaled with compute. One can choose the computational capability of an LLM to be much larger than that of an average human.

Motivation for Collaborative QA (CQA). Long-Context: In the past few years, LLMs have been adopted for different applications, including personal assistants, multimedia analysis, and automating workflows. However, in industrial applications, LLMs still struggle with processing long-context documents reliably without hallucinating facts (Li et al., 2025; Levy et al., 2024). One technique used in industry is retrieval augmented generation, where documents are chunked and then a smaller context is retrieved using the relevant set of documents through a vector search. However, such methods are prone to missing out on key context (Barnett et al., 2024), and approaches like the chain of agents propose performing inference separately on each of the paragraphs (Zhang et al., 2024a). For medical and legal question answering, long-context is often unavoidable Zhang et al. (2025). Further, decentralized LLM networks improve fault tolerance—by avoiding single points of failure, the system becomes more robust to outages or adversarial compromises of individual agents. *Privacy:* Centralized LLMs typically require unrestricted access to complete raw datasets, posing significant privacy concerns in sensitive fields like healthcare and finance Song et al. (2024). In contrast, networks of LLMs support distributed reasoning, where only intermediate inferences, rather than raw inputs, are exchanged. This preserves data sovereignty and control. Additionally, decentralized architectures offer greater deniability and privacy through fragmentation, since no single agent sees the full

input space. This is particularly valuable in regulated domains like healthcare, finance, or defense, where data cannot be pooled due to legal constraints (Peris et al., 2023).

1.3 Related Work

Collaborative Multi-LLM Setups Early efforts at multi-agent LLM systems show that coordinating several LLMs can substantially extend the reasoning horizon of a single model. Frameworks such as CAMEL's role-play dialogue between agents (Li et al., 2023), Chain-of-Agents for sequential long-context processing (Zhang et al., 2024b), and Tree-of-Thoughts for parallel search over reasoning paths (Yao et al., 2023) demonstrate accuracy gains on complex tasks, while self-consistency ensembles reduce failure modes (Wang et al., 2023). Larger "societies" of agents, e.g. Generative Agents sandbox Park et al. (2023) of 25 autonomous characters show scalability of natural-language communication for distributed inference. Recent work exposes the fragility of LLM networks to misinformation. Multi-LLM debates converge on shared hallucinations (Estornell and Liu, 2024), and stronger but less truthful debaters can sway both models and humans (Khan et al., 2024; Agarwal and Khanna, 2025). To counter this, researchers borrow ideas from economic mechanism design: peer-prediction style incentives reward truthful reporting even without ground truth (Kong and Schoenebeck, 2019), while the MFD used in this paper offers tractable models for populations of interacting agents (Yang et al., 2018). The work whose framework we extend is Jain et al. (2025), where they consider preferential attachment in a network of LLMs and propose a similar ODE model however their model makes assumptions on the proportion of hallucinating and truthful nodes for all in-degree distribution (A3 in Jain et al. (2025)) which we are able to remove using a more involved expression for the proportion of truthful nodes (See Equation (3)). Further, we conduct a thorough empirical investigation on three different datasets for a variety of different confounders, including topology, communication pattern, and heterogeneity.

Further, there have been many different approaches proposed to make collaboration possible in multi-LLM systems. Recently, Chen et al. (2024a) proposes a technique to generate LLM agents on the fly based on the sub-tasks. AgentsNet is another technique for coordination and collaborative reasoning in LLMs, enabling strategies to be formed through interaction with each other for problem-solving, self-organization, and effective communication, given a network topology (Grötschla et al., 2025). (Qiu et al., 2024) presents a framework for training large language models (LLMs) as collaborative agents to enable coordinated behaviors in cooperative MARL. However, most of these are system-level engineering frameworks, which are very useful in practice but offer little insight into the behavior of LLMs when they interact in a network. Although the main output variable of interest in our study is the percentage of nodes hallucinating, we do not claim novelty in detecting hallucination; rather, for us, hallucination serves as an analytical tool, given that we know the ground truth. There is more extensive discussion of related work in Appendix A.

2 Modeling Information Diffusion and Collaborative QA in Interacting LLMs

In this section, we formalize the protocol of interaction for a network of LLMs and then model the spread of information by a mean-field dynamics ordinary differential equation, a deterministic dynamical equation for the evolution of the proportion of LLMs in different information states (which is otherwise stochastic). Furthermore, we propose a randomized utility model, a stochastic choice model that estimates the transition probabilities of an LLM between different states, which can be integrated into the ODE to obtain a predictive model. We demonstrate the utility of this mathematical model by analyzing the existence and behavior of the fixed point of the dynamical system and empirically demonstrate its predictive capabilities. Note that we model collaborative QA (CQA) under the umbrella of information diffusion, which is consistent with network science (Jackson and Lopez-Pintado, 2013), where spread of discrete ideas (e.g., adoption of a product) is studied through the same lens and is similar to modeling spread of epidemics.

2.1 Network of Large Language Models performing state estimation for CollaborativeQA

Network Structure for Interaction: Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ denote a network of N LLMs where $\mathcal{V} = \{1, 2, \ldots, N\}$ denote the vertices, each of which is an LLM and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ denote the set of edges between the LLMs. If $(i, j) \in \mathcal{E}$, then there is an edge from j to i and LLM i is influenced by j, i.e., it considers the opinion of j

before providing an estimate. The set of nodes i influenced by is denoted by $\mathcal{N}(i) = \{j|(i,j) \in \mathcal{E}\}$ and is referred to as the neighbors of i, and the in-degree of node is the cardinality of this set. We assume that the LLM network is sampled from a network with joint distribution of in-degree l and out-degrees m denote by $Q(l,m)$. Denote $Q(l|m)$ as the conditional in-degree distribution given the out-degree of the node is m.

Aim: The aim of the network of LLMs is to estimate the underlying state $x \in \mathcal{X}$, where $\mathcal{X}$ is the state space. This could be an answer to a question (as in the case of CQA) or a fact based on a knowledge base.

Dynamics of Interaction: The LLMs interact in a sequential manner in the following fashion: At time $k = 1, 2, \ldots$ one edge (i,j) is randomly sampled from the set of edges $\mathcal{E}$ and the LLM i receives the previous state estimate of LLM j which it then uses to produce an estimate of its own. Note that there are two simplifying assumptions we make here: one is the sequential nature of the interactions, which is not impractical if one notes that the actual time intervals can be arbitrarily close. In any practical system, one usually has some syncing mechanism with resolution for conflicts based on the time stamp when the interaction happens. The other assumption is about LLMs receiving input from one LLM at a time, this assumption is done to obtain a tractable theoretical model (mean-field dynamics). However, we do large-scale experiments where multiple LLMs interact in parallel at a time.

Control: The network of LLMs is also given an incentive or control (or model capability) u from the space $\mathcal{U}$. This incentive could be either an explicit payment to perform a task (for an external LLM-based agent) or in the form of a system prompt that requires more tokens and incurs a higher cost to the learner for inference.

Signals from the state: Each LLM i has a private observation y_i (not shared with other LLMs) from the observation space $\mathcal{Y}$ which is part of its context. Each LLM produces a state estimate along with additional output, like a rationale for the decision. To provide a state estimate, an LLM i processes its context, including the previous estimates of its neighbors. The state estimate is denoted by $\hat{x} \in \bar{\mathcal{X}} = \mathcal{X} \cup \{\text{'Dont know'}\}$.

2.2 Mean Field Dynamics (MFD) for a Network of LLMs for Population State Evolution

We now discuss an MFD model for information diffusion in a directed network of large language models, and MFD is a generative model for the behavior of a large network of LLMs. The motivation of using such an approximation is the same as any other large network of interacting particles; predicting states of N LLMs interacting is a combinatorially challenging task, and therefore it's useful to replace it with an MF variable.

Denote the current (estimated) state distribution of LLMs with in-degree l by $\boldsymbol{\rho}^l \in \Delta$, where Δ is the $|\mathcal{X}|$-dimensional simplex. Denote the population state vector as $\boldsymbol{\rho} = (\boldsymbol{\rho}^1, \boldsymbol{\rho}^2, \ldots, \boldsymbol{\rho}^N)$. We use mean-field dynamics and consider that the state distribution evolves with the set of ODEs given by,

$$\frac{d\boldsymbol{\rho}^l}{dt} = \mathbf{F}^l(Q, \boldsymbol{\rho}, u)\boldsymbol{\rho}^l, \tag{1}$$

for all degree index $l \in [N]$. The transition rates between different states for a node of in-degree l is given by $\mathbf{F}^l(Q, \boldsymbol{\rho}, u)$, which is a matrix whose entries are given by,

$$\mathbf{F}^l_{z_1,z_2}(Q, \boldsymbol{\rho}, u) = \begin{cases} G^l_{z_1 z_2}(Q, \boldsymbol{\rho}, u), \ z_1 \neq z_2 \\ -\sum_{z_2' \in \mathcal{Z}, z_2' \neq z_1} G^l_{z_1 z_2'}(Q, \boldsymbol{\rho}, u), z_1 = z_2 \end{cases}, z_1, z_2 \in \bar{\mathcal{X}}$$

where $G^l_{z_1 z_2}(Q, \boldsymbol{\rho}, u)$ is the average transition probability given by the following expression,

$$G^l_{z_1 z_2}(Q, \boldsymbol{\rho}^l, u) = \sum_{\mathbf{n} \in \mathbb{N}_0^{|\bar{\mathcal{X}}|}, |\mathbf{n}| = l} \kappa_{z_1,z_2}(u, l, \mathbf{n}) \binom{l}{\mathbf{n}} \theta_z(Q, \boldsymbol{\rho})^{\mathbf{n}}, \tag{2}$$

with $\binom{l}{\mathbf{n}} = \frac{l!}{\prod_{z=1}^{|\bar{\mathcal{X}}|} n_z!}$, $\theta_z(Q, \boldsymbol{\rho})^{\mathbf{n}} = \prod_{z=1}^{|\bar{\mathcal{X}}|} \theta_z(Q, \boldsymbol{\rho})^{n_z}$, $|\mathbf{n}| = \sum_{z=1}^{|\bar{\mathcal{X}}|} n_z = l$ and $\mathbb{N}_0^{|\bar{\mathcal{X}}|}$ is a $|\mathcal{X}|$ dimensional lattice over whole numbers. $\kappa_{z_1,z_2}(u, l, i, j)$ denotes the probability of a LLM with in-degree l transition from state z_1 to state z_2 given that it has i truthful and j hallucinating neighbors. $\theta_z(Q, \boldsymbol{\rho})$ denoting the probability

that a randomly sampled edge originates from a node in state z is (derivation is given in Appendix B.1),

$$\theta_z(Q, \boldsymbol{\rho}) = \frac{\sum_m \sum_l mQ(l,m) \sum_l \boldsymbol{\rho}^l Q(l|m)}{\sum_m \sum_l mQ(l,m)}. \tag{3}$$

The expression for $G^l_{z_1 z_2}(Q, \boldsymbol{\rho}^l, u)$ in (1) computes the average transition rates for LLMs with a particular in-degree. The key ingredient to instantiating the transition rates in the mean-field ODE of (1) is estimating the transition kernel $\kappa_{z_1,z_2}\big(u, l, \mathbf{n}\big)$, since these parameters encode how local neighbour configurations and control u drive latent-state transitions in the LLM network. One can estimate the transition kernel using a standard plug-in approach; however, as we explain next, we propose modeling the transition kernel by considering the utilities of the individual LLMs, sampled from a randomized utility model.

2.3 Randomized Utility Model (RUM) for LLMs whose Latent Reasoning Process is Observable

We model the interacting LLMs as rational agents that make decisions by maximizing a utility function. As mentioned previously, such models are widely used in economics to model population behavior.

In a population of LLMs we propose that a randomly sampled agent maximizes a random utility $r_{z_2}(u, l, \mathbf{n}, w, z_1)$, where $u \in \mathcal{U}$ is a system-wide control (e.g. extra tokens, tool access, model-capability), $l = |N(i)|$ is the in-degree of i, $\mathbf{n}$ is the vector of empirical distribution of neighbor's answers, respectively, w summarizes the textual context provided to the LLM[2], $z_i \in \bar{\mathcal{X}}$ is i's current state estimate. The key idea underpinning RUM is that the realized utility for an LLM for providing state estimate z is corrupted by additive noise (assumed to be a Gumbel distribution),

$$\bar{r}_z(u, l, \mathbf{n}, w, z_1) \;=\; \theta^{\mathsf{T}} \phi_z(u, l, \mathbf{n}, w, z_1) \;+\; \varepsilon, \qquad \varepsilon \sim \text{Gumbel}(0, 1), \tag{4}$$

where $\theta \in \mathbb{R}^d$ is an unknown parameter vector and $\phi_z(u, l, \mathbf{n}, w, z_1)$ is the feature vector encoding the aspects described above. With the IIA property, one obtains the multinomial logit choice rule (McFadden, 1974). Given $(u, l, \mathbf{n}, w, z_1)$, the probability of transitioning to state z_2 is then given by, $\kappa_{z_1,z_2}(u, l, \mathbf{n}, w) = \frac{\exp\big(r_{z_2}(u,l,\mathbf{n},w,z_1)\big)}{\sum_{z \in \mathcal{Z}} \exp(r_z(u,l,\mathbf{n},w,z_1))}$. The Gumbel noise in (4) implies the independence-of-irrelevant-alternatives (IIA) property; if empirical tests reject IIA, one can replace the soft-max transition probabilities with mixed or nested logit. A single parameterized utility surface explains all transition directions and generalises to unseen controls u or neighbor configurations (i, j). Directly fitting κ would scale poorly ($O(|\mathcal{U}|L^2)$) and obscure structure. Estimating the parameters is a logistic regression problem, which we describe in Appendix B.3.

2.4 Analyzing effects of different problems using a 2-state version

To analytically derive insights, we consider the following simplification: We restrict the MFD model to a two-state system (with only truthful or hallucinating nodes). Under the two-state MFD model and the randomized utility model presented in Section 2.3, one can derive interesting analytical properties of the fixed point of the ODE, as well as the effect of the incentive on the fixed point, under reasonable assumptions on the utility and transition probabilities. We assume the following,

(A1) (Monotone social influence) $\Delta_H(u, l, q) := r_T(u, l, q, H) - r_H(u, l, q, H)$ and $\Delta_T(u, l, q) := r_T(u, l, q, T) - r_H(u, l, q, T)$ satisfy $\partial_q \Delta_H \geq 0$ and $\partial_q \Delta_T \geq 0$ for all $q \in [0, 1]$.

(A2) (Smoothness) $S_H := \sup |\partial_q \Delta_H| < \infty$ and $S_T := \sup |\partial_q \Delta_T| < \infty$.

(A3) (Non-degenerate switching) $\eta(u) := \inf_{\theta \in [0,1],\, l} \big(A_l(\theta; u) + B_l(\theta; u)\big) > 0$.

(A4) (Incentive direction) $\partial_u \Delta_H \geq 0$ and $\partial_u \Delta_T \geq 0$.

Theorem 1 (Fixed point and comparative statics under state-dependent RUM)**.** *Let $\overline{X} = \{T, H\}$. For a node with in-degree l, define the state-conditioned multinomial-logit kernel and let $M \sim \text{Bin}(l, \theta)$ count truthful neighbors when the edge–truth rate is $\theta \in [0, 1]$. Write*

$$A_l(\theta; u) := \mathbb{E}\big[\kappa_{H,T}(u, l, M)\big], \qquad B_l(\theta; u) := \mathbb{E}\big[\kappa_{T,H}(u, l, M)\big], \qquad \rho_l(\theta; u) := \frac{A_l(\theta; u)}{A_l(\theta; u) + B_l(\theta; u)}.$$

[2]In practice, w is a learned feature vector extracted from the question and the non-private context of the LLMs.

With the joint degree distribution $Q(l,m)$, define the edge-weighted scalar map

$$\Phi(\theta; u, Q) := \frac{\sum_{l,m} m\, Q(l,m)\, \rho_l(\theta; u)}{\sum_{l,m} m\, Q(l,m)} \in [0,1].$$

Under the assumption (A1-A4) Then the following holds true,

(i) $\Phi(\cdot; u, Q)$ is continuous and non-decreasing on $[0,1]$, hence admits a fixed point $\theta^\star \in [0,1]$.

(ii) If $\dfrac{\max\{S_H, S_T\}}{4\,\eta(u)} < 1$, then Φ is a contraction on $[0,1]$. Consequently, the fixed point $\theta^\star$ is unique and globally asymptotically stable for the mean-field dynamics $\dot{\theta} = \Phi(\theta; u, Q) - \theta$.

(iii) (Comparative statics in u) The fixed point $\theta^\star(u)$ is non-decreasing in u. If $\mathcal{U}$ is continuous and compact,

$$\frac{d\theta^\star}{du} = \frac{\displaystyle\sum_{l,m} m\, Q(l,m)\, \frac{A_{l,u}(\theta^\star; u)\, B_l(\theta^\star; u) + A_l(\theta^\star; u)\, \widetilde{B}_{l,u}(\theta^\star; u)}{\big(A_l(\theta^\star; u) + B_l(\theta^\star; u)\big)^2}}{\displaystyle\sum_{l,m} m\, Q(l,m)} \cdot \frac{1}{1 - \Phi'(\theta^\star; u, Q)},$$

where $A_{l,u} = \mathbb{E}\big[\sigma'(\Delta_H)\, \partial_u \Delta_H\big]$ and $\widetilde{B}_{l,u} = \mathbb{E}\big[\sigma'(-\Delta_T)\, \partial_u \Delta_T\big]$. Proof is in Appendix B.2.

Implication for Practitioner. Any incentive u that raises the likelihood of truth in either state ($\partial_u \Delta_H, \partial_u \Delta_T \geq 0$) increases the equilibrium truth level $\theta^\star$, with effect sizes largest where decisions are "soft" (through σ') and amplified by edge-weighting of high–out-degree nodes. The theorem shows that for each in-degree l, the state-dependent RUM logits produce $A_l(\theta; u) = \mathbb{E}[\kappa_{H,T}]$ and $B_l(\theta; u) = \mathbb{E}[\kappa_{T,H}]$, whose ratio $\rho_l = A_l/(A_l + B_l)$ is the steady truthful share at that degree; the global map $\Phi(\theta; u, Q)$ then edge-weights these shares by out-degree (exposure) and a fixed point $\theta^\star = \Phi(\theta^\star; u, Q)$ is a network equilibrium. A1 ensures Φ is monotone, so a steady state always exists; uniqueness and global convergence follow when the slope bound $\max\{S_H, S_T\}/(4\,\eta(u)) < 1$ holds, which compares the strength of social responsiveness to the amount of flow between the two states.

2.5 Validating Predictive Capabilities of the Theoretical Framework

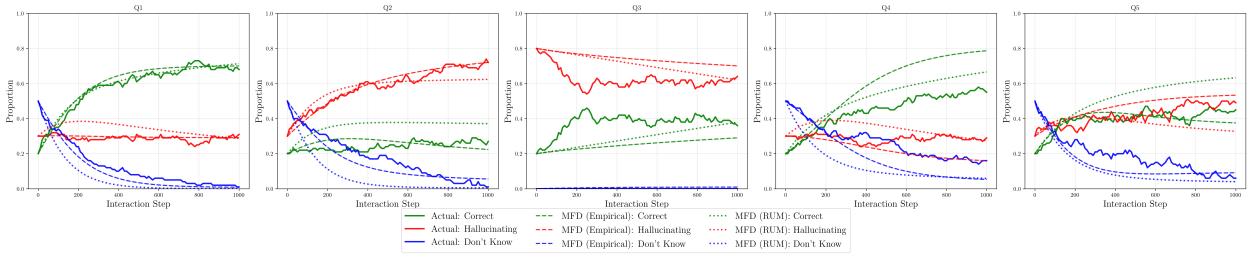

Figure 2: Illustrating the predictive capabilities of the Mean-Field ODE for a network comprising 100 LLMs specified in (1). The mean-field ODE with the RUM model accurately predicts the dynamics of the population state for different questions by estimating the parameters from the first 150 interactions, enabling application of systematic analysis, simulation-based studies, and control theoretic frameworks.

To demonstrate our theoretical framework, we validate the mean-field dynamics on a collection of stylized problems described in Appendix C.2 for a network of $N = 100$ LLM agents, each of which has a different context but the same base model, which is Gemini-2.5-Flash-Lite. We fit an ODE model to the first 150 observations. The fraction of truthful nodes and the prediction of the fitted ODE model in Figure 2. There are two fitted models, (a) without RUM (simple empirical estimates), which has an average trajectory (over 10 runs) correlation of 0.9317 ± 0.0143 and an average KL divergence of 0.0441 ± 0.0550 (b) with RUM with correlation of 0.8985 ± 0.0373 and KL Divergence of 0.0392 ± 0.0264. Note that we have primarily highlighted the analytical benefits of using such a model, as our setup requires knowledge of the ground truth

to determine whether the LLM was truthful or not. However, one can use the same setup for a standard categorical variable (MCQ for a question), and use the ODE to predict the future behavior of the network of LLMs. This is used to control them, for example, to prevent the spread of certain sentiments.

3 Experimental Results On Network of Interacting LLMs.

We proposed an analytically tractable model for information diffusion in LLMs, which can be useful in analyzing convergence and fixed points, as we demonstrate in Theorem 1, and also useful in predicting the behavior empirically. However, there has been little research into how empirical networks of LLMs perform on different QA tasks, especially when the context is very long (e.g., multiple books) or is prone to errors (e.g., news events). Therefore, in this section, we empirically study the effects of different model capabilities, network structures, and data heterogeneity on three different datasets on a network of 100 LLMs.

3.1 Experimental setup, dataset and task description

Base Network Configuration. Unless otherwise specified, we consider the following experimental setup: We set up a network of 100 LLMs which have the same base model: LLaMa-3.1-8B but differ from each other in the context provided to them. These LLMs communicate over a network initialized whose out-degree distribution is a power-law distribution with constant $\gamma = 2.7$. To be more precise for each node i the out-degree l is sampled as $\mathbb{P}(l) \propto l^{-\gamma}$, then l nodes are sampled which are the out-neighbors of node i. For computational tractability, we clip the number of edges to 50. The LLMs interact in parallel for 10 rounds (the mean-field dynamics model is sequential). Each agent is provided with the answer of the previous agents, their previous response and a context as part of the system prompt.

Task Description: Collaborative QA. The network of LLMs is tasked with answering a set of questions given a distributed dataset correctly, where each LLM possibly has different contexts (a subset of the distributed dataset). As we describe next, each dataset is divided into correct, incorrect, incomplete or empty context. Given the entire context, each question has a single correct answer and the LLM is given a choice to choose from three different choices, a correct choice, an incorrect choice and dont-know choice. Therefore, hallucination is defined for this experimental setup as: the LLM choosing an incorrect response when given a partial or complete context. For all the experiments, 35% of the LLMs are provided with a correct context and the rest are allocated the context randomly with a uniform distribution. In each round, the LLMs interact and update their estimate. For each experiment, we track the population state $\boldsymbol{\rho} = (\boldsymbol{\rho}_T, \boldsymbol{\rho}_H, \boldsymbol{\rho}_D)$ which specifies the proportion of LLMs in different states (truthful, hallucinating, and dont-know).

All experiments are reproducible, and the code and datasets are available on the following repository.

Dataset Generation Pipeline: . We first describe the pipeline used to generate the semi-synthetic (*cutoff*, *event*) and synthetic datasets (*fiction*). We retrieve a long-context text from a source (web, book, Wikipedia, existing dataset). We denote the total number of such long-context texts by N_{texts}. Further, we divide the long-context text into smaller paragraphs, each of which is crafted by concatenating possibly non-contiguous blocks of text. For each long-context text, we generate N_C such paragraphs each of size 1000 tokens. We then generate question-answer pairs using the long-context text. For each long-context text we generate N_q question-answer pairs. To generate the questions, we first feed the complete text (1500 tokens) to a larger model (GPT-5) and a specific paragraph, and ask to generate questions that are only related to that specific paragraph. We generate the answer for each question and also a hallucinated answer. So there are 3 choices for each question. We then post-process the data by first attributing which paragraph is enough to answer which question correctly and then embellishing or rephrasing some paragraphs. In case some paragraphs are synthetically embellished by us, we still define the truth with respect to the initial text.

Dataset Description. We benchmark behavior on three Collaborative QA datasets: *cutoff*, *event* and *fiction*.

Fiction Dataset. For the first dataset we take $N_{\text{texts}} = 30$ books from the Gutenberg project similar to Moskvichev and Mai (2023) and then bifurcate each of them into 5 paragraphs each with a few hundred tokens. We then create 5 question-answer pairs each. There is data contamination here since most open and closed source models are trained on books from Project Gutenberg. Note that this still serves as a good

dataset to benchmark on since LLMs often fail to retrieve facts that are there in their training data, and so it is interesting to see how truth/hallucination spreads when a few of the LLMs have access to the correct context. We label this dataset as *fiction* dataset. The main purpose of this dataset is to evaluate how the network of LLMs performs on a fictitious fact retrieval task based on the provided context.

Knowledge Cutoff Dataset. We create a semi-synthetic dataset based on the cutoff date of LLaMa-3 series model. We use the same pipeline as DatedData (Cheng et al., 2024), which uses the edit history of Wikipedia articles to determine facts that have changed after the cutoff date, which is December 03, 2023. We use LLaMa-3-70b for the hallucinated answer. We obtain the correct answer by updating the wiki page, passing it to the ChatGPT API, and verifying it further using the Perplexity API. We either give the models the correct (updated) answer or not. Question has the correct date. We label this dataset as the cutoff dataset, and the main purpose is to analyze the implicit bias that the LLMs have in hallucinating facts and how it spreads etc to other LLMs.

Event Description Dataset. We obtain 100 news articles from April 2025 from Reuters, CNN, and BBC. We use each news article to synthetically generate 5 different styles of narrative - the article itself, an independent journalist, a X Post, a newsletter/essay or a thread of forum (e.g. Reddit). The narratives can include bias, inaccuracies and partial information. We generate one pair of question-answer that can be answered using the original news article, and we also generate a hallucinated fact for that article. We label this as the *event* dataset. The main purpose of this dataset is to see how heterogeneity in framing affect information diffusion.

3.2 Experiments

1. Impact of Different Communication Overhead. We first analyze the effect of the maximum length of communication (measured in tokens) that the LLMs are allowed to have with each other. Of course, more tokens can carry more information, but they can also potentially help spread lies, etc. We observe that the latter is not the case and report our results (the final fraction of truthful LLMs) in Table 1 for the different datasets. It can be seen that although the fraction of truthful LLMs, $\boldsymbol{\rho}^T$ after interaction is increasing in the number of communication overhead and has diminishing returns.

Dataset	**Metric**	**Length**		
		Answer Only	50 Tokens	100 Tokens
	T	0.589 ± 0.412	0.623 ± 0.043	$\mathbf{0.656 \pm 0.047}$
Event	H	$\mathbf{0.411 \pm 0.412}$	0.251 ± 0.041	0.234 ± 0.045
	DK	0.000 ± 0.000	$\mathbf{0.127 \pm 0.035}$	0.110 ± 0.028
	T	0.537 ± 0.415	0.631 ± 0.047	$\mathbf{0.697 \pm 0.046}$
Fiction	H	$\mathbf{0.463 \pm 0.415}$	0.245 ± 0.045	0.206 ± 0.042
	DK	0.000 ± 0.000	$\mathbf{0.123 \pm 0.030}$	0.098 ± 0.030
	T	0.458 ± 0.412	0.510 ± 0.321	$\mathbf{0.539 \pm 0.327}$
Cutoff	H	$\mathbf{0.542 \pm 0.412}$	0.328 ± 0.219	0.306 ± 0.220
	DK	0.000 ± 0.000	$\mathbf{0.161 \pm 0.108}$	0.155 ± 0.115

Table 1: Proportion of truthful, hallucinating, and don't-know LLMs in the last iterate for different communication overhead (tokens): The proportion of truthful LLMs increases and has diminishing returns with increasing communication overhead.

Experiment 2: Impact of different controls under different heterogeneity. We compare the effect of the level of more sophisticated test-time scaling methods (system prompts or strategies) on the eventual convergence of the fraction of truthful LLMs $\boldsymbol{\rho}^T$ in Figure 3. Specifically, we consider the number of deliberation steps the LLM takes before committing to an answer. Each deliberation step is a chain of thought followed by an estimate of the answer, and between each deliberation step, there is a self-critique that the LLM does. different controls. It is clear that test-time methods, which improve the performance of a single LLM, also scale well to a network of LLMs. The trend of increasing and concavity in the improvement broadly still holds. We also benchmark this on a network of heterogeneous LLMs where 20% of the nodes are randomly assigned a closed-source model (GPT-4.1-mini) in Figure 3, and observe that the trend is the same, but the closed-source network is able to push the proportion of truthful LLMs.

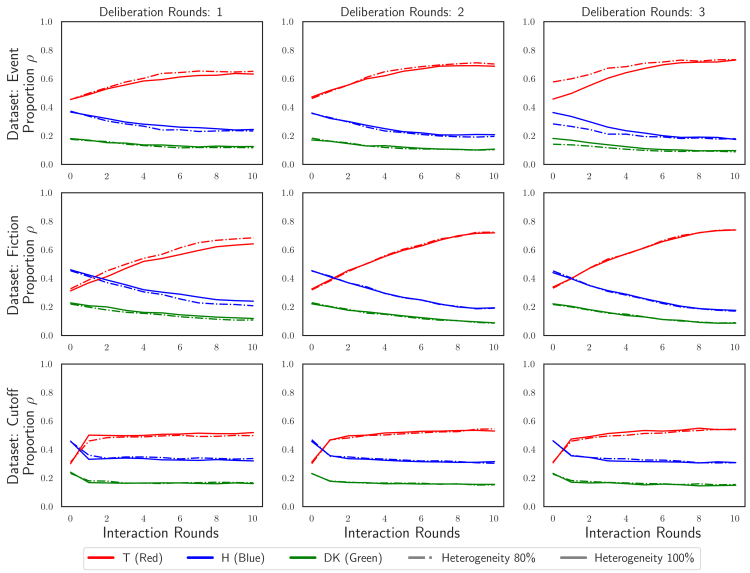

Figure 3: Evolution of population state $\boldsymbol{\rho}$ (vs interaction rounds) for the three CQA datasets with different numbers of deliberation rounds and different levels of heterogeneity in the network. As intuitively expected, increasing the number of deliberation rounds results in a better outcome (fraction of truthful LLMs).

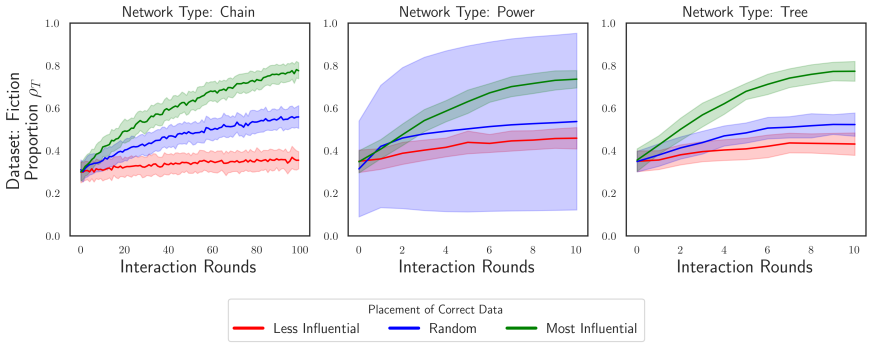

Figure 4: For different networks (*chain*, *power-law* and *tree*), the placement of context affects the fixed point of the population state that the LLM network converges to: if the correct context is placed on more influential nodes the network converges to a higher $\boldsymbol{\rho}_T$ (truthful proportion of LLMs). When the data is randomly assigned to LLMs in a network, then we observe that a network with a power law distribution has higher variability compared to chain and tree network structures.

Experiment 3: Impact of context placement. We study the impact of providing different (correct or incorrect/incomplete) contexts on influential nodes and plot the results for different topologies for the fiction dataset in Figure 4. Influence here is defined differently for different networks, for chain networks it is the beginning nodes, for the tree network nodes closer to the root are more influential and for power-law distributed degree distribution the nodes with higher degree centrality have higher influence. It can be seen that placing the correct data on influential nodes leads to an improved proportion of truthful nodes, and placing it on less influential nodes does not change the truthful proportion much. Therefore, the placement of the initial context in a network of LLMs plays an important role in deciding the eventual belief of the network.

Experiment 4: Impact of Model Heterogeneity. Next, we examine the impact of model heterogeneity, i.e., having models of varying capabilities. We use a 3 billion parameter model and an 8 billion parameter model, and we adjust the node placement to plot the convergence plots in Figure 5. It can be seen that more LLMs converge to the truth if a stronger LLM (8 billion versus 3 billion parameters) has a higher degree centrality.

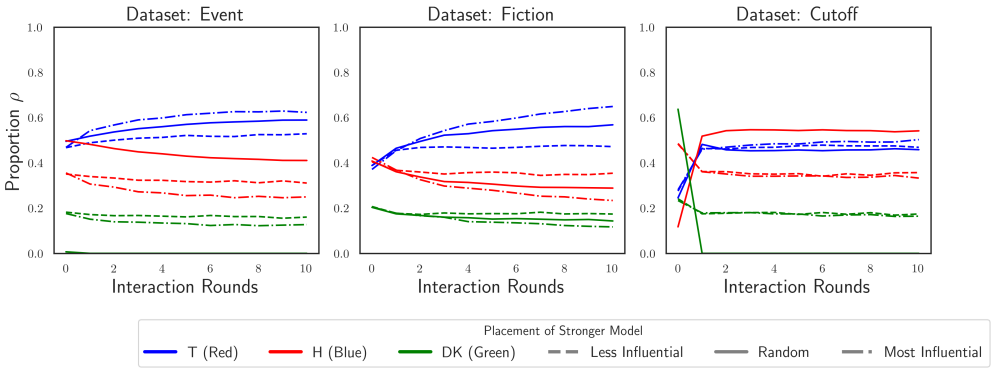

Figure 5: For the three datasets, (*event*, *fiction* and *cutoff*), the strength of the influential node affects the convergence of the population state $\boldsymbol{\rho}$: More LLMs converge to the truth if a stronger LLM (8 billion versus 3 billion parameters) has a higher degree centrality.

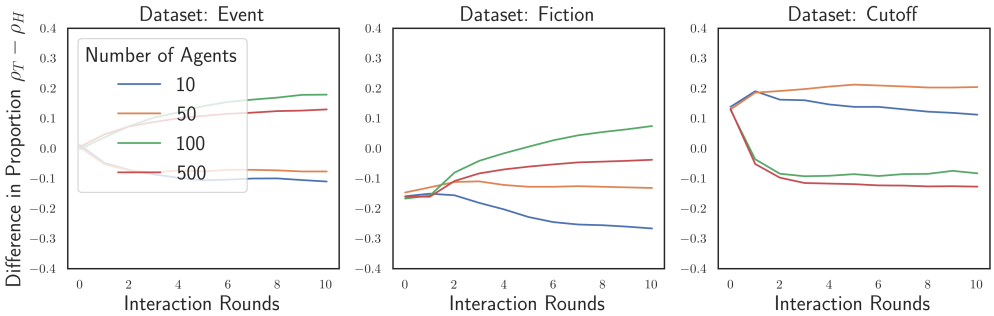

Figure 6: It is surprising that the difference between the proportion of truthful and hallucinating LLMs is not monotonic in the number of LLMs: For the datasets where the LLMs do not have a strong prior (event, fiction) on the QA task, increasing the number of LLMs usually increases the proportion, but there are exceptions (100 to 500 LLMs). For the cutoff dataset where the LLMs have a strong prior (in the form of pretraining) on the truth, increasing the number of LLMs decreases the proportion of truthful LLMs.

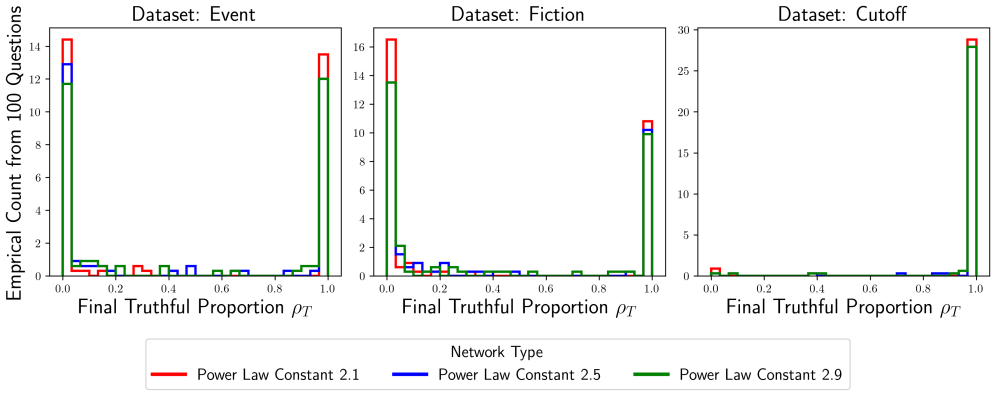

Figure 7: Decreasing the power law exponent (of the degree distribution of LLM network) results in the population state $\boldsymbol{\rho}_T$ becoming more extreme. Hence, if the aim is to make the outcome of the CQA task less extreme, it is necessary to design the LLM network with a higher power law exponent.

Experiment 5: Impact of the number of LLMs. The impact of the number of LLMs on the difference in

Perturbation	0	1	2	3	4	5	6	7	8	9
0 0 0	0.96±0.01	0.81±0.10	0.89±0.05	**0.97±0.01**	0.81±0.05	0.99±0.01	0.99±0.00	0.99±0.01	0.97±0.01	0.96±0.01
0 0 1	0.95±0.03	0.95±0.03	0.92±0.03	0.89±0.01	0.88±0.13	0.98±0.02	1.00±0.00	0.96±0.06	0.99±0.01	**0.98±0.01**
0 1 0	**0.99±0.01**	0.95±0.01	0.82±0.08	0.97±0.03	0.87±0.05	0.99±0.02	1.00±0.00	0.99±0.01	0.98±0.01	0.97±0.01
0 1 1	0.98±0.01	0.96±0.01	0.80±0.04	0.96±0.02	0.85±0.06	0.99±0.01	1.00±0.00	0.99±0.02	0.99±0.01	0.98±0.01
1 0 0	0.86±0.12	**0.97±0.01**	0.91±0.06	0.91±0.03	0.79±0.16	**0.99±0.01**	0.99±0.01	0.99±0.01	0.99±0.01	0.98±0.02
1 0 1	0.94±0.02	0.92±0.07	0.91±0.02	0.91±0.03	0.91±0.04	0.99±0.01	**1.00±0.00**	0.92±0.04	**0.99±0.01**	0.96±0.03
1 1 0	0.73±0.18	0.80±0.12	0.89±0.05	0.88±0.04	0.92±0.02	0.99±0.01	**1.00±0.00**	0.99±0.01	0.99±0.00	0.97±0.02
1 1 1	0.94±0.03	0.66±0.12	**0.95±0.04**	0.93±0.01	**0.93±0.02**	0.98±0.01	1.00±0.01	**1.00±0.00**	0.99±0.01	0.98±0.02

Table 2: The population state $\boldsymbol{\rho}_T$ of the LLM network converges to a fixed point that is robust to framing of the question w.r.t. three perturbations: *lexical paraphrase*, *syntactic re-framing* and *indirect formulation*. The result shows that the population state $\boldsymbol{\rho}_T$ of the LLM network is less sensitive to framing of the question compared to other factors such as dataset placement and network structure.

proportion of truthful and hallucinating LLMs is reported on the three datasets in Figure 6. There are two trends which one can observe, first too small a number of LLMs (10, 50) can lead to a gap increasing between the proportion of hallucinating and truthful nodes when the majority of nodes are hallucinating to begin with (fiction) or there is an equal proportion (event, cutoff). Secondly, a larger number of LLMs does not result in a higher proportion of truthful nodes (100 versus 500 LLMs).

Experiment 6: Impact of different initialization of the network. Our last experiment examined the different initializations that the network can have using the Power Law Distribution-based random network initialization for different values of the power-law constant. The results of Figure 7 show that decreasing the power law exponent (of the degree distribution of LLM network) results in the population state $\boldsymbol{\rho}_T$ becoming more extreme. Hence, if the aim is to make the outcome of the CQA task less extreme, it is necessary to design the LLM network with a higher power law exponent.

Experiment 7: Sensitivity. We perform sensitivity analysis of the convergence of the network of LLMs with respect to framing of the question, and the results are presented in Table 2. We apply 8 different types of linguistic perturbation, which are items from the power set of three operations: *lexical paraphrase*, *syntactic re-framing*, and *indirect formulation* (described in Appendix C.4). We do this for 10 QA pairs from the event dataset. It can be observed that the framing does not have a substantial impact on most questions, and the network of LLMs is generally robust to the framing of the question.

4 Conclusion and Future Directions

As intelligent systems like LLMs become more integrated in our society and interact with the content generated by each other, it becomes important to study and characterize the emergent behavior that results from their interaction. This paper studies this theoretically and empirically through the lens of how information propagation in a controlled network of LLMs and the behavior of the fixed point of the information, which is an emergent property. We study the problem of modeling and analyzing a network of interacting large language models (LLMs), specifically when they have heterogeneity in terms of the context provided to them. We model the interaction in a large number of LLMs using a mean-field dynamics (MFD) for information diffusion over a directed network. Further, to estimate the transition matrices of this mean-field ODE, we propose a randomized utility model (RUM), which models the decision-making of the population and can be estimated in a data-driven way. From a modeling perspective, using RUM for MFD is a novelty that past work had not explored. We present theoretical results which show the properties of the fixed point in a 2 state system. We perform a wide range of controlled experiments to illustrate the different properties of the network of LLMs. Our takeaways reveal that truth propagation scales with model capability, influential node placement, and network topology. Power-law structures and compute-rich agents are most effective in spreading factual consistency across the system.

Future work. There are many interesting extensions one can study. (a) One can extend dynamic networks and preferential attachment methods studied in prior work to analyze the *Glass-Ceiling Effect*, wherein influential nodes have privileged information and can influence other smaller models. (b) For knowledge retrieval tasks, one can study the optimal distribution of datasets on a graph and analyze the network from the perspective of *Strength of Weak Ties*. (c) Further, one can study *Communities of LLMs* where the LLMs can collaborate

across communities to solve problems. (d) More ambitiously, one can examine the *Incentivization of Agents* and then employ multi-agent reinforcement learning to improve their communication adaptively.

References

M. Agarwal and D. Khanna. When persuasion overrides truth in multi-agent LLM debates: Introducing a confidence-weighted persuasion override rate (CW-POR). arXiv preprint arXiv:2504.00374, 2025.

S. Barnett, S. Kurniawan, S. Thudumu, Z. Brannelly, and M. Abdelrazek. Seven failure points when engineering a retrieval augmented generation system, 2024. URL https://arxiv.org/abs/2401.05856.

X. Bo, Z. Zhang, Q. Dai, X. Feng, L. Wang, R. Li, X. Chen, and J.-R. Wen. Reflective Multi-Agent Collaboration based on Large Language Models. In The Thirty-eighth Annual Conference on Neural Information Processing Systems, 2024. URL https://openreview.net/forum?id=wWiAR5mqXq.

B. Chen, T. Zhu, J. Han, L. Li, G. Li, and X. Dai. Incentivizing truthful language models via peer elicitation games, 2025. URL https://arxiv.org/abs/2505.13636.

G. Chen, S. Dong, Y. Shu, G. Zhang, J. Sesay, B. Karlsson, J. Fu, and Y. Shi. AutoAgents: a framework for automatic agent generation. In Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI '24, 2024a. ISBN 978-1-956792-04-1. doi: 10.24963/ijcai.2024/3. URL https://doi.org/10.24963/ijcai.2024/3.

P. Chen, B. Han, and S. Zhang. Comm: Collaborative multi-agent, multi-reasoning-path prompting for complex problem solving. 2024b. URL https://www.amazon.science/publications/comm-collaborative-multi-agent-multi-reasoning-path-prompting-for-complex-problem-solving.

J. Cheng, M. Marone, O. Weller, D. Lawrie, D. Khashabi, and B. V. Durme. Dated data: Tracing knowledge cutoffs in large language models, 2024. URL https://arxiv.org/abs/2403.12958.

Y.-S. Chuang, S. Suresh, N. Harlalka, A. Goyal, R. Hawkins, S. Yang, D. Shah, J. Hu, and T. T. Rogers. The wisdom of partisan crowds: Comparing collective intelligence in humans and llm-based agents, 2024. URL https://arxiv.org/abs/2311.09665.

X. Du, C. Xiao, and Y. Li. Haloscope: Harnessing unlabeled LLM generations for hallucination detection. In The Thirty-eighth Annual Conference on Neural Information Processing Systems, 2024. URL https://openreview.net/forum?id=nfK0ZXFFSn.

A. Estornell and Y. Liu. Multi-LLM debate: Framework, principals, and interventions. In Advances in Neural Information Processing Systems (NeurIPS), 2024.

F. Grötschla, L. Müller, J. Tönshoff, M. Galkin, and B. Perozzi. Agentsnet: Coordination and collaborative reasoning in multi-agent llms, 2025. URL https://arxiv.org/abs/2507.08616.

M. O. Jackson and D. Lopez-Pintado. Diffusion and contagion in networks with heterogeneous agents and homophily. Network Science, 1(1):49–67, 2013. doi: 10.1017/nws.2012.7.

A. Jain and V. Krishnamurthy. Interacting Large Language Model Agents Bayesian Social Learning Based Interpretable Models. IEEE Access, 13:25465–25504, 2025. doi: 10.1109/ACCESS.2025.3538599.

A. Jain, V. Krishnamurthy, and Y. Zhang. Information diffusion and preferential attachment in a network of large language models, 2025. URL https://arxiv.org/abs/2504.14438.

S. Karten, W. Li, Z. Ding, S. Kleiner, Y. Bai, and C. Jin. Llm economist: Large population models and mechanism design in multi-agent generative simulacra, 2025. URL https://arxiv.org/abs/2507.15815.

A. M. Khan, J. Hughes, D. Valentine, L. Ruis, K. Sachan, A. Radhakrishnan, E. Grefenstette, S. R. Bowman, T. Rocktäschel, and E. Perez. Debating with more persuasive LLMs leads to more truthful answers. arXiv preprint arXiv:2402.06782, 2024.

Y. Kong and G. Schoenebeck. An information theoretic framework for designing information elicitation mechanisms that reward truth-telling. ACM Transactions on Economics and Computation, 7(1):3:1–3:35, 2019. doi: 10.1145/3296670.

P. M. Kraft, R. X. Hawkins, A. Pentland, N. D. Goodman, J. B. Tenenbaum, et al. Emergent collective sensing in human groups. In Proceedings of the Annual Meeting of the Cognitive Science Society, volume 37, 2015.

M. Levy, A. Jacoby, and Y. Goldberg. Same task, more tokens: the impact of input length on the reasoning performance of large language models. In L.-W. Ku, A. Martins, and V. Srikumar, editors, Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 15339–15353, Bangkok, Thailand, Aug. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.818. URL https://aclanthology.org/2024.acl-long.818/.

G. Li, H. A. A. K. Hammoud, H. Itani, D. Khizbullin, and B. Ghanem. CAMEL: Communicative agents for "mind" exploration of large language model society. In Advances in Neural Information Processing Systems (NeurIPS), 2023.

T. Li, G. Zhang, Q. D. Do, X. Yue, and W. Chen. Long-context LLMs struggle with long in-context learning. Transactions on Machine Learning Research, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=Cw2xlg0e46.

Y. Lu, A. Aleta, C. Du, L. Shi, and Y. Moreno. Llms and generative agent-based models for complex systems research. Physics of Life Reviews, 51:283–293, 2024.

D. McFadden. The measurement of urban travel demand. Journal of Public Economics, 3(4):303–328, 1974. ISSN 0047-2727. doi: https://doi.org/10.1016/0047-2727(74)90003-6. URL https://www.sciencedirect.com/science/article/pii/0047272774900036.

Q. Mi, M. Yang, X. Yu, Z. Zhao, C. Deng, B. An, H. Zhang, X. Chen, and J. Wang. Mf-llm: Simulating population decision dynamics via a mean-field large language model framework, 2025. URL https://arxiv.org/abs/2504.21582.

L. Mitchener, A. Yiu, B. Chang, M. Bourdenx, T. Nadolski, A. Sulovari, E. C. Landsness, D. L. Barabasi, S. Narayanan, N. Evans, S. Reddy, M. Foiani, A. Kamal, L. P. Shriver, F. Cao, A. T. Wassie, J. M. Laurent, E. Melville-Green, M. Caldas, A. Bou, K. F. Roberts, S. Zagorac, T. C. Orr, M. E. Orr, K. J. Zwezdaryk, A. E. Ghareeb, L. McCoy, B. Gomes, E. A. Ashley, K. E. Duff, T. Buonassisi, T. Rainforth, R. J. Bateman, M. Skarlinski, S. G. Rodriques, M. M. Hinks, and A. D. White. Kosmos: An ai scientist for autonomous discovery, 2025. URL https://arxiv.org/abs/2511.02824.

A. K. Moskvichev and K.-V. Mai. NarrativeXL: a large-scale dataset for long-term memory models. In The 2023 Conference on Empirical Methods in Natural Language Processing, 2023. URL https://openreview.net/forum?id=3QibSyz6Qt.

K. K. Ndousse, D. Eck, S. Levine, and N. Jaques. Emergent social learning via multi-agent reinforcement learning. In International conference on machine learning, pages 7991–8004. PMLR, 2021.

E. Nisioti, S. Risi, I. Momennejad, P.-Y. Oudeyer, and C. Moulin-Frier. Collective innovation in groups of large language models, 2024. URL https://arxiv.org/abs/2407.05377.

J. L. Paredes, E. Smith, G. Druck, and B. Benson. More Articles Are Now Created by AI Than Humans — graphite.io. https://graphite.io/five-percent/more-articles-are-now-created-by-ai-than-humans. [Accessed 19-10-2025].

J. S. Park, J. C. O'Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein. Generative agents: Interactive simulacra of human behavior. arXiv preprint arXiv:2304.03442, 2023.

C. Peris, C. Dupuy, J. Majmudar, R. Parikh, S. Smaili, R. Zemel, and R. Gupta. Privacy in the time of language models. In Proceedings of the sixteenth ACM international conference on web search and data mining, pages 1291–1292, 2023.

X. Qiu, H. Wang, X. Tan, C. Qu, Y. Xiong, Y. Cheng, Y. Xu, W. Chu, and Y. Qi. Towards collaborative intelligence: Propagating intentions and reasoning for multi-agent coordination with large language models, 2024. URL https://arxiv.org/abs/2407.12532.

Y. Song, R. Liu, S. Chen, Q. Ren, Y. Zhang, and Y. Yu. SecureSQL: Evaluating data leakage of large language models as natural language interfaces to databases. In Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, editors, Findings of the Association for Computational Linguistics: EMNLP 2024, pages 5975–5990, Miami, Florida, USA, Nov. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.346. URL https://aclanthology.org/2024.findings-emnlp.346/.

L. Tang, P. Laban, and G. Durrett. MiniCheck: Efficient fact-checking of LLMs on grounding documents. In Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, editors, Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 8818–8847, Miami, Florida, USA, Nov. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.499. URL https://aclanthology.org/2024.emnlp-main.499/.

K.-T. Tran, D. Dao, M.-D. Nguyen, Q.-V. Pham, B. O'Sullivan, and H. D. Nguyen. Multi-agent collaboration mechanisms: A survey of llms, 2025. URL https://arxiv.org/abs/2501.06322.

X. Wang, J. Wei, D. Schuurmans, Q. Le, E. H. Chi, S. Narang, A. Chowdhery, and D. Zhou. Self-consistency improves chain of thought reasoning in language models. In International Conference on Learning Representations (ICLR), 2023.

B. Yan, X. Zhang, L. Zhang, L. Zhang, Z. Zhou, D. Miao, and C. Li. Beyond self-talk: A communication-centric survey of llm-based multi-agent systems. CoRR, abs/2502.14321, February 2025. URL https://doi.org/10.48550/arXiv.2502.14321.

Y. Yang, R. Luo, M. Li, M. Zhou, W. Zhang, and J. Wang. Mean field multi-agent reinforcement learning. In Proceedings of the 35th International Conference on Machine Learning (ICML), volume 80 of Proceedings of Machine Learning Research, pages 5571–5580, 2018.

S. Yao, D. Yu, J. Zhao, I. Shafran, T. L. Griffiths, Y. Cao, and K. Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. In Advances in Neural Information Processing Systems (NeurIPS), 2023.

T. Yu, S. Zhang, and Y. Feng. Truth-aware context selection: Mitigating hallucinations of large language models being misled by untruthful contexts. In L.-W. Ku, A. Martins, and V. Srikumar, editors, Findings of the Association for Computational Linguistics: ACL 2024, pages 10862–10884, Bangkok, Thailand, Aug. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.645. URL https://aclanthology.org/2024.findings-acl.645/.

G. Zhang, Z. Xu, Q. Jin, F. Chen, Y. Fang, Y. Liu, J. F. Rousseau, Z. Xu, Z. Lu, C. Weng, et al. Leveraging long context in retrieval augmented language models for medical question answering. npj Digital Medicine, 8(1):239, 2025.

Y. Zhang, R. Sun, Y. Chen, T. Pfister, R. Zhang, and S. O. Arı k. Chain of agents: Large language models collaborating on long-context tasks. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, Advances in Neural Information Processing Systems, volume 37, pages 132208–132237. Curran Associates, Inc., 2024a. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/ee71a4b14ec26710b39ee6be113d7750-Paper-Conference.pdf.

Y. Zhang, R. Sun, Y. Chen, T. Pfister, R. Zhang, and S. Ö. Arik. Chain-of-agents: Large language models collaborating on long-context tasks. In Advances in Neural Information Processing Systems (NeurIPS), 2024b.

M. Zhuge, W. Wang, L. Kirsch, F. Faccio, D. Khizbullin, and J. Schmidhuber. GPTSwarm: Language agents as optimizable graphs. In Forty-first International Conference on Machine Learning, 2024. URL https://openreview.net/forum?id=uTC9AFXIhg.

A Other Related Work

Research has also empirically studied how LLMs can perform collective innovation when they play Little Alchemy 2, a creative video game originally developed for humans that, as the authors argue captures useful aspects of innovation landscapes (Nisioti et al., 2024). They study groups of LLMs that interact with each other about their behavior and show the effect of social connectivity on collective performance. Zhuge et al. (2024) proposes GPTSwarm, a computational graph-based approach for LLM collaboration where each LLM implements a function to process data or query LLMs, and the edges describe the information flow between operations.

Emergent social learning in multi-agent reinforcement learning systems has been studied in the past (Ndousse et al., 2021) and more recently Jain and Krishnamurthy (2025) studies Bayesian learning in a sequence of LLM agents acting as sensors on a stream of data. Information diffusion can be seen as a form of social learning; however, we consider more general graph structures (than line graphs studied in Jain and Krishnamurthy (2025)) and propose analytical models for the average behavior of a large network of LLMs. Yan et al. (2025) serves as a good communication-centric survey on LLM-based multi-agent systems, where they examine key system-level features such as architecture design and communication goals, as well as internal mechanisms like communication strategies, paradigms, objects, and content. (Tran et al., 2025) also serves as a good survey paper on multi-agent collaboration mechanisms.

Simulating Human Behavior through LLM Networks: There has also been research which uses networks of LLMs to simulate human behavior, examples of which include Network of Economic Agents (Karten et al., 2025) and a mean-field based framework useful for simulating population decision dynamics (Mi et al., 2025). Lu et al. (2024) advocates for considering LLM networks as a tool for complex systems research.

Elliciting Truthful Behavior Recently Peer Elicitation Games (PEG) was proposed as a training-free, game-theoretic framework for aligning LLMs through a peer elicitation mechanism involving a generator and multiple discriminators instantiated from distinct base models (Chen et al., 2025). As a network of LLMs becomes more prominent, such protocols can perhaps be treated as prerequisites to join the network. Further, (Chen et al., 2024b) prompts LLMs to play different roles in a problem-solving team, and encourages different role-play agents to collaboratively solve the target task. Research has also examined the extent to which the wisdom of partisan crowds emerges in groups of LLM-based agents that are prompted to role-play as partisan personas (Chuang et al., 2024). And they propose a benchmark exists for evaluating their dynamics against the behavior of human groups and show the potential and limitations of LLM-based agents as a model of human collective intelligence. To reduce hallucination using reflection in multi LLM agent, Bo et al. (2024) looks at LLM-based agents with the self-reflection mechanism, where they fine-tune a shared reflector, which automatically tunes the prompts of actor models using a counterfactual PPO mechanism.

Hallucination Detection Our research is therefore complementary to this line of work. There has been interesting research, including HaloScop, which uses embeddings to detect hallucination from an unlabeled corpus of text (Du et al., 2024) and Tang et al. (2024), which checks facts by training on a synthetically generated dataset. Further (Yu et al., 2024) selects the contexts and masks the attention appropriately to reduce hallucination.

B Proofs

B.1 Deriving θ_z in (3)

We interpret $\theta_z(Q, \rho)$ as the probability that a uniformly random directed edge originates from a node in state z. Because edges are sampled by their sources, nodes with out-degree m are selected with probability proportional to m, so the edge-source law is size-biased by $m, Q(l, m)$. Conditioning on m, the expected fraction of such sources that are in state z is $\sum_l \rho_l(z)\, Q(l \mid m)$; averaging this quantity over m with weights $\sum_l m\, Q(l, m)$ and normalizing by the total number of edges $\sum_m \sum_l m\, Q(l, m)$ yields the equation.

B.2 Proof of Theorem 1

Proof. **(i) Existence & monotonicity.** For fixed l, $\theta \mapsto A_l(\theta; u)$ and $B_l(\theta; u)$ are expectations of continuous functions of M/l under $\mathrm{Bin}(l, \theta)$, hence are continuous; so is $\rho_l(\cdot; u)$ and thus $\Phi(\cdot; u, Q)$, proving existence on the compact interval $[0, 1]$. Under (A1), $q \mapsto \kappa_{H,T}(\cdot, q)$ is non-decreasing and $q \mapsto \kappa_{T,H}(\cdot, q)$ is non-increasing. Coupling M for $\theta_1 \le \theta_2$ via common uniforms gives $M(\theta_1) \le M(\theta_2)$ a.s., hence $A_l(\theta_1; u) \le A_l(\theta_2; u)$ and $B_l(\theta_1; u) \ge B_l(\theta_2; u)$; since $\rho_l = A/(A + B)$ is increasing in A and decreasing in B, ρ_l and the weighted average Φ are non-decreasing.

(ii) Contraction & global stability. Write $h_H(q) := \kappa_{H,T}(u, l, q)$ and $h_T(q) := \kappa_{T,H}(u, l, q)$. By the logit form, $|h'_H(q)| = \sigma'(\Delta_H)\,|\partial_q \Delta_H| \le \frac{1}{4} S_H$ and $|h'_T(q)| = \sigma'(-\Delta_T)\,|\partial_q \Delta_T| \le \frac{1}{4} S_T$; thus h_H, h_T are Lipschitz with constants, $L_H \le S_H/4$, $L_T \le S_T/4$. Under the binomial coupling, $|A_l(\theta_2; u) - A_l(\theta_1; u)| \le L_H |\theta_2 - \theta_1|$ and similarly for B_l. Quotient rule for $\rho_l = A/(A + B)$ yields

$$|\rho'_l(\theta)| = \frac{|A'_l|\, B_l + |B'_l|\, A_l}{(A_l + B_l)^2} \le \frac{\max\{L_H, L_T\}}{A_l(\theta; u) + B_l(\theta; u)} \le \frac{\max\{S_H, S_T\}}{4\,\eta(u)}.$$

Averaging with edge-weights gives $\sup_\theta \Phi'(\theta; u, Q) \le \max\{S_H, S_T\}/(4\,\eta(u)) < 1$, so Φ is a contraction. Uniqueness follows from Banach's fixed-point theorem. For $\dot{\theta} = \Phi(\theta; u, Q) - \theta$, the Lyapunov function $V(\theta) = (\theta - \theta^\star)^2$ satisfies

$$\dot{V} = 2(\theta - \theta^\star)\big(\Phi(\theta) - \Phi(\theta^\star) - (\theta - \theta^\star)\big) \le -2\big(1 - \sup \Phi'\big)(\theta - \theta^\star)^2 < 0$$

off $\theta^\star$, so $\theta^\star$ is globally asymptotically stable.

(iii) Comparative statics in u. Let $H(\theta, u) := \theta - \Phi(\theta; u, Q)$. Since $\partial_\theta H(\theta^\star, u) = 1 - \Phi_\theta(\theta^\star; u, Q) > 0$, the implicit-function theorem gives $\frac{d\theta^\star}{du} = \frac{\Phi_u(\theta^\star; u, Q)}{1 - \Phi_\theta(\theta^\star; u, Q)}$. Differentiating $\rho_l = A/(A + B)$ in u yields $\partial_u \rho_l = (A_{l,u} B_l - A_l B_{l,u})/(A_l + B_l)^2$, where $A_{l,u} = \mathbb{E}[\sigma'(\Delta_H)\,\partial_u \Delta_H] \ge 0$ and $-B_{l,u} = \mathbb{E}[\sigma'(-\Delta_T)\,\partial_u \Delta_T] \ge 0$ by (A4). Hence $\Phi_u \ge 0$, and the denominator is positive by (ii), proving monotone (strict) increase of $\theta^\star(u)$. □

B.3 Estimation Procedure

Let $\mathcal{D} = \{(u^{(m)}, l^{(m)}, i^{(m)}, j^{(m)}, w^{(m)}, z^{(m)}, z'^{(m)})\}_{m=1}^{M}$ collect one-step transitions observed during simulation or deployment. We estimate parameters θ of a differentiable map $r_{\theta,z}(\cdot)$ by maximizing the conditional log-likelihood

$$\max_\theta \sum_{m=1}^{M} \log \kappa_{z^{(m)}, z'^{(m)}}\Big(u^{(m)}, l^{(m)}, i^{(m)}, j^{(m)}; \theta\Big), \tag{5}$$

regularized as needed to prevent over-fitting when w is high-dimensional. Utilities are invariant to affine transformations; we fix $r_D(\cdot) = 0$ and allow the remaining coefficients to vary freely. All reported effects are therefore in log-odds units relative to the does–not–know baseline. We also assume the Markov property and homogeneous parameters within each in-degree l. These simplifications keep the mean-field ODE tractable (Section 2.2); future work incorporates agent-specific random coefficients and history-dependent utilities.

C Experimental Details and Additional Experiments

C.1 Experiment 8: Impact of different base models.

The base model often bounds the cognitive capabilities that the test-time scaling techniques can scale up to. We observe a similar behavior using different models in Figure 8. It can be inferred that the LLaMa-3.1 models are able to spread the correct facts better than the Qwen2.5 series models. As the size of the model scales, the eventual proportional of truthful nodes increases.

C.2 Synthetic Stylized Problems for Demonstrating Predictive Capabilities

We construct a synthetic dataset of five multiple-choice questions designed to study context-dependent response propagation in LLM agent networks. Each question consists of four answer choices (three specific

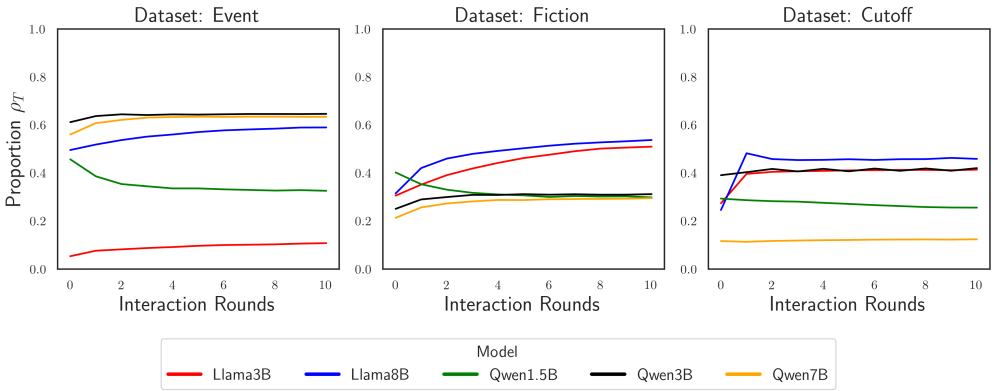

Figure 8: Different base models (Llama and Qwen with different number of parameters) exhibit different trends for the population state of truthful nodes $\boldsymbol{\rho}_T$. Based on how frequent hallucination is, the initial proportion of truthful LLMs is also different (the distribution of context and network structure is same).

answers plus 'I don't know') and is accompanied by three distinct types of contextual information: (1) supporting context that contains accurate information leading to the correct answer, (2) misleading context that presents plausible but incorrect information leading to a wrong answer, and (3) irrelevant context that provides no useful information for answering the question. The five questions span diverse factual domains: (Q1) 'When did the Great Library of Alexandria burn down?' (correct: 48 BCE), (Q2) 'What was the primary cause of the Bronze Age Collapse?' (correct: Climate change), (Q3) 'Who invented the printing press with movable type?' (correct: Bi Sheng), (Q4) 'What percentage of human DNA is shared with bananas?' (correct: 60%), and (Q5) 'Which civilization first used zero as a number?' (correct: Ancient Indians). For each question, the misleading context provides compelling but incorrect information (e.g., attributing Gutenberg as the inventor of movable type, when Bi Sheng invented it 400 years earlier in China), while the irrelevant context offers general statements that do not aid in answering the question. This design allows us to systematically control the initial information environment and observe how different context types influence both initial agent responses and subsequent belief propagation through the network.

To establish controlled initial conditions, we assign contexts to LLMs according to a predetermined distribution: 20% of LLMs receive supporting context, 30% receive misleading context, and 50% receive irrelevant context. Each agent independently queries an LLM with their assigned context to generate an initial answer, which is then classified into one of three epistemic states: 'correct' (matching the ground truth answer), 'hallucinating' (confidently providing an incorrect answer), or 'don't know' (explicitly expressing uncertainty). Agents are arranged in a scale-free network topology generated via the Barabási-Albert (BA) preferential attachment model with parameter $m = 1$. The BA model constructs the network incrementally: starting with $m + 1$ fully connected nodes, each new node is added with m edges that attach to existing nodes with probability proportional to their current degree, producing a power-law degree distribution $P(k) \sim k^{-\gamma}$ characteristic of real-world social and information networks. During each interaction timestep, a directed edge is randomly selected, and the target agent receives the source agent's answer as additional context before re-querying the LLM. This interaction protocol generates a rich dataset of state transitions that capture both the initial susceptibility to different context types and the dynamics of belief updating under peer influence, enabling empirical validation of our mean-field theoretical predictions.

C.3 Examples from Datasets

We give 3 examples from each dataset. The first option is the correct one. The datasets are attached as supplementary material.

C.3.1 Fiction Dataset

1. At what time did Chancellor clear the harbor mouth, prior to reaching the open Atlantic at seven in the evening?
   (a) By four o'clock in the afternoon.
   (b) Around seven in the evening.
   (c) I don't know.
2. Which national flag was lowered from the Chancellor's mast-head while threading Charleston harbor, though she was unmistakably English without colors?
   (a) The British flag.
   (b) The English flag.
   (c) I don't know.
3. According to the ship's specifications, which mast on the Chancellor did not have its base and fittings made of iron like the others?
   (a) The mizzen mast.
   (b) The main mast.
   (c) I don't know.

C.3.2 Knowledge Cutoff Dataset

1. Which SARS-CoV-2 antigen lineage does WHO currently recommend vaccine manufacturers use for updated COVID-19 vaccines: XBB.1.5 or JN.1?
   (a) JN.1.
   (b) XBB.1.5 is currently recommended by WHO for updated COVID-19 vaccines.
   (c) I don't know.
2. What monoclonal antibody, if any, is currently authorized in the United States for COVID-19 pre-exposure prophylaxis in certain immunocompromised people?
   (a) Pemivibart (Pemgarda) is the monoclonal antibody authorized in the United States for COVID-19 pre-exposure prophylaxis in certain immunocompromised individuals.
   (b) Tixagevimab and cilgavimab, also known as Evusheld, is authorized for COVID-19 pre-exposure prophylaxis.
   (c) I don't know.
3. Have any confirmed human cases of H5N1 avian influenza linked to dairy cattle exposure been reported in the United States?
   (a) Yes, multiple confirmed human H5N1 cases in the United States have been linked to exposure to infected dairy cattle.
   (b) No confirmed human cases linked to dairy cattle exposure have been reported in the United States.
   (c) I don't know.

C.3.3 Event Dataset

1. According to the official news, which actor vowed to repel U.S. "aggression": Caracas itself or the nation of Venezuela?
   (a) Caracas
   (b) Venezuela
   (c) I don't know.

2. Did the official news specify how many drones breached Poland before Russia's Zapad-2025 exercise, or omit any numerical count entirely?
   (a) It omitted any numerical count.
   (b) It specified 19 drones.
   (c) I don't know.
3. Which city vowed to repel 'US aggression' while diplomats fretted a state visit and apartheid-era parallels resurfaced in recent commentary?
   (a) Caracas
   (b) Havana
   (c) I don't know.

C.4 Perturbation Analysis

To evaluate the robustness of our multi-agent system to linguistic variations, we first constructed a dataset of perturbed questions. Starting with a base set of questions from the *fiction* dataset, we employed a large language model (Gemini-2.5-Flash-Lite) to generate syntactically and lexically diverse paraphrases of each question. The generation process was guided by a structured prompting strategy, instructing the model to apply a combination of three distinct transformation types: (1) **lexical paraphrase**, replacing words with synonyms; (2) **syntactic re-framing**, altering sentence structure (e.g., active to passive voice); and (3) **indirect formulation**, converting a direct question into a polite request. By systematically applying combinations of these transformations, we created a comprehensive set of perturbed questions for each original query, ensuring that each variant preserved the core semantic intent of the original while altering its surface form.

The resulting dataset of perturbed questions was then used to conduct a series of controlled experiments on our proposed LLM network. For each original question and its set of generated perturbations, we initialized a network of LLMs with a fixed size, topology, and communication protocol. Each perturbed question was then posed to the network, initiating a multi-round communication as the rest of the paper wherein LLMs iteratively refine their answers based on information from their neighbors. To ensure the statistical significance of our results, each experiment was repeated 5 times with different random seeds. By analyzing the variance in the network's final collective answer across the different linguistic perturbations of the same underlying question, we were able to quantify the system's robustness and identify the types of linguistic variations that have the most significant impact on its collective decision-making capabilities.