NON-BAYESIAN SOCIAL LEARNING FOR MODELING INTERACTING LARGE LANGUAGE MODEL AGENTS

Zoya Firsova[⋆] *Malek Khammassi*[⋆] *Mert Kayaalp*[⋆] *Ali H. Sayed*[⋆]

[⋆] Institute of Electrical and Micro Engineering, EPFL

ABSTRACT

Non-Bayesian social learning (NBSL) is a framework in distributed inference that describes how agents in a network combine local observations and their neighbors' beliefs to infer a hidden state. While the framework has been extensively analyzed in theory, its role as a model of belief formation in realistic multi-agent systems remains under-explored. In this work, we investigate whether NBSL can capture the belief dynamics of interacting large language model (LLM) agents. We design controlled experiments in which LLM agents revise their beliefs sequentially based on local evidence and exchanges with their neighbors. Then, we compare their belief trajectories to those predicted by NBSL. Our results provide the first empirical evaluations of NBSL on modern AI collectives, and show that LLM networks can exhibit belief evolution patterns that closely follow NBSL dynamics.

***Index Terms*—** Non-Bayesian social learning, multi-agent systems, LLM agents, opinion dynamics, information aggregation.

1. INTRODUCTION

Non-Bayesian social learning (NBSL)[1–3] models how agents in a network collectively perform global inference through local interactions. It applies in settings where each agent repeatedly observes private signals about an unknown phenomenon and exchanges information with its neighbors to improve its *belief* about that phenomenon.

The framework of NBSL has been examined from two complementary perspectives. The first perspective is engineering and control, where NBSL is studied as a distributed inference algorithm in the context of multi-agent decisions making: the agents can be nodes equipped with sensors, and they collaboratively estimate a hidden state by combining local measurements with information received from neighboring sensors. The second perspective is economics, where NBSL is seen as a behavioral model of belief dynamics, intended to capture how individuals update their views when exposed to private information and the opinions of others.

Over the past two decades, this dual perspective has motivated an extensive theoretical literature on NBSL. Under mild assumptions, results in [1–8] established that agents identify the true state almost surely, with well-characterized convergence rates and error decay. Consequently, NBSL stands as a theoretically mature framework for distributed inference and belief formation over networks.

However, an important gap remains. NBSL can be tested for control and engineering applications, for instance, in sensor networks where agents' observations directly correspond to sensor measurements. In contrast, its interpretation as a model of belief formation has not been sufficiently validated on empirical data. This is because NBSL is a model of how private information and interactions with neighbors shape one's belief about the truth, and the main difficulty about validating such model lies in observability: agents' private information (e.g., what an individual is exposed to) cannot be tracked, and the likelihood models that govern belief updates (e.g., an individual's cognitive processes) cannot be specified or measured. These limitations have so far made the direct empirical validation of NBSL as a model of opinion dynamics challenging.

Recent developments in multi-agent large language models (LLMs) offer an opportunity to bridge this gap. LLMs are now increasingly deployed as autonomous agents that interact with each other and humans in open-ended environments, giving rise to behaviors that surpass simple question-answer exchanges. Much of this work has focused on designing collaborative protocols—such as debate, voting, or role specialization—to improve the reasoning and factual accuracy of LLM models [9–12]. On the other hand, multi-agent LLM environments are increasingly attracting attention as proxies for human societies, providing a way to explore phenomena such as polarization, convention formation, and cooperation [13–15]. A distinctive advantage of these systems is that both their inputs and outputs can be fully observed, in contrast to human or economic networks, where private information and cognitive processes are hidden. This makes multi-agent LLMs an attractive experimental platform for the empirical validation of opinion dynamics models such as NBSL.

The contribution of this paper is to use multi-agent LLM systems as a testbed to empirically validate NBSL as a model of belief dynamics. We design a protocol in which LLM agents receive local observations and exchange their beliefs with neighbors. This environment enables us to assess whether the trajectories of their beliefs align with those predicted by NBSL. In doing so, we aim to connect the mature body of theory from NBSL with the emerging study of LLM collectives.

2. NON-BAYESIAN SOCIAL LEARNING

We start by presenting the NBSL model, which we aim to empirically validate in this work. Consider a set of K agents connected through a directed graph $G = (\mathcal{K}, E)$, where $\mathcal{K} = \{1, \dots, K\}$ denotes the set of agents and $E \subseteq \mathcal{K} \times \mathcal{K}$ the set of edges. An edge $(\ell, k) \in E$ indicates that agent ℓ is a *neighbor* of agent k, and information can flow from ℓ to k. We denote by $\mathcal{N}_k$ the set of neighbors of agent k including k itself. Interactions among agents are governed by a *combination matrix* $A = [a_{\ell k}]$, where $a_{\ell k} > 0$ if $(\ell, k) \in E$ and $a_{\ell k} = 0$ otherwise. Each column of A sums to one, so that agent k forms a convex combination of the information received from its neighbors. The weights $a_{\ell k}$ capture the level of trust that agent k assigns to agent ℓ.

The agents face an unknown state of the world $\theta^\star$ that belongs to a finite set of hypotheses Θ. At each time $i = 1, 2, \dots$, each agent k observes a private signal $\boldsymbol{x}_{k,i}$, drawn independently from the distribution $L_k(\cdot \mid \theta^\star)$. Each agent is equipped with a collection of likelihood models $\{L_k(\cdot \mid \theta)\}_{\theta \in \Theta}$, which specify, for every hypothesis θ, the probability of observing any given signal. This enables each agent to assess how consistent its private observations are with the competing hypotheses in Θ.

At each time i, agent k holds a belief $\boldsymbol{\mu}_{k,i} \in \Delta(\Theta)$, where $\Delta(\Theta)$ denotes the probability simplex over the set of hypotheses Θ. Upon receiving a new signal at time i, each agent k updates its belief in two stages. First, each agent performs a *self-learning step*, where the new observation is incorporated into the belief by means of a local Bayesian update:

$$\boldsymbol{\psi}_{k,i}(\theta) = \frac{L_k(\boldsymbol{x}_{k,i} \mid \theta)\boldsymbol{\mu}_{k,i-1}(\theta)}{\sum_{\theta'} L_k(\boldsymbol{x}_{k,i} \mid \theta')\boldsymbol{\mu}_{k,i-1}(\theta')}, \quad \theta \in \Theta. \tag{1}$$

Second, each agent performs a *combination step* to aggregate beliefs from its neighbors:

$$\boldsymbol{\mu}_{k,i}(\theta) = \frac{\prod_{\ell=1}^{K} \left[\boldsymbol{\psi}_{\ell,i}(\theta)\right]^{a_{\ell k}}}{\sum_{\theta'} \prod_{\ell=1}^{K} \left[\boldsymbol{\psi}_{\ell,i}(\theta')\right]^{a_{\ell k}}}, \quad \theta \in \Theta. \tag{2}$$

Thus, the NBSL model describes belief dynamics in which agents iteratively alternate between a local Bayesian update and a social aggregation step, refining their beliefs over time. Standard convergence guarantees for this model rely on the following assumptions.

Assumption 1 (**Strong Connectivity**). *The graph G is strongly connected, which means that there exists a directed path between any agent pair and there is at least one agent that does not discard its own information.*

This ensures that information from any agent can eventually reach every other agent in the network.

Assumption 2 (**Finiteness of the KL divergence**). *For all $\theta \neq \theta'$ and all $k \in \mathcal{K}$, the Kullback–Leibler (KL) divergence between $L_k(\cdot \mid \theta)$ and $L_k(\cdot \mid \theta')$ is finite.*

This condition ensures that no single observation can fully determine the truth, so agents update their beliefs in a well-defined and incremental way.

Assumption 3 (**Global Identifiability**). *For any $\theta \neq \theta^\star$, there exists at least one agent $k \in \mathcal{K}$ such that the KL divergence between $L_k(\cdot \mid \theta)$ and $L_k(\cdot \mid \theta^\star)$ is positive.*

This condition rules out degenerate cases by requiring that the true state is distinguishable from any false hypothesis by at least one agent.

Assumption 4 (**Positive Initial Beliefs**). *The initial belief of all agents assigns strictly positive probability to every hypothesis in Θ.*

This prevents the true state from being excluded at the start of the belief formation process.

Theorem 1 (**Truth Learning** [3]). *Under Assumptions 1–4, we have*

$$\lim_{i \to \infty} \boldsymbol{\mu}_{k,i}(\theta^\star) = 1 \quad \textit{a.s., for all } k \in \mathcal{K}. \tag{3}$$

That is, as time progresses, the beliefs of all agents concentrate almost surely on the true state $\theta^\star$.

3. EMPIRICAL VALIDATION

In order to empirically validate NBSL as a model of opinion dynamics, we design controlled experiments in which a set of LLM agents interact over a network over multiple rounds. At each round, each LLM agent updates its belief based on two sources of information: (i) private text evidence related to an unknown state and (ii) the current beliefs of their neighbors. Then, we compare the resulting belief trajectories of the LLM agents with the predictions of the NBSL model presented in (1)–(2).

3.1. Experimental Set up

We consider a movie classification task. Movies offer a natural, discrete hypothesis space. Moreover, each movie can be characterized by textual attributes or *tags* (e.g., *crime*, *romance*, *violence*) that capture thematic or stylistic elements. These tags, obtained from datasets linking movies to descriptive terms, serve as partially informative private observations for the agents. Within this formulation, the unknown state $\theta^\star$ corresponds to the identity of a movie, and the agents' objective is to infer this state by combining their private observations with information exchanged with their neighbors.

We conduct two experiments, each defined on a different random network and dataset. Both datasets are selected from the MovieLens collection [16]. In the *first experiment*, we consider $K_1 = 12$ agents connected according to the graph shown in Fig. 1, where all agents have a self-loop that is not depicted. The set of hypotheses is chosen as $\Theta_1 = \{1, 2, 3\}$, corresponding to the movies *Pulp Fiction* ($\theta = 1$), *Forrest Gump* ($\theta = 2$), and *Jurassic Park* ($\theta = 3$), with the true state fixed to $\theta^\star = 1$ (*Pulp Fiction*). Private observations $\{\boldsymbol{x}_{k,i}\}$ for each agent k consist of textual tags from the `ml-20m` dataset [16].

In the *second experiment*, we use a larger network of $K_2 = 23$ agents connected according to the graph shown in Fig. 1, where all agents have a self-loop that is not displayed. The set of hypotheses is chosen as $\Theta_2 = \{1, 2, 4\}$, where $\theta = 4$ corresponds to the movie *Fight Club*, and the true state is again fixed to $\theta^\star = 1$ (*Pulp Fiction*). Here, private observations $\boldsymbol{x}_{k,i}$ for each agent k are drawn from the `ml-32m` dataset [16]. To make the inference task more challenging, we introduce additional noise into these collections in two ways: (i) a subset of multi-word tags is split into single-word tokens, reducing their informativeness, and (ii) with a small probability, agents receive a tag associated with an incorrect movie. As a result, each clue offers only partial and potentially misleading evidence about the true state, forcing agents to rely more heavily on social interactions to refine their beliefs.

3.1.1. Data Preprocessing

The `ml-20m` dataset that we use for the first experiment provides approximately 1100 unique tags for the three selected movies in Θ_1. Each tag and movie pair is given a relevance score that quantifies how well the tag fits the movie. For our experiment, we retain only tags with a relevance score greater than or equal to 0.145, thereby filtering some of the weak associations. This results in 691 tags for the three movies, which are then grouped into $K_1 = 12$ groups, each group serving as the pool of private observations for one of the agents.

The `ml-32m` dataset provides approximately 2058 unique tags for the three selected movies in Θ_2. Unlike the `ml-20m` dataset, this dataset does not assign relevance scores for each movie and tag; instead, each movie is associated with a fixed collection of tags. For the experiment, we partition these tags into $K_2 = 23$ random

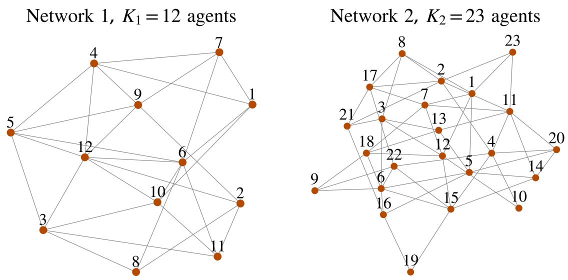

Fig. 1: Network structures for the first and second experiments, respectively. Self-loops are omitted for visual clarity.

groups, one per agent. To further increase the difficulty of the inference task, we take a subset of multi-word tags associated with the true movie $\theta^\star$, split them into single-word tokens, and randomly select 140 of these tokens to augment the tag collection of $\theta^\star$.

3.1.2. LLM Belief Evolution

We elicit belief trajectories from the LLM agents using the protocol in Algorithm 1. At round $i = 0$, each agent is prompted to set its belief distribution uniformly over the three candidate movies, which we record as the initial belief vector $\boldsymbol{\mu}^{\text{LLM}}_{k,0}$. For each subsequent round $i \geq 1$, each agent k is prompted twice. Each prompt provides the agent with specific information and instructs it to output a probability distribution over the candidate movies, which we then record as a belief vector. In the first prompt, agent k is given its belief vector from the previous round $\boldsymbol{\mu}^{\text{LLM}}_{k,i-1}$ together with a new private tag $\boldsymbol{x}_{k,i}$, randomly sampled at test time from the pool of tags assigned to that agent. The LLM is instructed to update its belief based on this information, and its output is recorded as the intermediate belief $\boldsymbol{\psi}^{\text{LLM}}_{k,i}$. In the second prompt, agent k is provided with its intermediate belief $\boldsymbol{\psi}^{\text{LLM}}_{k,i}$ along with the set of neighbor beliefs $\{\boldsymbol{\psi}^{\text{LLM}}_{\ell,i}\}_{\ell\in\mathcal{N}_k}$, all obtained from the first prompt. The LLM is instructed to revise its belief in light of both its own and its neighbors' beliefs, and the resulting output is recorded as the updated belief $\boldsymbol{\mu}^{\text{LLM}}_{k,i}$.

Repeating this procedure for $i = 1, \ldots, T = 15$ allows us to record a sequence of beliefs for each agent, $\{\boldsymbol{\psi}^{\text{LLM}}_{k,i}\}_{i=1}^{T}$ for $k \in \mathcal{K}$. To account for stochasticity in the LLM's outputs, each experiment is repeated $L = 10$ times and the reported trajectories correspond to the average across simulation runs.

Algorithm 1: LLM Protocol

Input: Number of iterations T, initial belief $\boldsymbol{\mu}^{\text{LLM}}_{k,0}$, neighbors $\mathcal{N}_k$ for every $k \in \mathcal{K}$
Output: Belief trajectory $\{\boldsymbol{\psi}^{\text{LLM}}_{k,i}\}_{i=1}^{T}$
for $i \leftarrow 1$ **to** T **do**
  `// Self-learning step`
  for $k \leftarrow 1$ **to** K **do**
    $\boldsymbol{\psi}^{\text{LLM}}_{k,i} \leftarrow$ **LLM output when prompted with:** its previous belief $\boldsymbol{\mu}^{\text{LLM}}_{k,i-1}$ and its new tag $\boldsymbol{x}_{k,i}$;
  `// Combination step`
  for $k \leftarrow 1$ **to** K **do**
    $\boldsymbol{\mu}^{\text{LLM}}_{k,i} \leftarrow$ **LLM output when prompted with:** its own and neighbors' beliefs $\{\boldsymbol{\psi}^{\text{LLM}}_{\ell,i}\}_{\ell\in\mathcal{N}_k}$;

3.1.3. NBSL Belief Evolution

We compute the model-predicted belief trajectories using the NBSL update rules (1)–(2). This requires specifying both the likelihood models $\{L_k(\cdot \mid \theta)\}$ and the combination matrix A.

Likelihood approximation. The likelihood $L_k(\boldsymbol{x}_{k,i} \mid \theta)$ represents the probability that the LLM assigns to observing the tag $\boldsymbol{x}_{k,i}$ under hypothesis θ. This quantity required to run NBSL cannot be accessed directly. What we can elicit instead is the posterior $\mathbb{P}[\theta \mid \boldsymbol{x}_{k,i}]$, which represents the probability that the movie is θ given that the tag $\boldsymbol{x}_{k,i}$ was observed. In practice, this posterior can be obtained by prompting the LLM with the tag $\boldsymbol{x}_{k,i}$ and instructing it to output the probabilities assigned to each $\theta \in \Theta$.

Now note that, under a uniform prior $\mathbb{P}[\theta] = 1/|\Theta|$, Bayes' rule gives

$$\mathbb{P}[\theta \mid \boldsymbol{x}_{k,i}] = \frac{L_k(\boldsymbol{x}_{k,i} \mid \theta)\,\mathbb{P}[\theta]}{\sum\limits_{\theta'\in\Theta} L_k(\boldsymbol{x}_{k,i} \mid \theta')\,\mathbb{P}[\theta']} = \frac{L_k(\boldsymbol{x}_{k,i} \mid \theta)}{\sum\limits_{\theta'\in\Theta} L_k(\boldsymbol{x}_{k,i} \mid \theta')}, \tag{4}$$

which implies

$$L_k(\boldsymbol{x}_{k,i} \mid \theta) \;\propto\; \mathbb{P}[\theta \mid \boldsymbol{x}_{k,i}]. \tag{5}$$

This shows that under a uniform prior, the posterior $\mathbb{P}[\theta \mid \boldsymbol{x}_{k,i}]$ is proportional to the likelihood $L_k(\boldsymbol{x}_{k,i} \mid \theta)$. Therefore, we approximate the likelihood of the LLM by the posterior. To enforce a uniform prior, each query is run in a fresh context so the model has no memory of earlier tags, and the LLM is instructed to treat all movies as equally likely before observing $\boldsymbol{x}_{k,i}$. We denote the resulting estimate by $\widehat{L}_k(\boldsymbol{x}_{k,i} \mid \theta) = \mathbb{P}[\theta \mid \boldsymbol{x}_{k,i}]$, and collect these values for all agents $k \in \mathcal{K}$, hypotheses $\theta \in \Theta$, and tags $\boldsymbol{x}_{k,i}$ in the dataset.

Combination matrix. For the aggregation step (2), we adopt equal neighbor weights, i.e., $a_{\ell k} = 1/|\mathcal{N}_k|$ for $\ell \in \mathcal{N}_k$, reflecting the assumption that agents assign equal credibility to all neighbors.

Given the estimated likelihoods and the combination matrix, the NBSL recursion is implemented following the recursions (1)–(2), producing a trajectory $\{\boldsymbol{\psi}^{\text{SL}}_{k,i}\}_{i=1}^{T}$ for each agent k, and for $T = 15$. To account for stochasticity in the LLM's outputs in likelihood estimation, each experiment is repeated $L = 10$ times and the reported trajectories correspond to the average across simulation runs.

3.2. Results

To evaluate the agreement between the belief dynamics of LLM agents and the predictions of the NBSL model, Fig. 2 shows the belief evolution of eight representative agents from the networks in Fig. 1. Two main observations emerge. First, the belief trajectories of the LLM agents (dashed lines) closely follow the predictions of NBSL (continuous lines), indicating that the model accurately captures their dynamics. Second, in both networks the beliefs concentrate on the correct hypothesis, $\theta = 1$ (*Pulp Fiction*). These experiments are designed so that Assumptions 1–4 hold. Therefore, as established by Theorem 1, the agents uncover the true hypothesis $\theta^\star$, further supporting that the behavior of LLM agents is consistent with the theory of NBSL.

Furthermore, we compute the total variation (TV) distance [3] between the beliefs of each LLM agent $\boldsymbol{\psi}^{\text{LLM}}_{k,i}$ and the corresponding NBSL prediction $\boldsymbol{\psi}^{\text{SL}}_{k,i}$:

$$d_{\text{TV}}(k,i) = \tfrac{1}{2}\sum_{\theta\in\Theta} \left|\boldsymbol{\psi}^{\text{LLM}}_{k,i}(\theta) - \boldsymbol{\psi}^{\text{SL}}_{k,i}(\theta)\right|. \tag{6}$$

We then plot in Fig. 3 the average TV distance across agents:

$$\bar{d}_{\text{TV}}(i) = \tfrac{1}{K}\sum_{k=1}^{K} d_{\text{TV}}(k,i), \tag{7}$$

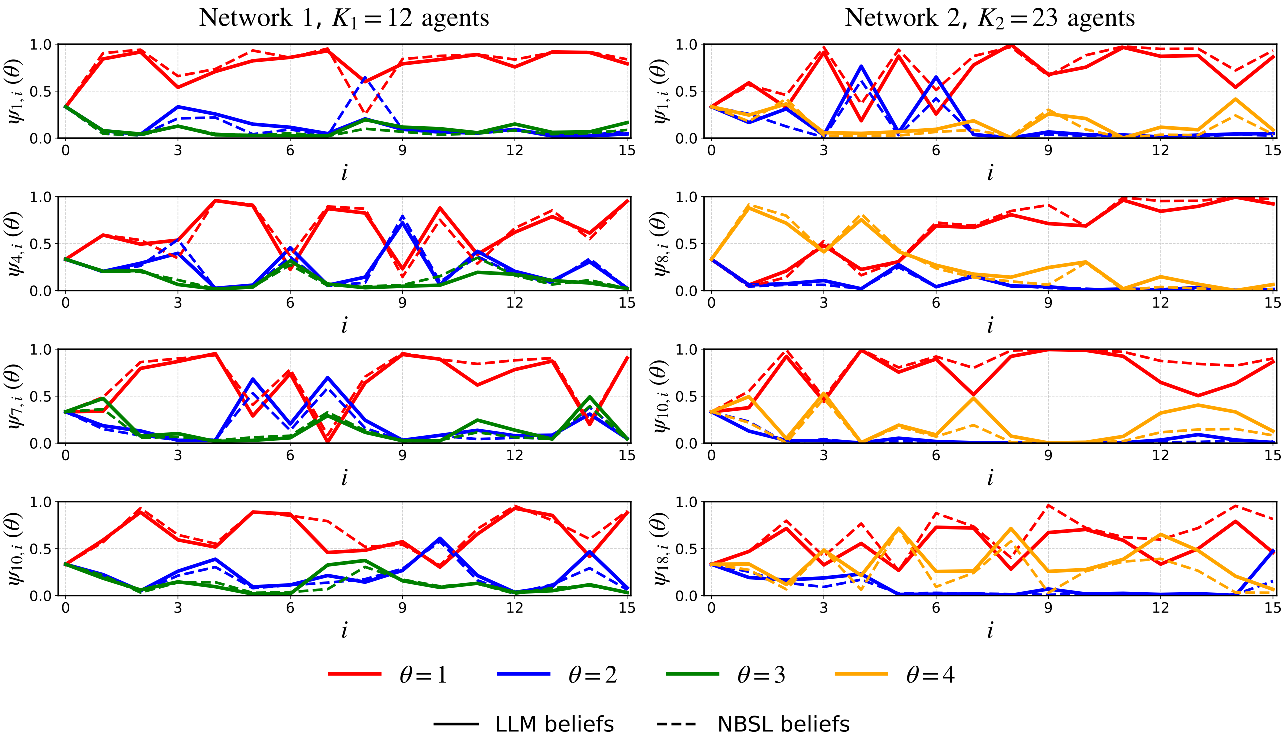

Fig. 2: Belief evolutions for agents 1, 4, 7 and 10 in network 1, and for agents 1, 8, 10, and 18 in network 2.

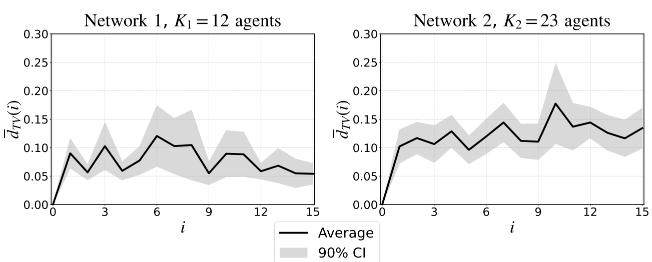

Fig. 3: Total variation evolution of agents over $T = 15$ time steps with confidence interval computed with Student's t-test [17].

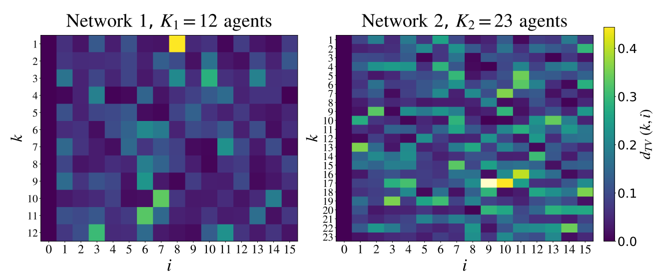

Fig. 4: Total variation evolution for every agent over all $T = 15$ time steps.

together with a shaded area representing the 90% confidence interval (CI) obtained from a Student's t-test [17] over all agents. First, we observe that $\bar{d}_{\text{TV}}(i)$ remains relatively stable over time, indicating that the NBSL model consistently tracks the belief evolution without accumulating error. In addition, the confidence interval is narrow, suggesting that agents exhibit similar behavior across the network.

Finally, to examine the results at a finer granularity, we plot a heat map where rows correspond to each agent k, columns to each time step i, and cell values to the TV distance $d_{\text{TV}}(k, i)$. The heat map confirms that the TV distance remains stable across agents and over time, providing further evidence that NBSL reliably captures the belief dynamics of the LLM agents.

Reproducibility & Artifacts

We use OpenAI's framework and the gpt-oss-120b model with temperature set to 0 and top_p set to 0.5 [18]. The full code, prompt templates, and evaluation scripts are available at: GitHub.

4. CONCLUSION

In this paper, we empirically validate NBSL as a model of belief dynamics in multi-agent LLM environments. By comparing belief trajectories from LLM agents with the predictions of NBSL, we showed that the model reliably tracks the evolution of LLM agents' beliefs across different networks and datasets. In our experiments, tasks were selected at random at test time. Consequently, the LLM agents could not rely on information memorized during pretraining, which minimizes potential bias in the results. Moreover, we assumed that all agents use constant, uniform confidence weights when aggregating information. Exploring how to model biases in LLM behavior and how to capture heterogeneous or adaptive aggregation of information remains an important direction for future research.

References

[1] A. Jadbabaie, P. Molavi, A. Sandroni, and A. Tahbaz-Salehi, "Non-Bayesian social learning," *Games Econ. Behav.*, vol. 76, no. 1, pp. 210–225, Sep 2012.

[2] X. Zhao and A. H. Sayed, "Learning over social networks via diffusion adaptation," in Proc. *Asilomar Conf. Signals, Syst. and Comput.*, 2012, pp. 709–713.

[3] V. Matta, V. Bordignon, and A. H. Sayed, *Social Learning: Opinion Formation and Decision-Making over Graphs*. NOW Book Series on Information and Learning Sciences, 2024.

[4] A. Nedić, A. Olshevsky, and C. A. Uribe, "Fast convergence rates for distributed non-bayesian learning," *IEEE Trans. Autom. Control*, vol. 62, no. 11, pp. 5538–5553, Mar. 2017.

[5] P. Molavi, A. Tahbaz-Salehi, and A. Jadbabaie, "A theory of non-bayesian social learning," *Econometrica*, vol. 86, pp. 445–490, 2018.

[6] A. Lalitha, T. Javidi, and A. D. Sarwate, "Social learning and distributed hypothesis testing," *IEEE Trans. Inf. Theory*, vol. 64, no. 9, pp. 6161–6179, May 2018.

[7] P. Molavi, K. R. Rad, A. Tahbaz-Salehi, and A. Jadbabaie, "On consensus and exponentially fast social learning," in *Proc. Amer. Control Conf. (ACC)*, 2012, pp. 2165–2170.

[8] V. Bordignon, S. Vlaski, V. Matta, and A. H. Sayed, "Learning from heterogeneous data based on social interactions over graphs," *IEEE Trans. Inf. Theory*, vol. 69, no. 5, pp. 3347–3371, May 2023.

[9] Q. Wu, G. Bansal, J. Zhang, Y. Wu, B. Li, E. Zhu, L. Jiang, X. Zhang, S. Zhang, J. Liu, A. H. Awadallah, R. W. White, D. Burger, and C. Wang, "AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation," 2023. [Online]. Available: https://arxiv.org/abs/2308.08155

[10] G. Li, H. A. K. Hammoud, H. Itani, D. Khizbullin, and B. Ghanem, "Camel: Communicative agents for "mind" exploration of large language model society," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2023, pp. 51 991–52 008.

[11] Y. Du, S. Li, A. Torralba, J. B. Tenenbaum, and I. Mordatch, "Improving factuality and reasoning in language models through multi-agent debate," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2024, pp. 11 733–11 763.

[12] A. Jain and V. Krishnamurthy, "Interacting large language model agents. Bayesian social learning based interpretable models," *IEEE Access*, vol. 13, pp. 25 465–25 504, 2025.

[13] J. Park, J. C. O'Brien, C. Cai, M. Morris, C. Liang, J. Liu, and M. Bernstein, "Generative agents: Interactive simulacra of human behavior," in *Proc. ACM Symp. User Interface Softw. Technol. (UIST)*, 2023, pp. 1–22.

[14] M. Ashery, T. Chen, R. Kumar, K. Clark, and J. Andreas, "Emergent social conventions and biases in decentralized llm populations," *Sci. Adv.*, vol. 11, no. 8, p. eabc1234, 2025.

[15] L. P. Argyle, E. C. Busby, N. Fulda, J. R. Gubler, C. Rytting, and D. Wingate, "Out of one, many: Using language models to simulate human samples," *Polit. Anal.*, pp. 337–351, 2023.

[16] F. M. Harper and J. A. Konstan, "The movielens datasets: History and context," *ACM Trans. Interact. Intell. Syst.*, 2015.

[17] E. L. Lehmann and J. P. Romano, *Testing Statistical Hypotheses*, 3rd ed. Springer, 2005.

[18] OpenAI, "Gpt-4 technical report," *arXiv:2303.08774*, 2023.