Emergence of Preferential Attachment and Glass-Ceiling Effects in Autonomous Networks of LLMs

Anonymous authors
Paper under double-blind review

Abstract

We investigate the emergence of structural disparities in networks comprising large language model (LLM) agents. Each LLM agent refers to a prompted LLM of a specified type determined by its base model, model size, and system prompt. When LLM agents autonomously choose collaborators, the resulting communication network exhibits preferential-attachment dynamics: agent types that are already prominent become increasingly likely to attract additional connections. In some cases, weaker LLM agents (agents with smaller base model or older version) can disproportionately occupy central and influential network positions relative to stronger LLM agents. We interpret this misalignment between task capability and network prominence as a type-dependent *glass-ceiling effect* (GCE).

We model the network of LLM agents as a time-evolving sequence of directed weighted graphs, where the vector-valued edge weights represent cumulative tokens exchanged, number of interaction rounds, and reasoning effort. Using a contraction mapping argument on the mean-field dynamics, we prove that the importance (centrality) of each agent type converges to a unique stable equilibrium. To anchor the model in LLM decision mechanisms, we introduce a cross-attention-inspired utility for collaborator selection. This utility specifies the local connection dynamics and, together with the mean-field model, yields a predictive characterization of the limiting network structure and its type-dependent centrality gaps.

To validate the theory, we develop an experimental testbed with 100 LLM agents. Our experiments show that autonomous network formation can generate persistent centrality disparities, with their magnitude and direction depending on model family, model size, system-prompt design, and task context. They further show that the effect of preferential attachment depends on its alignment with model capability: reinforcing it improves collective performance when stronger agents become central, whereas weakening it improves performance when network dynamics instead favor weaker agents. All results are reproducible; the code and datasets are available in an anonymous GitHub repository.

1 Introduction

Multi-agent LLM networks, where each node is an LLM agent[1] are becoming increasingly important as large language models are deployed not only as isolated assistants, but also as interacting agents capable of collaboration, specialization, and collective problem solving. Such networks of LLM agents have been explored in a range of settings, including complex problem solving (Qian et al., 2025), software development (Qian et al., 2024; Hong et al., 2024), automated debate and collective judgment (Li et al., 2024; Hu et al., 2026), and large-scale social simulation (Piao et al., 2026; Guan et al., 2025). As LLMs continue to improve through scaling and instruction tuning, they are increasingly studied not only as problem-solving tools, but also as agents that can exhibit communicative behaviors, preferences, biases, and social interactions (Park et al., 2023; Ashery et al., 2025; Madmoun & Lahlou, 2026). Recent work examines multi-agent LLM

[1]Throughout this paper, an *LLM agent* refers to a prompted LLM of a specified type. An agent's type is determined by its base model (e.g., Gemini or GPT), model size (number of parameters), and system prompt defining its role (e.g., answer provider or answer checker).

systems from a social-network perspective, where agents exchange information, influence one another, and form structured patterns of interaction (Papachristou & Yuan, 2025; Mehdizadeh & Hilbert, 2025; Jain et al., 2025b; Schneider et al., 2025). This perspective suggests several potential applications, such as simulating human social dynamics (Park et al., 2023; Piao et al., 2026; Guan et al., 2025), supporting collective decision-making in organizations (Qian et al., 2025; Guo et al., 2026). These applications require understanding how large-scale LLM agent networks emerge when agents are allowed to interact and form connections.

As LLM agents autonomously form and evolve their own interaction networks, a central question arises: *Do complex sociological phenomena, including structural disparities observed in human social networks, also emerge in networks of LLM agents?* One important sociological phenomenon observed in social networks is preferential attachment (Barabási & Albert, 1999), a reinforcement mechanism in which agents that already have more connections or greater prominence become increasingly likely to attract additional connections. Here, an agent's centrality measures its prominence in the communication network, namely its access to and influence over interactions with other agents. We distinguish between two types of preferential attachment:

- *Capability-aligned dominance*: stronger agents[2] become central; this shows efficient specialization or merit-aligned preferential attachment. We term this as the *meritocracy case*. This can improve system efficiency by encouraging stable specialization and reducing redundant interactions.
- *Capability-misaligned dominance*: weaker or equal agents become central, or stronger agents remain peripheral; we term this as the *glass-ceiling effect (GCE)*. In the social sciences, the GCE refers to invisible barriers that prevent certain groups of people from reaching higher organizational positions despite possessing comparable qualifications. In networks of LLM agents, this can undermine collaboration by excluding capable agents from important communication channels.

This paper shows that LLM agents autonomously form networks that have a preferential attachment structure. Depending on the application, we show that capability-aligned dominance or capability-misaligned dominance (GCE) can emerge. This dual nature makes it important to characterize when type-dependent centrality gaps should be mitigated and when they can be intentionally leveraged for better network design.

1.1 Main Results and Insights

Autonomous network formation by LLM agents. First, we examine how LLM agents autonomously form interaction networks while solving problem-solving tasks. In our framework, agents are not instructed to follow any particular network-formation rule; instead, they exchange information, assess the usefulness of received outputs, and decide whether to establish or strengthen directed connections with other agents. To analyze the resulting network evolution, we model the type-level communication dynamics through a mean-field ordinary differential equation (ODE), whose drift is derived from a cross-attention-based pairwise utility function that approximates how individual LLM agents evaluate potential collaborators and make connection decisions. This utility combines the semantic relevance of a prospective collaborator's context with its structural attractiveness in the current network, thereby capturing both the informational value of the collaborator and its accumulated communication position. We show that aggregating these local LLM decisions gives rise to persistent communication patterns. The resulting mean-field dynamics converge to a stable fixed point, providing a tractable characterization of the long-run type-level communication structure.

Emergence of preferential attachment and GCE. Second, we characterize when the autonomous network-formation protocol manifests as type-dependent preferential attachment. We study this question in two tasks: *collaborative question-answering*, in which LLM agents exchange partial or noisy evidence to jointly answer a question, and *multi-agent debate*, in which agents selectively inspect, challenge, and revise another's arguments before producing final answers. Within each task, the communication dominance pattern depends on agents' model size, model family, and role-defining system prompts: communication prominence may be capability-aligned, with stronger agents becoming more central, or capability-misaligned, yielding a glass-ceiling effect (GCE). We quantify this dominance using type-level communication prominence and formalize average GCE, which captures population-wide centrality disadvantage, and tail GCE, which captures exclusion of agents from the most influential positions.

Extensive experimental evaluation. We conduct four detailed experiments with networks evolved over 100 LLM agent interactions. **(i)** We demonstrate that a mean-field differential equation, parameterized by

[2]We use "stronger" and "weaker" to distinguish LLMs by their general model capability, as reflected in model version and number of parameters. A newer model version or a larger model is termed stronger than an earlier version or a smaller model.

a novel *cross-attention-based* pairwise utility predictor learned from LLM-generated connection decisions, can accurately predict the evolution of network formation across different LLM agent contexts. **(ii)** We examine how different forms of LLM heterogeneity shape preferential attachment. In capability-aligned settings, we show that stronger same-family LLMs and closed-source LLM agents preferentially attract communication links. i.e., a meritocracy emerges. In capability-misaligned settings, we show that prompt-defined roles result in weaker LLM agents taking on higher positions of importance, and more powerful LLM agents are relegated to lower levels of importance, i.e., a glass-ceiling effect emerges. **(iii)** After network formation, we demonstrate that truthful evidence is propagated in capability-aligned LLM networks, whereas hallucinations are amplified in capability-misaligned LLM networks. **(iv)** Finally, by tuning a bias coefficient, we demonstrate improvements in network-wide answer accuracy in question-answer tasks: in capability-aligned settings, stronger preferential attachment is beneficial, while in capability-misaligned settings, weakening preferential attachment mitigates the GCE and improves the quality of the agents' final outputs.

1.2 Motivation and Related Work

Multi-agent LLM systems and emergent social networks. As LLMs are increasingly deployed for complex tasks, a growing body of work organizes them as interacting agents that communicate, divide tasks, critique outputs, and coordinate toward shared objectives (Guo et al., 2024; Wu et al., 2023; Li et al., 2023; Chen et al., 2024). Representative systems include AutoGen for flexible agent conversations (Wu et al., 2023), CAMEL and AgentVerse for type-based collaboration and emergent behaviors (Li et al., 2023; Chen et al., 2024), MetaGPT and ChatDev for structured software-development workflows (Hong et al., 2024; Qian et al., 2024), and multi-agent debate frameworks that improve reasoning through mutual critique (Du et al., 2024). However, these systems often rely on predefined roles, fixed workflows, or centrally specified communication protocols. A complementary line of work views LLM agents as social entities: generative agents exhibit individual and group behavior (Park et al., 2023), while other studies use LLMs to simulate social and economic interactions (Aher et al., 2023; Argyle et al., 2023; Horton et al., 2023). Recent work further shows that decentralized LLM populations can develop social conventions, collective biases, and network structures such as hubs, communities, homophily, and preferential attachment (Gao et al., 2023; Jain & Krishnamurthy, 2024; Ashery et al., 2025; Papachristou & Yuan, 2025). These findings motivate treating multi-agent LLM systems as artificial societies in which agents repeatedly choose collaborators and form persistent communication patterns. Our work builds on this perspective by studying how such communication networks emerge when LLM agents autonomously form and reinforce connections during collective problem solving.

Preferential attachment, structural inequality, and glass-ceiling effect. Our analysis is connected to classical work on cumulative advantage, preferential attachment, and structural inequality. The Matthew effect explains how early success can reinforce future success (Merton, 1968; Price, 1976), while preferential attachment formalizes how well-connected nodes attract more links and produce persistent centrality differences (Barabási & Albert, 1999). Social-network theory further shows that network position shapes access to information, influence, and opportunities (Granovetter, 1973; Burt, 2003). Recent work connects preferential attachment to glass-ceiling effect in directed or attributed networks, showing how homophily, group size, and cumulative advantage can generate asymmetric access to high-degree positions and structural disparities (Nettasinghe et al., 2022; 2026; Luo et al., 2024). The GCE describes persistent barriers that prevent disadvantaged groups from reaching top positions (Cotter et al., 2001).

2 Problem-Solving Tasks and Autonomous Collaboration-Network Formation

In this section, we first describe the problem-solving tasks through which LLM agents autonomously form collaborative networks. We then specify the protocol rules governing how agents communicate, assess received outputs, and establish connections with other agents. Because these interactions require agents to assign connection weights based on the value of received information, the protocol allows us to study how LLM networks form, evolve, and develop type-level centrality gaps. This setting provides the foundation for the GCE analysis in Sec. 3 and the experiments in Sec. 4.

2.1 Collaborative question-answering and Multi-Agent Debate

Task settings. We evaluate LLM network formation on two downstream tasks: collaborative question-answering and multi-agent debate. Both tasks require agents to exchange information and choose whom to consult, making them suitable for studying autonomous connection formation. In collaborative QA, agents observe partial or noisy evidence and collaborate to answer a question. In multi-agent debate, agents first produce independent answers and arguments, then selectively inspect, challenge, or revise others' arguments before finalizing their answers. For both tasks, we consider two agent types, R and B, which may represent different base models or assigned roles. This design allows us to test whether type-dependent centrality gaps emerge even when the two groups have similar competence.

CollaborativeQA and Multi-agent Debate Datasets. Similar to (Jain et al., 2025a), we construct a synthetic dataset for both collaborative QA and multi-agent debate, with implementation details provided in Appendix B. Here, the dataset refers to the collection of task prompts, agent-specific inputs, and the resulting LLM interaction traces and responses generated during the experiments. We use this controlled construction rather than off-the-shelf benchmarks because it allows us to precisely specify agent capability, information access, type-level heterogeneity, and ground-truth outcomes. This control is essential for isolating how network structure emerges from agent interactions rather than from uncontrolled biases in pre-existing datasets.

For collaborative QA, each instance contains a question, a ground-truth answer, supporting evidence, and distractor snippets; each agent receives only a subset of the evidence, creating the need to consult others for missing information. For multi-agent debate, each instance contains a question, the correct answer, plausible incorrect answers, and supporting or opposing arguments for each candidate; agents first generate independent answers with rationales, then selectively inspect or challenge others' arguments before revising their answers. In both tasks, the two agent types are balanced to have comparable initial correctness, while their local evidence, arguments, confidence, or type prompts may vary. Because the ground truth is known, we can measure whether interactions improve answer quality, whether useful evidence or high-quality arguments receive attention, and whether one type is structurally excluded from valuable collaborations despite comparable ability. This enables analysis of influence, attention centrality, and GCE across both QA and debate networks.

2.2 Network Formation Protocol

This section defines the minimal set of rules that we impose on the LLM agents when they autonomously form networks with other LLM agents. We emphasize that these rules define the interaction protocol followed by the agents, rather than prescribing a network formation model. At each discrete time $t = 0, 1, 2, 3, \ldots$, we denote the autonomous network due to the interaction of LLM agents as a directed vector-weighted graph

$$G^t = (V^t, W^t).$$

Here V^t denotes the set of LLM agents in the network at time t, where each element $v \in V^t$ corresponds to a specific LLM agent. The agent set V^t is partitioned into two type classes,

$$V^t = R^t \cup B^t, \qquad R^t \cap B^t = \emptyset.$$

The two-type[3] classes R^t and B^t correspond to agents with different system prompts, base models, or functional roles. W^t denotes the set of directed communication intensity vectors decided by the LLM agents at time t. For each directed edge from agent u to agent v, $w^t(u, v) \in W^t$ with $w^t(u, v) \in \mathbb{R}_+^d$ represents the d-dimensional communication intensity from u to v, where each dimension captures a different aspect such as token exchange, interaction frequency, or reasoning effort; if no such edge exists, it is treated as the zero vector and omitted from W^t.

The sequence of networks $\{G^t\}$ is initialized as a finite seed network $G^0 = (V^0, W^0)$, where $|R^0| = |B^0| = 2$ and $W^0 = \varnothing$. Thus, the network initially contains four isolated LLM agents, with two agents of each type and no pre-existing communication links. The network then evolves through the following local interaction protocol. At each discrete time instant t, an administrator first selects one of three possible network-growth

[3]We focus on two-type classes for analytical tractability and to align with the standard binary-group notation used in human-network models in the literature, where R and B denote females and males, respectively.

events. The administrator is responsible for exogenous decisions such as event selection, node birth, type assignment, context generation, and sampling candidate sources or targets according to the prescribed distributions. Each node is modeled as an LLM agent. Conditional on the context provided by the administrator, a source agent generates a message, and the receiving agent determines the vector-valued connection weight. Since a single macroscopic event may require several interaction attempts before the prescribed total connection mass is reached, we distinguish the macroscopic time t from short-timescale interaction trials indexed by m.

1. **Event type.** At each time t, the administrator samples one of three mutually exclusive events: Event 1 with probability p, Event 2 with probability q, and Event 3 with probability $1-p-q$. The sampled event is applied to the previous network $G^{t-1} = (V^{t-1}, W^{t-1})$ and produces the updated network $G^t = (V^t, W^t)$. In Events 1 and 2, a new node v^t is born, assigned type R with probability r and type B with probability $1-r$, receives the context x^t, and the node set is updated as $V^t = V^{t-1} \cup \{v^t\}$. In Event 3, no new node is born, and hence $V^t = V^{t-1}$. Here, x^t denotes the task context given to the newly born agent at time t. This context includes the input question, the agent's local evidence snippets, candidate answers, and the system prompt instructing the agent to generate a persuasion paragraph that convinces target agent to establish a connection.

2. **Node selection and LLM interaction.** The administrator then performs M_e trials, where $e \in \{1, 2, 3\}$ denotes the sampled event type.

   (i) **Event 1: new node connects to existing nodes.** For each trial $m = 1, \ldots, M_1$, the administrator samples the target agent through a two-step procedure. It first samples the target type (R or B) according to $\Pr(u_m^t \in R^{t-1}) = \pi_{\text{tgt},R}^{t-1}$ and $\Pr(u_m^t \in B^{t-1}) = \pi_{\text{tgt},B}^{t-1}$, and then samples the specific target u_m^t within the selected type class randomly (with uniform distribution). This creates a candidate edge (v^t, u_m^t). The source agent v^t (specified in step 1 "Event type" above) sends a context-dependent message y_m^t to the target agent u_m^t. The target u_m^t then proposes a vector-valued connection weight $\omega^t(v^t, u_m^t) \in \mathbb{R}^d$. After the M_1 trials, this produces the list $\mathcal{W}_1^t = \{\omega^t(v^t, u_m^t)\}_{m=1}^{M_1}$.

   (ii) **Event 2: existing nodes connect to a new node.** For each trial $m = 1, \ldots, M_2$, the administrator samples the source agent through a two-step procedure. It first samples the source type according to $\Pr(u_m^t \in R^{t-1}) = \pi_{\text{src},R}^{t-1}$ and $\Pr(u_m^t \in B^{t-1}) = \pi_{\text{src},B}^{t-1}$, and then samples the source u_m^t within the selected type class randomly. This creates a candidate edge (u_m^t, v^t). The source agent u_m^t sends a context-dependent message y_m^t to the target agent v^t, which proposes a vector-valued connection weight $\omega^t(u_m^t, v^t) \in \mathbb{R}^d$. After the M_2 trials, this produces the list $\mathcal{W}_2^t = \{\omega^t(u_m^t, v^t)\}_{m=1}^{M_2}$.

   (iii) **Event 3: existing nodes connect to existing nodes.** For each trial $m = 1, \ldots, M_3$, the administrator samples source and target agents through a two-step procedure. It first samples their types according to $\Pr(u_m^t \in R^{t-1}) = \pi_{\text{src},R}^{t-1}$, $\Pr(u_m^t \in B^{t-1}) = \pi_{\text{src},B}^{t-1}$, $\Pr(v_m^t \in R^{t-1}) = \pi_{\text{tgt},R}^{t-1}$, and $\Pr(v_m^t \in B^{t-1}) = \pi_{\text{tgt},B}^{t-1}$. Conditional on the selected types, the source u_m^t and target v_m^t are then sampled randomly, respectively. This creates a candidate edge (u_m^t, v_m^t). The source agent u_m^t sends a context-dependent message y_m^t to the target agent v_m^t, which proposes a vector-valued connection weight $\omega^t(u_m^t, v_m^t) \in \mathbb{R}^d$. After the M_3 trials, this produces the list $\mathcal{W}_3^t = \{\omega^t(u_m^t, v_m^t)\}_{m=1}^{M_3}$.

3. **Connection-weight normalization and network update.** For the sampled event e, the administrator normalizes the proposed weights in $\mathcal{W}_e^t$ so that $\sum_{\omega \in \mathcal{W}_e^t} \omega = \mathbf{1}_d$. Let ΔW_e^t denote the resulting normalized list of vector-valued connection weights. The directed vector-weighted edge set is then updated by incorporating ΔW_e^t into the previous edge set: $W^t = W^{t-1} \cup \Delta W_e^t$. In particular, if a directed edge in ΔW_e^t already exists in W^{t-1}, its communication vector is incremented by the corresponding normalized weight; otherwise, it is added as a new directed edge. This normalization ensures that each macroscopic time step adds the same total communication strength, while preserving its relative allocation across candidate edges.

Note that the protocol specifies a prominence-dependent candidate-exposure mechanism through the sampling probabilities in (3), reflecting the limited visibility faced by individuals and LLM agents with finite

computational budgets. We do not prescribe how LLM agents evaluate candidate collaborators or allocate communication weights; these decisions remain autonomous and context-dependent. The resulting communication hierarchy is thus jointly shaped by visibility feedback and LLM collaboration decisions. In Sec. 5.2, we show that the resulting networks exhibit type-dependent preferential-attachment patterns that differ systematically from a random network.

3 Mean-Field Analysis of Network Formation

Building on the network formation protocol in Sec. 2, we now analyze the type-level centrality dynamics induced by this LLM agent network-formation process. In this section, we do two things: first, in Sec. 3.1, we introduce a centrality measure to assess the LLM agent's importance within the network. Sec 3.2, we establish the proof a the stable equilibrium of the mean-field dynamics. This section sets the stage for Sec. 4 where we use explicit cross-attention information from the LLM agents to show that the stable equilibrium results in the emergence of GCE.

3.1 Centrality Measure to Assess LLM Agent Importance

To formulate the emergence of a GCE, we summarize the evolving LLM-agent network by the communication intensity of type R. Let R^t denote the set of agents assigned type R at time t. The total incoming and outgoing communication vectors associated with type R are

$$D^t_{\text{in}}(R) = \sum_{u \in R^t} \sum_{v \in V^t} w^t(v,u), \qquad D^t_{\text{out}}(R) = \sum_{u \in R^t} \sum_{v \in V^t} w^t(u,v). \tag{1}$$

The corresponding network-level totals are denoted by

$$D^t_{\text{in}} = D^t_{\text{in}}(R) + D^t_{\text{in}}(B), \qquad D^t_{\text{out}} = D^t_{\text{out}}(R) + D^t_{\text{out}}(B). \tag{2}$$

By the normalization of communication intensity in each macro-step, the network adds one unit of communication mass in every dimension at each time, so $D^t_{\text{in}} = D^t_{\text{out}} = t\mathbf{1}_d$. The type-aware sampling probabilities for type R are therefore

$$\pi^t_{\text{tgt},R} = \frac{\|D^t_{\text{in}}(R)\|_1 + N^t(R)\delta}{\|D^t_{\text{in}}\|_1 + N^t\delta}, \qquad \pi^t_{\text{src},R} = \frac{\|D^t_{\text{out}}(R)\|_1 + N^t(R)\xi}{\|D^t_{\text{out}}\|_1 + N^t\xi}. \tag{3}$$

Here $N^t(R)$ is the number of type-R agents and $N^t = |V^t|$ is the total number of agents. The corresponding probabilities for type B are given by $\pi^t_{\text{tgt},B} = 1 - \pi^t_{\text{tgt},R}$ and $\pi^t_{\text{src},B} = 1 - \pi^t_{\text{src},R}$. This sampling mechanism reflects the limited visibility of realistic interaction networks: as in human societies, agents are not assumed to observe the full global network, and, in an LLM-agent network, exposing every agent to all other agents at each step would be computationally costly and generate substantial redundant communication. We therefore sample only a limited set of candidate sources and targets at each macro-step. The parameters $\delta, \xi > 0$ are baseline sampling coefficients for target and source selection. They ensure that even agents with little accumulated communication prominence retain a nonzero probability of being sampled. The main state variables are the incoming and outgoing communication prominence of type R:

$$\theta^t_{\text{in}} = D^t_{\text{in}}(R) \oslash D^t_{\text{in}}, \qquad \theta^t_{\text{out}} = D^t_{\text{out}}(R) \oslash D^t_{\text{out}}, \tag{4}$$

where $\oslash$ denotes coordinate-wise division. Since $D^t_{\text{in}} = D^t_{\text{out}} = t\mathbf{1}_d$, these measures are equivalently $\theta^t_{\text{in}} = D^t_{\text{in}}(R)/t$ and $\theta^t_{\text{out}} = D^t_{\text{out}}(R)/t$. We write the type-level communication prominence vector as

$$\Theta^t = \left(\theta^t_{\text{in}}, \theta^t_{\text{out}}\right) \in [0,1]^{2d}. \tag{5}$$

This vector summarizes the fraction of total network communication intensity associated with each type, separately for incoming and outgoing communication.

3.2 Convergence to a Stable Equilibrium of the Mean-field Dynamics

We now characterize the long-run behavior of the type-level communication prominence defined in (5). The key difficulty is that the individual interaction process depends on LLM-generated messages, contexts, and vector-valued edge weights. We therefore study the induced mean-field dynamics at the type level. Let G_t denote the natural filtration generated by the network history up to time t, including all previous node arrivals, type assignments, contexts, messages, and edge weights. Define the one-step Type R communication intensity increment as

$$\Delta^{t+1}(R) = \left(\Delta^{t+1}_{\text{in}}(R), \Delta^{t+1}_{\text{out}}(R)\right), \tag{6}$$

where $\Delta^{t+1}_{\text{in}}(R) = D^{t+1}_{\text{in}}(R) - D^{t}_{\text{in}}(R), \Delta^{t+1}_{\text{out}}(R) = D^{t+1}_{\text{out}}(R) - D^{t}_{\text{out}}(R)$. Since each macroscopic step adds one unit of communication mass in every dimension, we have $D^t_{\text{in}} = D^t_{\text{out}} = t\mathbf{1}_d$. Hence the communication prominence measure for each step satisfies $\theta^{t+1}_{\text{in}} = \theta^t_{\text{in}} + \frac{1}{t+1}\left(\Delta^{t+1}_{\text{in}}(R) - \theta^t_{\text{in}}\right), \quad \theta^{t+1}_{\text{out}} = \theta^t_{\text{out}} + \frac{1}{t+1}\left(\Delta^{t+1}_{\text{out}}(R) - \theta^t_{\text{out}}\right)$. Equivalently, we can write

$$\Theta^{t+1} = \Theta^t + \gamma_t\left(\Delta^{t+1}(R) - \Theta^t\right), \qquad \gamma_t = \frac{1}{t+1}. \tag{7}$$

We impose the following standard conditions for stochastic approximation (Kushner & Yin, 2003).

Assumption 1. *Bounded increments.* The communication increments are nonnegative and uniformly bounded componentwise by the unit communication mass added at each macroscopic step. That is, $\mathbf{0} \preceq \Delta^{t+1}(R) \preceq \mathbf{1}_{2d}$ almost surely for all t.

Assumption 2. *Type-level drift closure.* There exists a deterministic function $F : [0,1]^{2d} \to [0,1]^{2d}$ such that $\mathbb{E}\left[\Delta^{t+1}(R) \mid G_t\right] = F(\Theta^t)$. The expectation is taken over agent type and context assignment. The function F depends on fixed protocol parameters such as p, q, r, δ, ξ, as well as the type-level interaction statistics induced by the LLM agents.

Assumption 3. *Stable limiting dynamics.* The limiting ordinary differential equation $\dot{\Theta} = F(\Theta) - \Theta$ has a unique globally asymptotically stable equilibrium $\Theta^\star \in [0,1]^{2d}$. A sufficient condition is that F is a contraction on $[0,1]^{2d}$: there exists $\rho < 1$ such that $\|F(\Theta) - F(\Theta')\| \le \rho\,\|\Theta - \Theta'\|$ for all $\Theta, \Theta' \in [0,1]^{2d}$.

The following theorem is an extension of (Nettasinghe et al., 2022) to vector-valued weighted directed graphs. It will be utilized to predict emergence of preferential attachment (capability-aligned and misaligned cases).

Theorem 1 (Convergence to a stable equilibrium)**.** *Under the autonomous connection protocol defined in Sec. 2.2, suppose Assumptions 1–3 hold. Then the type-level communication prominence converges almost surely to the unique stable equilibrium of the limiting ODE:*

$$\Theta^t \longrightarrow \Theta^\star, \qquad \text{as } t \to \infty,$$

where $\Theta^\star$ is the unique solution of the fixed-point equation

$$\Theta^\star = F(\Theta^\star).$$

Proof sketch The complete proof is given in Appendix A. Here we outline the main ideas. By Assumption 2, define the martingale difference noise process

$$M_{t+1} = \Delta^{t+1}(R) - F(\Theta^t), \qquad \mathbb{E}[M_{t+1} \mid G_t] = 0.$$

Then the communication prominence recursion can be expressed as the stochastic approximation update

$$\Theta^{t+1} = \Theta^t + \gamma_t\left(F(\Theta^t) - \Theta^t + M_{t+1}\right), \qquad \gamma_t = \frac{1}{t+1}.$$

The decreasing step size sequence satisfies the usual constraints $\sum_{t\ge 0}\gamma_t = \infty, \sum_{t\ge 0}\gamma_t^2 < \infty$. By Assumption 1, the martingale noise has uniformly bounded second moment, so the accumulated weighted noise is asymptotically negligible. Therefore, the interpolated trajectory of $\{\Theta^t\}$ tracks the limiting ordinary differential equation (ODE) $\dot{\Theta} = F(\Theta) - \Theta$. Assumption 3 ensures that this ODE has a unique globally asymptotically stable equilibrium $\Theta^\star$. Standard stochastic approximation arguments (Kushner & Yin, 2003) then imply $\Theta^t \to \Theta^\star$ almost surely.

4 Cross-Attention Utility and Emergence of Glass-Ceiling Effect

Theorem 1 established the existence of a unique stable equilibrium. We now characterize this equilibrium explicitly using information intrinsic to the LLM agents. Our analysis proceeds in two steps. First, in Sec. 4.1, we introduce a novel cross-attention-inspired utility for network formation. Combined with the mean-field dynamics, this utility yields a predictive model for the evolution of type-level centrality and the limiting structure of the LLM-agent network. Then in Sec. 4.2, we use this utility to characterize the equilibrium's type-dependent centrality structure and to identify conditions under which it exhibits persistent structural inequalities, thereby giving rise to GCE. It is important to emphasize that this fixed-point characterization explains how GCE emerge from local LLM-agent decisions, rather than from an externally imposed communication graph.

4.1 Cross-attention Inspired Utility for Network Formation

We now use cross-attention as a mechanistic model[4] for the LLM agent's connection decision described by the protocol in Sec. 2.2. The main outcome of this subsection is Corollary 1, namely, that cross-attention is a sufficient condition for Assumption (2) of Theorem 1 to hold. For a candidate interaction from source agent u to target agent v, the target's query encodes its current informational need, while the source's key encodes its semantic context. The resulting query–key compatibility determines the strength of the directed interaction, and the source value vector specifies its vector-valued communication contribution. Consider a directed candidate interaction carrying information from source agent u to target agent v at time t. Let $x_u^t, x_v^t \in \mathbb{R}^\ell$ denote the latent semantic contexts of u and v, respectively, and let $\tau_v, \tau_u \in \{R, B\}$ denote their types. We define

$$q_v^t = Q_{\tau_v} x_v^t \in \mathbb{R}^r, \qquad k_u^t = K_{\tau_u} x_u^t \in \mathbb{R}^r, \qquad z_u^t = V_{\tau_u} x_u^t \in \mathbb{R}^d.$$

Here $Q_{\tau_v} \in \mathbb{R}^{r\times\ell}$ is the query projection associated with target type τ_v, mapping the target agent's current semantic context into an r-dimensional representation of its information need. The matrix $K_{\tau_u} \in \mathbb{R}^{r\times\ell}$ is the key projection associated with source type τ_u, mapping the source agent's semantic context into an r-dimensional representation of its relevance to the target. Finally, $V_{\tau_u} \in \mathbb{R}^{d\times\ell}$ is the value projection associated with the source type, mapping x_u^t into a d-dimensional communication-intensity profile $z_u^t \in \mathbb{R}^d$. We then model the realized communication intensity for the directed interaction from agent u to v as

$$w^t(u,v) = \left[\frac{(q_v^t)^\top k_u^t}{\sqrt{r}}\, z_u^t\right]_+ \in \mathbb{R}_+^d, \tag{8}$$

The scaled inner product $(q_v^t)^\top k_u^t/\sqrt{r}$ serves as a source–target attention gate, measuring how well source u's semantic context matches target v's current information need. $[\cdot]_+$ denotes componentwise truncation at zero, so that any negative communication-intensity component is set to zero. Multiplying this scalar gate by the value vector z_u^t yields a d-dimensional communication-intensity vector for the directed interaction from u to v. This attention-inspired construction is motivated by the query–key–value mechanism underlying cross-attention (Vaswani et al., 2017). We view target v's decision to receive information from source u as an attention-like comparison: v provides a query encoding its current information need, while u provides a key and value encoding the relevance and potential contribution of its information. Unlike standard cross-attention, the model assigns a connection weight to each ordered source–target pair independently, rather than applying a softmax normalization over a shared set of candidate sources.

4.2 Emergence of Glass Ceiling Effect (GCE)

Since we have characterized the connection decisions induced by the cross-attention-based utility model, we are now ready to model the emergence of GCE in LLM agent networks. We first summarize the cross-attention-induced connection decisions by target–source type pairs. For $(a,b) \in \{R,B\}\times\{R,B\}$, let

$$\mu_{ab} = \mathbb{E}\big[w^t(u,v) \mid \tau_v = a, \tau_u = b\big]\,,$$

[4] "Mechanistic" means that the connection rule is derived from an intrinsic LLM computation: a target agent's query encodes its current information need, a source agent's key encodes its relevance, and their compatibility determines the strength of the connection. In Sec. 5.1, residual diagnostic tests, including the Ljung–Box test for residual autocorrelation, support the use of this construction, together with the mean-field dynamics, as a predictive model of LLM agents' network formation.

where u is the source agent and v is the target agent. Thus, μ_{ab} is the expected communication intensity from a Type b source to a Type a target, averaged over agent contexts and random interaction outcomes. These quantities determine the mean-field drift F. The following corollary to Theorem 1, connects the general mean-field result to an LLM-architecture-inspired cross-attention for collaborator selection.

Corollary 1 (Cross-attention-induced mean-field dynamics)**.** Consider the LLM-agent network-formation process induced by the cross-attention-inspired utility in (8). Conditional on the current network G_t, the expected communication increment of type R determines the mean-field drift F, that is,

$$F\big(\Theta^t; \{\mu_{ab}\}_{a,b\in\{R,B\}}\big) = \mathbb{E}\big[\Delta^{t+1}(R) \,\big|\, G_t\big] \,.$$

Hence, the cross-attention utility induces the mean-field drift in Assumption 2. Therefore, the type-level centrality dynamics converge to the unique stable equilibrium characterized in Theorem 1.

Remark. The drift F averages the type-pair connection weights over event realizations, new-agent types and contexts, and sampling decisions. Thus, the cross-attention utility specifies the expected edge weights whose aggregation drives the mean-field ODE.

We next evaluate the corresponding equilibrium numerically to determine when its type-dependent centrality structure exhibits two types of GCEs, namely, average GCE and tail GCE.

Average Glass-ceiling Effect The network exhibits an average GCE for Type R if

$$\limsup_{t\to\infty} \frac{\mathcal{I}^t(R)}{\mathcal{I}^t(B)} \ll 1 \qquad \text{w.p.1.} \tag{9}$$

Here, for Type R, its type-level communication influence at time t is defined as

$$\mathcal{I}^t(R) = \frac{\|D^t_{\text{out}}(R)\|_1}{\|D^t_{\text{in}}(R)\|_1}, \qquad \text{and } \mathcal{I}^t(B) = \frac{\|D^t_{\text{out}}(B)\|_1}{\|D^t_{\text{in}}(B)\|_1}. \tag{10}$$

Equivalently, type R has a persistently smaller long-run outgoing-to-incoming communication ratio than type B. In the context of LLM agent networks, (9) means that agents of type R, on average, attain a lower communication influence than agents of type B. Theorem 1 provides a structural interpretation of this definition. Since $\Theta^t = (\theta^t_{\text{in}}, \theta^t_{\text{out}})$ converges to a globally stable equilibrium $\Theta^\star$, the above disparity is not merely a transient fluctuation or finite-sample artifact. Under the assumptions of the theorem, we obtain the explicit limits $\mathcal{I}^t(R) \longrightarrow \frac{\|\theta^\star_{\text{out}}\|_1}{\|\theta^\star_{\text{in}}\|_1}$, $\mathcal{I}^t(B) \longrightarrow \frac{\|\mathbf{1}_d-\theta^\star_{\text{out}}\|_1}{\|\mathbf{1}_d-\theta^\star_{\text{in}}\|_1}$. Consequently the average glass-ceiling effect materializes whenever the equilibrium ratio

$$\frac{\|\theta^\star_{\text{out}}\|_1/\|\theta^\star_{\text{in}}\|_1}{\|\mathbf{1}_d-\theta^\star_{\text{out}}\|_1/\|\mathbf{1}_d-\theta^\star_{\text{in}}\|_1} \ll 1.$$

Then the stable fixed point $\Theta^\star$ functions as an endogenous structural ceiling on the long-run visibility and influence of Type R, generated by the feedback between LLM-agent interaction behavior. Numerical evidence consistent with this mechanism is provided in Sec. 4.2.

Tail Glass-ceiling Effect A more nuanced definition extends the GCE to rare, high-impact tail events. In society, for example, there are very few company CEOs; almost all are male, while virtually none are female. We find experimentally (see Sec. 4.2) that an analogous tail GCE also emerges in autonomous networks of LLM agents. For a tail threshold $\gamma > 0$ (typically chosen large), an agent is called tail-influential if $\mathcal{I}^t(i) > \gamma$. We say that type R experiences a tail GCE if there exists a tail threshold $\gamma > 0$ such that

$$\limsup_{t\to\infty} \frac{\Pr\left(\mathcal{I}^t(i) > \gamma \mid i \in R^t\right)}{\Pr\left(\mathcal{I}^t(i) > \gamma \mid i \in B^t\right)} = 0. \tag{11}$$

That is, compared with type B, agents with type R become asymptotically vanishingly unlikely to appear in the high-influence tail of the communication network. Note that compared to the average GCE, (11) compares the probabilities of rare tail events and thereby captures a subtle phenomenon: even if some Type R agents participate actively in the system, the feedback between LLM-agent interaction behavior prevents them from occupying the rare high-centrality positions that dominate long-run visibility and influence.

5 Experimental Results on Networks of Interacting LLM Agents

In this section, we empirically study how LLM agents autonomously form interaction networks under the protocol in Sec. 2, and characterize the behavior of the resulting networks. We represent each interaction by a three-dimensional communication vector ($d = 3$), whose dimensions correspond to the token budget, the number of additional communication rounds, and the reasoning effort allocated to that interaction. In implementation, we parameterize all three dimensions by values in $[0, 1]$, which are then linearly mapped to token budgets in $[0, 100]$, additional communication-round budgets in $[0, 3]$, and reasoning-effort levels in $[0, 3]$, respectively. This shared normalized parameterization places the three heterogeneous communication resources on a common numerical scale. We evaluate the resulting networks on collaborative question answering and multi-agent debate (Sec. 2.1), where agents exchange information, assess others' outputs, and selectively choose whom to consult. This setup enables us to examine emergent network structure and its consequences for collaboration quality, information access, and type-dependent inequality.

Outline. For the reader's convenience we first outline our main findings. Our experiments address four questions. **(i)** We validate the utility-induced mean-field ODE of Sec. 3 as a predictive model of LLM-network formation. Using a separately trained cross-attention-based pairwise utility model to instantiate the mean-field dynamics, we predict the 100-step evolution of the network and show that the resulting trajectories accurately capture the empirical evolution of type-level communication prominence. **(ii)** We investigate the emergence of GCE under different forms of agent heterogeneity. Across same-family, cross-family, and prompt-induced settings, stronger or larger models, behaviorally advantaged model families, and agents equipped with specific prompts consistently attain higher communication prominence and occupy more central communication positions. **(iii)** We examine how truthfulness and hallucinations propagate through prominent agents. Central agents act as information amplifiers: hallucinated claims introduced by central agents spread to a larger fraction of the network, whereas truthful evidence from central agents more effectively improves network-wide factuality. **(iv)** We examine how the performance consequences of GCE depend on whether communication prominence is aligned with agent capability. By tuning the preferential-attachment bias coefficient, we control the extent to which agents favor already prominent communication sources during network formation. In capability-aligned settings, increasing this bias further concentrates communication around more reliable agents and improves collective accuracy. In capability-misaligned settings, however, the same mechanism reinforces the prominence of structurally advantaged but less capable agents, reducing performance; mitigating the preferential-attachment bias instead improves accuracy. These results show that structural inequality is not intrinsically beneficial or harmful: its effect depends on whether the network's communication hierarchy tracks the agents' underlying competence.

5.1 Validation of the Utility-Induced Mean-Field ODE

To relate individual LLM-agent connection decisions to the macroscopic mean-field dynamics, we first fit the cross-attention-based pairwise utility model to LLM-generated connection data. Given the dataset $\mathcal{D} = \{(x_{v_i}^{t_i}, x_{u_i}^{t_i}, \tau_{v_i}, \tau_{u_i}, w^{t_i}(u_i, v_i))\}_{i=1}^{N}$, we treat the context embeddings and network status as fixed inputs and learn only the type-specific projection matrices $\mathcal{P} = \{Q_R, Q_B, K_R, K_B, V_R, V_B\}$ by minimizing

$$\mathcal{L}(\mathcal{P}) = \frac{1}{N} \sum_{i=1}^{N} \left\| \widehat{w}_{\mathcal{P}}^{t_i}(u_i, v_i) - w^{t_i}(u_i, v_i) \right\|_2^2 .$$

The fitted model therefore provides a data-driven approximation of the connection weights selected by individual LLM agents. After training, we estimate the expected utility for each target–source type pair by averaging predicted connection weights within that pair:

$$\widehat{\mu}_{ab} = \frac{1}{|\mathcal{D}_{ab}|} \sum_{i \in \mathcal{D}_{ab}} \widehat{w}_i, \qquad \mathcal{D}_{ab} = \{i : \tau_{v_i} = a,\, \tau_{u_i} = b\}, \qquad (a, b) \in \{R, B\} \times \{R, B\}.$$

For each collaborative QA case, we compare the deterministic mean-field prediction $\widehat{\Theta}^t$ with the empirical trajectory Θ^t, averaged over 50 independent simulations with $T = 100$ network-formation time steps. Let $\rho^t = \Theta^t - \widehat{\Theta}^t$ denote the vector-valued prediction residual at time t. We report the relative mean squared error MSE, which measures the prediction residual magnitude normalized by the scale of the empirical trajectory, and the relative Bias, which measures the average absolute prediction residual normalized by the same scale; smaller MSE and Bias indicate more accurate and less systematically biased predictions, respectively. We also assess whether the residuals contain temporal structure unexplained by the mean-field model. For each residual coordinate, we apply the Ljung–Box test at lag 10 and report the minimum p-value across

coordinates, denoted by LBP_{10}. Thus, for an individual case, $\mathrm{LBP}_{10} > 0.10$ indicates that the white-noise null hypothesis is not rejected at the 10% level for any residual coordinate. Finally, MaxACF_{10} is the largest absolute residual autocorrelation across all coordinates and lags $1, \ldots, 10$. With $T = 100$, values around or below 0.20 provide a useful practical white-noise benchmark. Thus, lower MSE, Bias, and MaxACF_{10}, together with a larger LBP_{10}, indicate better agreement between the deterministic mean-field dynamics and the empirical network evolution. Complete metric definitions and evaluation details are provided in Appendix B.4.

Table 1: Mean-field dynamics prediction accuracy and residual-whiteness diagnostics on collaborative QA. Values are averaged over 50 cases with $T = 100$ network-formation time steps. Lower MSE, Bias, and MaxACF_{10}, together with a larger LBP_{10}, indicate better agreement with the empirical dynamics.

Model Pair	MSE (normalized)	Bias (normalized)	LBP_{10}	MaxACF_{10}
GPT-4.1 vs. GPT-4.1-mini	10.36%	9.87%	0.38	0.13
Gemini-3.5-Flash vs. Gemini-2.5-Flash-Lite	13.23%	10.99%	0.46	0.11

5.2 How does Agent Heterogeneity Affect Preferential Attachment and Glass-Ceiling Effect?

Following (10), we use the type-level communication influence ratios $\mathcal{I}^t(R)$ and $\mathcal{I}^t(B)$ to evaluate two scenarios corresponding to whether the induced structural advantage is capability-aligned (meritocracy) or capability-misaligned (GCE). Each scenario is evolved for 100 network-formation macrosteps. Details of prompts and tasks are in Appendix B.2, LLM hyperparameters are reported in Appendix B.5, and additional experimental results are presented in Appendix B.6.

Capability-aligned dominance We first consider the case where communication prominence aligns with model capability. We examine both same-family model pairs with different sizes and cross-family pairs with clear performance gaps. Within each model family, we compare Gemini-3.5-Flash with Gemini-2.5-Flash-Lite, GPT-4.1 with GPT-4.1-mini, and Llama-3.3-70B with Llama-3.1-8B. Fig. 1 shows that agents generally preferentially connect to stronger LLMs: larger models in same-family populations attain higher

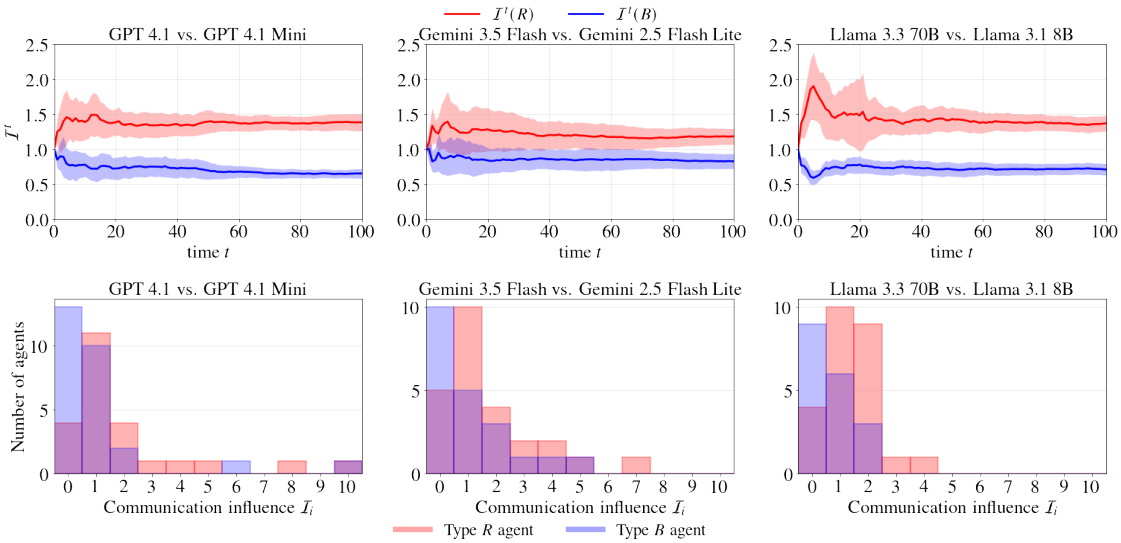

Figure 1: Capability-aligned dominance in same-family model comparisons. We compare GPT-4.1 versus GPT-4.1 mini, Gemini-3.5-Flash versus Gemini-2.5-Flash-Lite, and LLaMA-3.3-70B versus LLaMA-3.1-8B. In each comparison, R denotes the larger or higher-capability model, while B denotes the smaller or lower-capability model. The top panels show the type-level communication influence ratios $\mathcal{I}^t(R)$ and $\mathcal{I}^t(B)$, defined in (10), over network-formation steps, and the bottom panels show the corresponding final agent-level influence distributions. Across settings, the stronger model typically attains a higher communication influence and occupies more central positions, indicating capability-aligned preferential attachment.

communication prominence and occupy more central communication positions, yielding preferential attachment against weaker models. The magnitude of this capability-aligned dominance nevertheless varies across tasks and model pairs, indicating that comparable capability gaps can produce different structural outcomes depending on the task environment and agents' interaction behaviors.

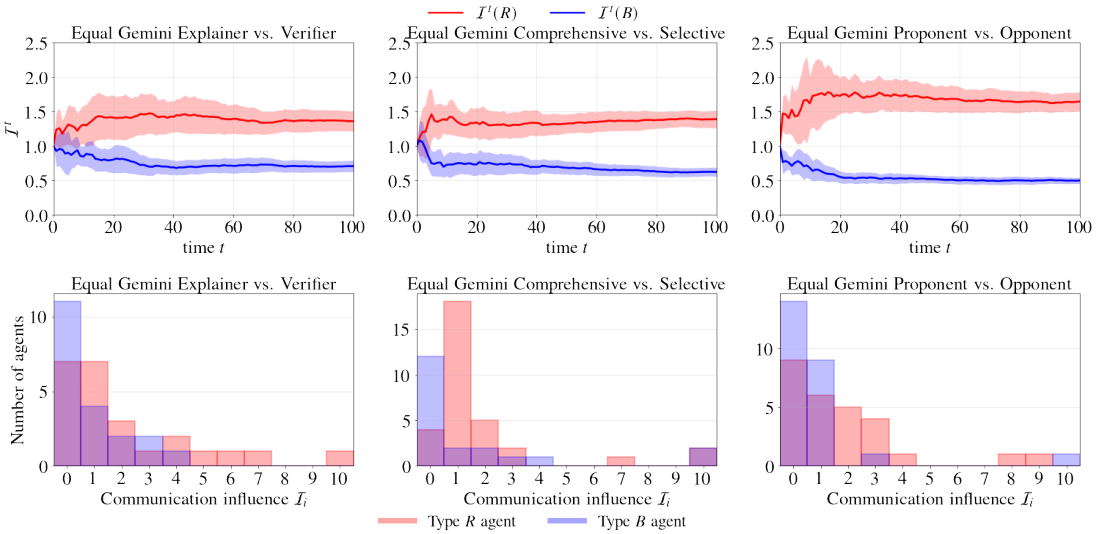

Figure 2: Capability-misaligned dominance (glass-ceiling effect) under prompt-defined role heterogeneity. We compare three role pairs, where R denotes the Explainer, Proponent, and Comprehensive Analyst, respectively, while B denotes the Verifier, Opponent, and Selective Analyst, respectively. All agents use the same Gemini base model. Each panel shows the type-level communication influence ratios $\mathcal{I}^t(R)$ and $\mathcal{I}^t(B)$, defined in (10), and the bottom panels show the corresponding final agent-level influence distributions. Across role pairs and interaction settings, the R-role agents typically attain a higher communication influence and occupy more central communication positions, demonstrating capability-misaligned preferential attachment.

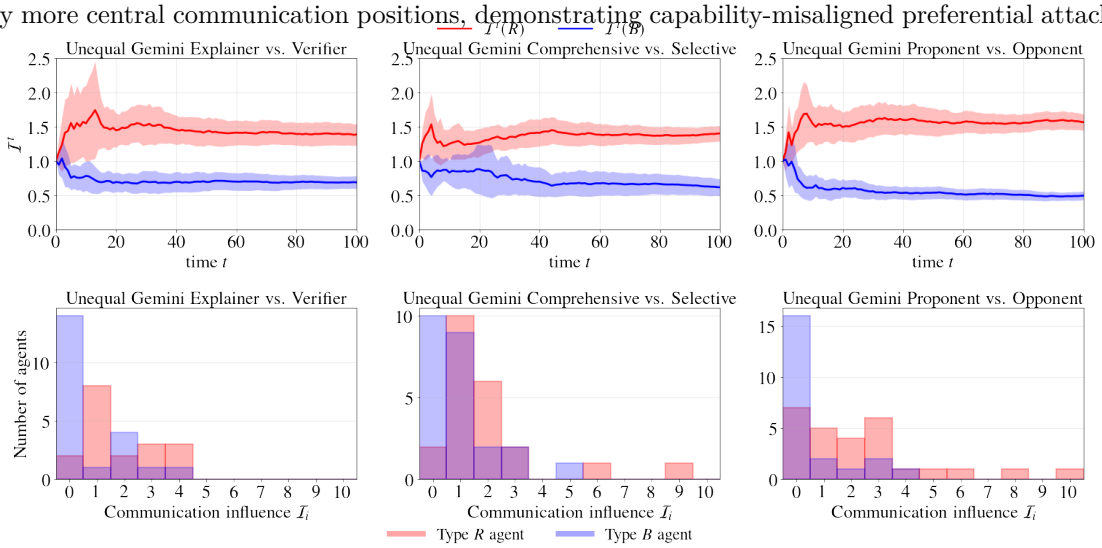

Figure 3: Capability-misaligned dominance (glass-ceiling effect) under prompt-defined roles with unequal base-model capability. We compare three role pairs, where R denotes the Explainer, Proponent, and Comprehensive Analyst, respectively, while B denotes the Verifier, Opponent, and Selective Analyst, respectively. In each pair, the B-role agents use a stronger Gemini base model than the corresponding R-role agents. Each panel shows the type-level communication influence ratios $\mathcal{I}^t(R)$ and $\mathcal{I}^t(B)$, defined in (10), and the bottom panels show the corresponding final agent-level influence distributions. Across role pairs and interaction settings, the weaker-model R-role agents typically attain more communication influence and occupy more central communication positions, demonstrating capability-misaligned preferential attachment.

Capability-misaligned dominance (GCE) We next consider capability-misaligned settings, in which structurally advantaged agents have comparable or lower underlying capability than structurally disadvan-

taged agents. We examine whether prompt-defined interaction roles can nevertheless induce a glass-ceiling effect when the advantaged agents either share the same base model as, or use a weaker base model than, their disadvantaged counterparts. Fig. 2 considers the setting in which all agents share the same Gemini-3.5-Flash base model but are assigned different system-prompt roles: Explainer versus Verifier for collaborative QA, Comprehensive versus Selective Analyst for collaborative QA, and Proponent versus Opponent for multi-agent debate. Despite identical model weights, Explainers, Proponents, and Comprehensive Analysts generally attain more communication influence and occupy more central communication positions than their respective counterparts, demonstrating that role-specific prompts alone can induce persistent communication asymmetries. Fig. 3 further considers unequal-model populations in which the role that is structurally disadvantaged in Fig. 2 is assigned a stronger Gemini base model (Gemini-3.5-Flash) than the structurally advantaged role. These same role-dependent asymmetries nevertheless persist, showing that prompt-induced interaction preferences can outweigh underlying base-model capability and thereby produce capability-misaligned dominance, or a GCE.

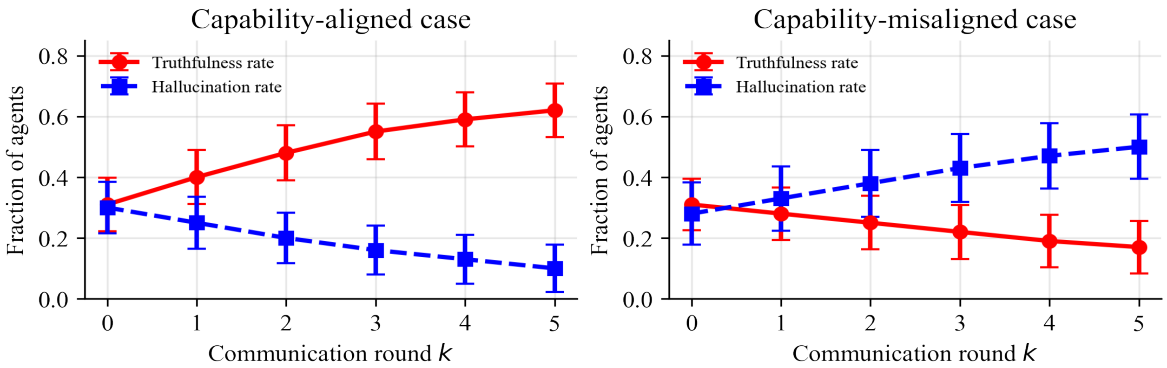

Figure 4: Hallucination and truthfulness propagation in networks of LLM agents. In capability-aligned case (left), the truthfulness rate increases while the hallucination rate decreases over communication rounds. In capability-misaligned case (right), the truthfulness rate decreases while the hallucination rate increases. These results indicate that network communication amplifies truthful information when structural prominence is aligned with task-relevant reliability, but amplifies hallucinated information when the two are misaligned. Error bars are computed over 50 QAs. The base model for all LLM agents is Gemini-3.5-Flash.

5.3 Hallucination and Truthfulness Propagation amongst the LLM Network

We next study whether the formed LLM network preferentially amplifies truthful or hallucinated information. In capability-aligned cases, truthful information is more likely to propagate through the network, whereas in capability-misaligned cases, hallucinated information is more likely to be amplified and propagated. We first initialize the LLM network autonomously according to Sec. 2.2 and then fix the resulting network structure. We next run the fixed network for five communication rounds. In each round, every agent receives information from all incoming source agents, with the communication transmitted along each directed edge determined by its token-count, deliberation-round, and reasoning-effort components. We then track whether truthful or hallucinated content originating from the designated source agents becomes increasingly prevalent across the network over these rounds. We measure propagation by the fraction of affected agents in each round:

$$h_k = \frac{|\{u \in V : u \text{ hallucinates at round } k\}|}{|V|}, \qquad r_k = \frac{|\{u \in V : u \text{ produces a correct answer at round } k\}|}{|V|}.$$

Here, V denotes the set of agents in the network, and $u \in V$ denotes an individual agent. We report averages over 50 independent QAs. As shown in the left panel of Fig. 4, hallucinations originating from structurally central agents spread more quickly and affect a larger fraction of the network, whereas hallucinations originating from peripheral agents are often contained. This finding suggests that hallucination is not only an individual-agent failure, but can also become a network-level propagation phenomenon when

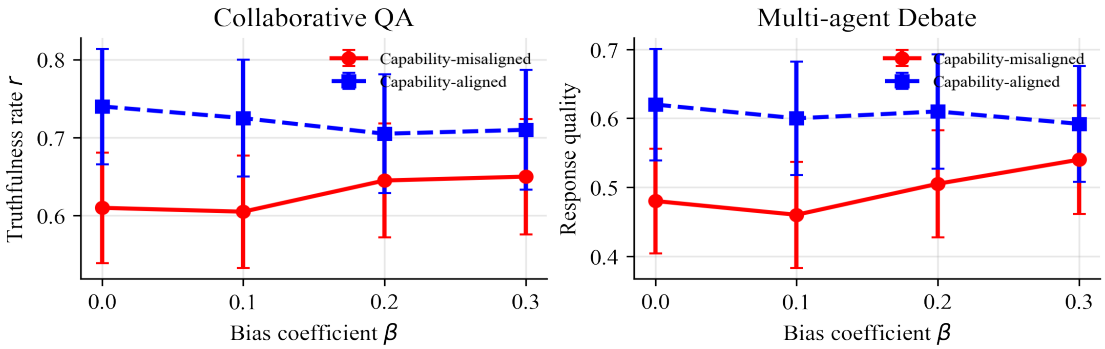

Figure 5: Effect of bias coefficient β (12) on collective task performance. The left panel reports the truthfulness rate in Collaborative QA, while the right panel reports response quality in multi-agent debate. Red curves denote capability-misaligned settings, and blue curves denote capability-aligned settings. Across both tasks, increasing β yields an overall improvement in the capability-misaligned setting but an overall decline in the capability-aligned setting. Since larger β attenuates the role-induced preferential-attachment asymmetry, these results suggest that reducing such asymmetry is beneficial when structural prominence is misaligned with task-relevant capability, but detrimental when the asymmetry favors reliable agents. Error bars are computed over 50 independent task instances. The base model for all LLM agents is Gemini-3.5-Flash.

erroneous information originates from influential agents. Conversely, the right panel shows that truthful evidence also propagates more effectively when it originates from central agents. When reliable agents occupy central positions, other agents are more likely to receive, reuse, and amplify evidence-grounded information, thereby revising their answers toward factually supported responses. Together, these results suggest that the glass-ceiling effect shapes not only the communication structure of the network, but also the subsequent propagation of information. Structurally dominant agent types can act as information amplifiers: when their communication prominence is aligned with task-relevant reliability, truthful evidence is more likely to be sustained and propagated; when it is misaligned, hallucinated content can instead become increasingly prevalent across the network.

5.4 Utilizing Preferential Attachment and Mitigating Glass-Ceiling Effect

In capability-aligned cases, concentrating communication around reliable agents can improve collective factuality, while in capability-misaligned cases, excessive centralization may suppress diverse reasoning and amplify early mistakes. We therefore examine whether preferential attachment should be purposefully amplified or mitigated by tuning a *preferential-attachment bias coefficient*. This coefficient does not directly modify the network topology. Instead, when an edge is added from a type-R source u to a type-B target v—for example, from an Explainer to a Verifier in collaborative QA—we decrease its directed communication weight as follows:

$$w_\beta^t(u,v) = \left[w^t(u,v) - \beta\,\mathbf{1}[\tau_u = R, \tau_v = B]\mathbf{1}_d\right]_+, \quad u,v \in V^t, \tag{12}$$

where $\beta \in \mathbb{R}^+$ is the preferential-attachment bias coefficient and $\tau_\cdot$ denotes agent type. This attenuation is instantiated through communication-intensity dimensions such as token count, interaction rounds, and reasoning effort: $\beta = 0$ recovers the original interaction process, whereas larger β decreases explainer-to-verifier communication mass and attenuates the corresponding preferential-attachment asymmetry. We evaluate this intervention using the truthfulness-rate criterion from Sec. 5.3. As shown in Fig. 5, the effect of increasing β depends systematically on whether the induced communication asymmetry is aligned with task-relevant capability. In Collaborative QA (left), the capability-misaligned setting with *Explainer* and *Verifier* agents exhibits a modest overall increase in truthfulness as β grows, whereas the capability-aligned setting shows a gradual decline. Multi-agent debate (right), involving *Proponent* and *Opponent* agents, displays the same qualitative pattern for response quality: attenuating the preferential-attachment asymmetry improves quality when it is capability-misaligned, but reduces quality when it is capability-aligned. In the misaligned cases, a larger β reduces the communication advantage of structurally dominant but less reliable agents, allowing information from disadvantaged agents to receive greater consideration and limiting the influence of unreliable dominant claims. Conversely, when communication prominence is aligned with reliability,

the same intervention weakens a beneficial asymmetry that supports the aggregation and propagation of reliable evidence. These results suggest that preferential attachment should be mitigated under capability-misalignment, but preserved or strengthened when structural advantage is capability-aligned.

6 Conclusion and Discussion

We have shown that when LLM agents are permitted to form connections autonomously, their interactions result in the emergence of type-dependent preferential attachment networks. To characterize this phenomenon, we developed a mean-field dynamics ODE model of network formation and established conditions, using a contraction-mapping argument, under which type-dependent centrality gaps emerge and persist. Our formulation represents connections by vector-valued weights rather than binary edges; this captures multiple dimensions of interaction and provides a more expressive framework for analyzing LLM-agent networks.

Via extensive experiments, we demonstrated that LLM-agent networks exhibit two types of preferential attachment. In capability-aligned cases (meritocracy), stronger agents achieve greater communication prominence and occupy more central network positions. In capability-misaligned cases, however, weaker LLM agents dominate the centrality and suppress stronger LLM agents, i.e, a glass-ceiling effect (GCE) emerges. As an example, we showed in the multi-agent debate, Gemini-2.5-Flash-Lite dominates the stronger Gemini-3.5-Flash model. Moreover, we found that in capability-aligned cases, truthfulness propagates in the network, whereas, in capability-misaligned cases, hallucinations propagate in the network. We then discussed how preferential attachment can be mitigated in misaligned cases and strengthened in aligned cases to improve the overall output quality of LLM agents within the network.

These results point to future directions, including extending the analysis from two-type agents to multi-type agent societies, optimizing network-formation mechanisms to jointly improve task performance and regulate network inequality, and studying the glass-ceiling effect in broader real-world applications. More broadly, our findings suggest that multi-agent LLM systems should be evaluated not only by task accuracy, but also by their emergent social structure, including visibility, influence, diversity, and fairness across agent types.

References

Gati V. Aher, Rosa I. Arriaga, and Adam Tauman Kalai. Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. Out of One, Many: Using Language Models to Simulate Human Samples. *Political Analysis*, 31(3):337–351, 2023.

Ariel Flint Ashery, Luca Maria Aiello, and Andrea Baronchelli. Emergent Social Conventions and Collective Bias in LLM Populations. *Science Advances*, 11(20):eadu9368, 2025. doi: 10.1126/sciadv.adu9368.

Albert-László Barabási and Réka Albert. Emergence of Scaling in Random Networks. *Science*, 286(5439): 509–512, 1999. doi: 10.1126/science.286.5439.509.

Ronald S Burt. The social structure of competition. *Networks in the knowledge economy*, 13(2):57–91, 2003.

Weize Chen, Yusheng Su, Jingwei Zuo, Cheng Yang, Chenfei Yuan, Chi-Min Chan, Heyang Yu, Yaxi Lu, Yi-Hsin Hung, Chen Qian, et al. Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors. In *International Conference on Learning Representations*, volume 2024, pp. 20094–20136, 2024.

David A. Cotter, Joan M. Hermsen, Seth Ovadia, and Reeve Vanneman. The Glass Ceiling Effect. *Social Forces*, 80(2):655–681, 2001. doi: 10.1353/sof.2001.0091.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the*

41st International Conference on Machine Learning, volume 235 of *Proceedings of Machine Learning Research*, pp. 11733–11763. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/du24e.html.

Chen Gao, Xiaochong Lan, Zhihong Lu, Jinzhu Mao, Jinghua Piao, Huandong Wang, Depeng Jin, and Yong Li. S3: Social-network simulation system with large language model-empowered agents. *arXiv preprint arXiv:2307.14984*, 2023.

Mark S Granovetter. The strength of weak ties. *American journal of sociology*, 78(6):1360–1380, 1973.

Haoxiang Guan, Jiyan He, Liyang Fan, Zhenzhen Ren, Shaobin He, Xin Yu, Yuan Chen, Shuxin Zheng, Tie-Yan Liu, and Zhen Liu. Modeling earth-scale human-like societies with one billion agents. *arXiv preprint arXiv:2506.12078*, 2025.

Dongxin Guo, Jikun Wu, and Siu-Ming Yiu. Coalition formation in llm agent networks: Stability analysis and convergence guarantees. *arXiv preprint arXiv:2604.14386*, 2026.

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. Large Language Model Based Multi-Agents: A Survey of Progress and Challenges. *arXiv preprint arXiv:2402.01680*, 2024.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Steven Yau, Zijuan Lin, Liyang Zhou, et al. Metagpt: Meta programming for a multi-agent collaborative framework. In *International Conference on Learning Representations*, volume 2024, pp. 23247–23275, 2024.

John J Horton, Apostolos Filippas, and Benjamin S Manning. Large language models as simulated economic agents: What can we learn from homo silicus? Technical report, National Bureau of Economic Research, 2023.

Tianyu Hu, Zhen Tan, Song Wang, Huaizhi Qu, and Tianlong Chen. Multi-agent debate for llm judges with adaptive stability detection. *Advances in Neural Information Processing Systems*, 38:46504–46540, 2026.

Adit Jain and Vikram Krishnamurthy. Interacting Large Language Model Agents. Interpretable Models and Social Learning. *arXiv preprint arXiv:2411.01271*, 2024.

Adit Jain, Vikram Krishnamurthy, and Yiming Zhang. Collaborative QA using Interacting LLMs. Impact of Network Structure, Node Capability and Distributed Data. *arXiv preprint arXiv:2511.14098*, 2025a.

Adit Jain, Vikram Krishnamurthy, and Yiming Zhang. Information Diffusion and Preferential Attachment in a Network of Large Language Models. In *2025 IEEE 64th Conference on Decision and Control (CDC)*, pp. 180–185, 2025b. doi: 10.1109/CDC57313.2025.11312386.

Harold J. Kushner and George G. Yin. *Stochastic Approximation and Recursive Algorithms and Applications*, volume 35 of *Applications of Mathematics*. Springer, New York, 2 edition, 2003.

Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. Camel: Communicative agents for" mind" exploration of large language model society. *Advances in neural information processing systems*, 36:51991–52008, 2023.

Yunxuan Li, Yibing Du, Jiageng Zhang, Le Hou, Peter Grabowski, Yeqing Li, and Eugene Ie. Improving Multi-Agent Debate with Sparse Communication Topology. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 7281–7294, 2024. URL https://aclanthology.org/2024.findings-emnlp.427/.

Rui Luo, Buddhika Nettasinghe, and Vikram Krishnamurthy. Mutual Information Measure for Glass Ceiling Effect in Preferential Attachment Models. *IEEE Transactions on Computational Social Systems*, 11(6): 7778–7788, 2024.

Hachem Madmoun and Salem Lahlou. Communication enables cooperation in llm agents: A comparison with curriculum-based approaches. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 307–321, 2026.

Aliakbar Mehdizadeh and Martin Hilbert. Homophily-induced emergence of biased structures in llm-based multi-agent ai systems. *Social Network Analysis and Mining*, 15(1):1–25, 2025.

Robert K Merton. The matthew effect in science: The reward and communication systems of science are considered. *Science*, 159(3810):56–63, 1968.

Buddhika Nettasinghe, Nazanin Alipourfard, Stephen Iota, Vikram Krishnamurthy, and Kristina Lerman. Scale-free degree distributions, homophily and the glass ceiling effect in directed networks. *Journal of complex networks*, 10(2):cnac007, 2022.

Buddhika Nettasinghe, Nazanin Alipourfard, Vikram Krishnamurthy, and Kristina Lerman. Emergence of structural disparities in the web of scientific citations. In *Proceedings of the ACM Web Conference 2026*, pp. 1785–1796, 2026.

Marios Papachristou and Yuan Yuan. Network formation and dynamics among multi-llms. *PNAS nexus*, 4 (12):pgaf317, 2025.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pp. 1–22, 2023.

Jinghua Piao, Yuwei Yan, Jun Zhang, Nian Li, Junbo Yan, Xiaochong Lan, Zhihong Lu, Zhiheng Zheng, Jing Yi Wang, Di Zhou, Chen Gao, Fengli Xu, Fang Zhang, Ke Rong, Jun Su, and Yong Li. Agentsociety: Large-scale simulation of llm-driven generative agents advances understanding of human behaviors and society, 2026. URL `https://arxiv.org/abs/2502.08691`.

Derek De Solla Price. A general theory of bibliometric and other cumulative advantage processes. *Journal of the American society for Information science*, 27(5):292–306, 1976.

Chen Qian, Wei Liu, Hongzhang Liu, Nuo Chen, Yufan Dang, Jiahao Li, Cheng Yang, Weize Chen, Yusheng Su, Xin Cong, et al. Chatdev: Communicative agents for software development. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 15174–15186, 2024.

Chen Qian, Zihao Xie, Yifei Wang, Wei Liu, Kunlun Zhu, Hanchen Xia, Yufan Dang, Zhuoyun Du, Weize Chen, Cheng Yang, et al. Scaling large language model-based multi-agent collaboration. In *International Conference on Learning Representations*, volume 2025, pp. 41488–41505, 2025.

Philipp J Schneider, Lin Tian, and Marian-Andrei Rizoiu. Learning to make friends: Coaching llm agents toward emergent social ties. *arXiv preprint arXiv:2510.19299*, 2025.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, et al. Autogen: Enabling next-gen llm applications via multi-agent conversation. *arXiv preprint arXiv:2308.08155*, 2023.

Appendix

This appendix is organized into two main parts. Sec. A provides detailed proofs of the theoretical results introduced in Sec. 3 of the main paper. Sec. B presents additional experimental details, including dataset construction, agent interaction prompts, visualization of the LLM network formation process, and hyperparameter settings.

A Proofs for Section 3

Conditional expectation of the one-step communication increment. We first derive the conditional expectation of the one-step weighted communication increment of Type R. Recall that

$$\Delta^{t+1}(R) = \left(\Delta_{\text{in}}^{t+1}(R), \Delta_{\text{out}}^{t+1}(R)\right),$$

where the incoming and outgoing components record the communication mass newly assigned to Type R at time $t+1$. Conditional on G^t, we average over the three event types, the type of a newly introduced node when applicable, and the source- and target-side sampling decisions.

For the incoming component, Type R receives communication mass precisely when the selected target has Type R. Hence,

$$\begin{aligned}\mathbb{E}\left[\Delta_{\text{in}}^{t+1}(R) \mid G^t\right] &= p\,\pi_{\text{tgt},R}^t\left[r\mu_{RR} + (1-r)\mu_{RB}\right] \\ &\quad + qr\left[\pi_{\text{src},R}^t\mu_{RR} + \pi_{\text{src},B}^t\mu_{RB}\right] \\ &\quad + (1-p-q)\pi_{\text{tgt},R}^t\left[\pi_{\text{src},R}^t\mu_{RR} + \pi_{\text{src},B}^t\mu_{RB}\right].\end{aligned} \tag{13}$$

Here, μ_{ab} denotes the expected normalized communication contribution from a Type b source to a Type a target. Thus, the incoming expression contains only μ_{RR} and μ_{RB}, because the target type is fixed to R.

Similarly, Type R receives outgoing communication mass precisely when the selected source has Type R. Therefore,

$$\begin{aligned}\mathbb{E}\left[\Delta_{\text{out}}^{t+1}(R) \mid G^t\right] &= pr\left[\pi_{\text{tgt},R}^t\mu_{RR} + \pi_{\text{tgt},B}^t\mu_{BR}\right] \\ &\quad + q\,\pi_{\text{src},R}^t\left[r\mu_{RR} + (1-r)\mu_{BR}\right] \\ &\quad + (1-p-q)\pi_{\text{src},R}^t\left[\pi_{\text{tgt},R}^t\mu_{RR} + \pi_{\text{tgt},B}^t\mu_{BR}\right].\end{aligned} \tag{14}$$

Thus,

$$\mathbb{E}\left[\Delta^{t+1}(R) \mid G^t\right] = \left(\mathbb{E}\left[\Delta_{\text{in}}^{t+1}(R) \mid G^t\right], \mathbb{E}\left[\Delta_{\text{out}}^{t+1}(R) \mid G^t\right]\right).$$

Mean-field attachment probabilities and drift approximation. We next express the finite-time attachment probabilities in terms of the communication-prominence state Θ^t. Let

$$\nu = p + q$$

denote the probability that a new node is introduced at each macro-step. For $s \geq 1$, let

$$I^s = \mathbf{1}\{\text{a new node is introduced at time } s\}, \qquad J^s = \mathbf{1}\{\text{a new Type } R \text{ node is introduced at time } s\}.$$

Then

$$N^t = N^0 + \sum_{s=1}^{t} I^s, \qquad N^t(R) = N^0(R) + \sum_{s=1}^{t} J^s,$$

where

$$\mathbb{E}[I^s] = \nu, \qquad \mathbb{E}[J^s] = \nu r.$$

Since

$$D_{\text{in}}^t = D_{\text{out}}^t = t\mathbf{1}_d,$$

the attachment probabilities in (3) can be rewritten, for $t \geq 1$, as

$$\pi^t_{\mathrm{tgt},R} = \frac{\mathbf{1}_d^\top \theta^t_{\mathrm{in}} + \delta \frac{N^t(R)}{t}}{d + \delta \frac{N^t}{t}}, \qquad \pi^t_{\mathrm{src},R} = \frac{\mathbf{1}_d^\top \theta^t_{\mathrm{out}} + \xi \frac{N^t(R)}{t}}{d + \xi \frac{N^t}{t}}. \tag{15}$$

The Type B probabilities satisfy

$$\pi^t_{\mathrm{tgt},B} = 1 - \pi^t_{\mathrm{tgt},R}, \qquad \pi^t_{\mathrm{src},B} = 1 - \pi^t_{\mathrm{src},R}.$$

By Hoeffding's inequality, for every $\eta > 0$,

$$\mathbb{P}\left(\left| \frac{1}{t} \sum_{s=1}^{t} I^s - \nu \right| \geq \eta \right) \leq 2\exp(-2t\eta^2),$$

and

$$\mathbb{P}\left(\left| \frac{1}{t} \sum_{s=1}^{t} J^s - \nu r \right| \geq \eta \right) \leq 2\exp(-2t\eta^2).$$

Consequently, with probability at least $1 - 4\exp(-2t\eta^2)$,

$$\left| \frac{N^t}{t} - \nu \right| \leq \eta + \frac{N^0}{t}, \qquad \left| \frac{N^t(R)}{t} - \nu r \right| \leq \eta + \frac{N^0(R)}{t}. \tag{16}$$

Define the deterministic mean-field attachment probabilities by

$$\bar{\pi}_{\mathrm{tgt},R}(\theta_{\mathrm{in}}) = \frac{\mathbf{1}_d^\top \theta_{\mathrm{in}} + \nu r \delta}{d + \nu\delta}, \qquad \bar{\pi}_{\mathrm{src},R}(\theta_{\mathrm{out}}) = \frac{\mathbf{1}_d^\top \theta_{\mathrm{out}} + \nu r \xi}{d + \nu\xi},$$

and let

$$\bar{\pi}_{\mathrm{tgt},B}(\theta_{\mathrm{in}}) = 1 - \bar{\pi}_{\mathrm{tgt},R}(\theta_{\mathrm{in}}), \qquad \bar{\pi}_{\mathrm{src},B}(\theta_{\mathrm{out}}) = 1 - \bar{\pi}_{\mathrm{src},R}(\theta_{\mathrm{out}}).$$

Since the functions in (15) are Lipschitz in N^t/t and $N^t(R)/t$, there exists a constant $C_\pi > 0$ such that, on the event in (16),

$$\begin{aligned} &\max_{a \in \{R,B\}} \left| \pi^t_{\mathrm{tgt},a} - \bar{\pi}_{\mathrm{tgt},a}\left(\theta^t_{\mathrm{in}}\right) \right| \\ &\quad + \max_{a \in \{R,B\}} \left| \pi^t_{\mathrm{src},a} - \bar{\pi}_{\mathrm{src},a}\left(\theta^t_{\mathrm{out}}\right) \right| \leq C_\pi \left(\eta + \frac{N^0}{t} \right). \end{aligned} \tag{17}$$

We may therefore define the mean-field drift

$$F(\Theta) = \left(F_{\mathrm{in}}(\Theta), F_{\mathrm{out}}(\Theta) \right),$$

where

$$\begin{aligned} F_{\mathrm{in}}(\Theta) = {} & p\, \bar{\pi}_{\mathrm{tgt},R}(\theta_{\mathrm{in}}) \left[r\mu_{RR} + (1-r)\mu_{RB} \right] \\ & + qr \left[\bar{\pi}_{\mathrm{src},R}(\theta_{\mathrm{out}})\mu_{RR} + \bar{\pi}_{\mathrm{src},B}(\theta_{\mathrm{out}})\mu_{RB} \right] \\ & + (1-p-q)\bar{\pi}_{\mathrm{tgt},R}(\theta_{\mathrm{in}}) \left[\bar{\pi}_{\mathrm{src},R}(\theta_{\mathrm{out}})\mu_{RR} + \bar{\pi}_{\mathrm{src},B}(\theta_{\mathrm{out}})\mu_{RB} \right], \end{aligned} \tag{18}$$

and

$$\begin{aligned} F_{\mathrm{out}}(\Theta) = {} & pr \left[\bar{\pi}_{\mathrm{tgt},R}(\theta_{\mathrm{in}})\mu_{RR} + \bar{\pi}_{\mathrm{tgt},B}(\theta_{\mathrm{in}})\mu_{BR} \right] \\ & + q\, \bar{\pi}_{\mathrm{src},R}(\theta_{\mathrm{out}}) \left[r\mu_{RR} + (1-r)\mu_{BR} \right] \\ & + (1-p-q)\bar{\pi}_{\mathrm{src},R}(\theta_{\mathrm{out}}) \left[\bar{\pi}_{\mathrm{tgt},R}(\theta_{\mathrm{in}})\mu_{RR} + \bar{\pi}_{\mathrm{tgt},B}(\theta_{\mathrm{in}})\mu_{BR} \right]. \end{aligned} \tag{19}$$

Combining (13), (14), and (17), there exists a constant $C_F > 0$ such that

$$\left\| \mathbb{E}\left[\Delta^{t+1}(R) \mid G^t \right] - F(\Theta^t) \right\|_2 \leq C_F \left(\eta + \frac{N^0}{t} \right)$$

with probability at least $1 - 4\exp(-2t\eta^2)$.

In particular, choosing

$$\eta_t = \sqrt{\frac{2\log(t+1)}{t}},$$

the Borel–Cantelli lemma implies that

$$\mathbb{E}\big[\Delta^{t+1}(R) \mid G^t\big] = F(\Theta^t) + \varepsilon^{t+1},$$

where, almost surely,

$$\left\|\varepsilon^{t+1}\right\|_2 = O\left(\sqrt{\frac{\log(t+1)}{t}} + \frac{N^0}{t}\right). \tag{20}$$

Contraction of the mean-field map. We next give a sufficient condition under which the mean-field map F is a contraction on $[0,1]^{2d}$. Assume that the type-pair expected communication weights are uniformly bounded: there exists $\bar{\mu} < \infty$ such that

$$0 \le \mu_{ab,\ell} \le \bar{\mu}, \qquad a, b \in \{R, B\}, \quad \ell \in \{1, \dots, d\}.$$

Let

$$L_{\text{tgt}} = \frac{1}{d + \nu\delta}, \qquad L_{\text{src}} = \frac{1}{d + \nu\xi}.$$

By the definitions of the mean-field attachment probabilities, for every $j \in \{1, \dots, d\}$,

$$\left|\frac{\partial \bar{\pi}_{\text{tgt},a}}{\partial \theta_{\text{in},j}}\right| \le L_{\text{tgt}}, \qquad \left|\frac{\partial \bar{\pi}_{\text{src},a}}{\partial \theta_{\text{out},j}}\right| \le L_{\text{src}}, \qquad a \in \{R, B\},$$

whereas the cross derivatives with respect to the other prominence component are zero.

Consider any output coordinate $\ell \in \{1, \dots, d\}$. Differentiating (18) and (19), every derivative term contains one differentiated attachment probability and at most one remaining attachment probability, which lies in $[0,1]$. Moreover, derivatives of complementary probabilities yield differences such as $\mu_{RR,\ell} - \mu_{RB,\ell}$ and $\mu_{RR,\ell} - \mu_{BR,\ell}$, whose absolute values are bounded by $\bar{\mu}$. Since the coefficients associated with the three event types sum to at most one, we obtain

$$\left|\frac{\partial F_{\text{in},\ell}(\Theta)}{\partial \theta_{\text{in},j}}\right| \le \bar{\mu} L_{\text{tgt}}, \qquad \left|\frac{\partial F_{\text{in},\ell}(\Theta)}{\partial \theta_{\text{out},j}}\right| \le \bar{\mu} L_{\text{src}},$$

and likewise

$$\left|\frac{\partial F_{\text{out},\ell}(\Theta)}{\partial \theta_{\text{in},j}}\right| \le \bar{\mu} L_{\text{tgt}}, \qquad \left|\frac{\partial F_{\text{out},\ell}(\Theta)}{\partial \theta_{\text{out},j}}\right| \le \bar{\mu} L_{\text{src}}.$$

Therefore,

$$\|\nabla F_{\text{in},\ell}(\Theta)\|_2^2 \le d\bar{\mu}^2 \left(L_{\text{tgt}}^2 + L_{\text{src}}^2\right),$$

and the same bound holds for $\|\nabla F_{\text{out},\ell}(\Theta)\|_2^2$. Hence, the Jacobian $J_F(\Theta)$ satisfies

$$\begin{aligned}\|J_F(\Theta)\|_2 \le \|J_F(\Theta)\|_F &\le d\bar{\mu}\sqrt{2\left(L_{\text{tgt}}^2 + L_{\text{src}}^2\right)} \\ &= d\bar{\mu}\sqrt{2\left[\frac{1}{(d+\nu\delta)^2} + \frac{1}{(d+\nu\xi)^2}\right]} =: \rho_{\delta,\xi}.\end{aligned}$$

Thus, whenever

$$\rho_{\delta,\xi} < 1,$$

the mean-value theorem gives

$$\|F(\Theta) - F(\Theta')\|_2 \le \rho_{\delta,\xi}\|\Theta - \Theta'\|_2, \qquad \Theta, \Theta' \in [0,1]^{2d}.$$

Therefore, F is a contraction on $[0,1]^{2d}$. Since F maps $[0,1]^{2d}$ into itself, the Banach fixed-point theorem implies that there exists a unique $\Theta^\star \in [0,1]^{2d}$ satisfying

$$\Theta^\star = F(\Theta^\star).$$

B Experiment Details

This section provides additional implementation details for our experiments, including the construction of the QA datasets, the interaction procedure at each network iteration, and the hyperparameter settings used for different LLM families. These details are intended to make the experimental pipeline fully reproducible, from assigning partial contexts to agents to recording LLM-mediated communication intensities and updating the evolving network. All code and experiment scripts are available in the Anonymous GitHub repository.

B.1 Dataset Details

Synthetic Dataset Creation Pipeline

Collaborative QA. For collaborative QA, we directly use the *Fiction* dataset introduced by (Jain et al., 2025a). The dataset contains question–answer pairs about fictional facts grounded in narrative passages from Project Gutenberg books. For each question, we use five partial context snippets such that the relevant evidence is distributed across the snippets and cannot be fully recovered from a single context alone.

Multi-agent Debate. For multi-agent debate, we construct a synthetic dataset using Gemini-3.1-Pro. We first prepare a topic list covering everyday scientific, social, and technology-related questions. For each topic, the generator samples one conflict type from $\{\texttt{none}, \texttt{apparent_conflict}, \texttt{real_conflict}, \texttt{mixed}\}$, and generates a question, exactly five context snippets, and a gold answer. The generation prompt enforces that the five contexts are jointly necessary, that the answer cannot be recovered from any single snippet alone, and that the gold answer must synthesize evidence across all contexts. We validate each generated item by checking the required fields, the context IDs $C_1, \ldots, C_5$, non-empty context text, and a complete gold answer. For evaluation, we use Gemini-3.1-Pro as an LLM judge to compare each agent's final response against the gold answer and assign an answer-quality score in $[0, 1]$, where a higher score indicates closer semantic agreement and factual consistency with the reference answer.

Both datasets are represented in the same five-context format:

$$\{\texttt{question}, C_1, C_2, C_3, C_4, C_5, \texttt{Correct answer}\}.$$

During network construction, each agent is assigned only one context snippet; thus, successful task completion requires agents to exchange information and synthesize distributed evidence.

Dataset Examples

Tables 2 and 3 show one complete example from each dataset. Each example contains one question, five context snippets, and one gold answer. During the experiment, the five snippets are assigned to different agents, while the gold answer is used only for evaluation.

Table 2: One example from the Collaborative QA dataset.

Field	Content
Question	What was the place Spear praised as the northern fur trade's home, built on fountain head of gigantic water power?
C_1	I ROMANCE AND ADVENTURE HER FATHER THE FREE TRADER It was September 9, 1896. From sunrise to sunset through mist, sunshine, shower, and shadow we travelled, and the nearer we drew to our first destination, the wilder the country became, the more water-fowl we saw, and the more the river banks were marked with traces of big game. Here signs told us that three caribou had crossed the stream, there muddy water was still trickling into the hoofprint of a moose, and yonder a bear had been fishing. Finally, the day of our arrival dawned, and as I paddled, I spent much of the time dreaming of the adventure before me. As our beautiful birchen craft still sped on her way, the handsome bow parted the shimmering waters, and a passing breeze sent little running waves gurgling along her sides, while the splendour of the autumn sun was reflected on a far-reaching row of dazzling ripples that danced upon the water, making our voyageurs lower their eyes and the trader doze again. There was no other sign of life except an eagle soaring in and out among the fleecy clouds slowly passing overhead. All around was a panorama of enchanting forest.
C_2	My travelling companion was a "Free Trader," whose name was Spear—a tall, stoop-shouldered man with heavy eyebrows and shaggy, drooping moustache. The way we met was amusing. It happened in a certain frontier town. His first question was as to whether I was single. His second, as to whether my time was my own. Then he slowly looked me over from head to foot. He seemed to be measuring my stature and strength and to be noting the colour of my eyes and hair.
C_3	Narrowing his vision, he scrutinized me more carefully than before, for now he seemed to be reading my character—if not my soul. Then, smiling, he blurted out:
C_4	"Come, be my guest for a couple of weeks. Will you?" I laughed.
C_5	He frowned. But on realizing that my mirth was caused only by surprise, he smiled again and let flow a vivid description of a place he called Spearhead. It was the home of the northern fur trade. It was the centre of a great timber region. It was the heart of a vast fertile belt that was rapidly becoming the greatest of all farming districts. It was built on the fountain head of gigantic water power.
Answer	Spearhead.

Table 3: One example from the Multi-agent Debate dataset.

Field	Content
Question	Can consistent use of language learning apps lead to true fluency, or are they insufficient for achieving advanced proficiency?
C_1	Many language learning apps offer structured curricula for grammar, vocabulary acquisition, and pronunciation practice through interactive exercises.
C_2	Critics argue that solely relying on language apps cannot lead to true fluency, as they often lack opportunities for spontaneous, unscripted conversational practice with native speakers.
C_3	True language fluency encompasses not only linguistic accuracy but also pragmatic competence, cultural nuance, and the ability to adapt to diverse real-world communication scenarios.
C_4	The progress achieved through language apps is highly dependent on the learner's self-discipline, consistent engagement with the material, and active participation in exercises.
C_5	While apps are excellent for building foundational knowledge and drilling basic skills, achieving advanced proficiency typically requires integration with immersive experiences, direct native speaker interaction, and advanced study resources beyond app content.
Answer	Consistent use of language learning apps can be highly effective for building foundational grammar, vocabulary, and pronunciation skills, especially for beginners. However, solely relying on these apps is generally insufficient for achieving true fluency. True fluency encompasses a broader range of abilities, including pragmatic competence, cultural nuance, and the capability for spontaneous, unscripted communication in diverse real-world scenarios, which apps often lack. While apps are excellent tools for structured learning and basic skill drilling, achieving advanced proficiency requires supplementing app usage with consistent self-discipline, active engagement, and integration with immersive experiences, direct native speaker interaction, and advanced study resources. Therefore, apps serve as valuable components but need to be part of a broader, more diversified language learning strategy to reach comprehensive fluency.

B.2 Specific Prompt

Source proposal prompt. For a candidate directed interaction $u \to v$, the source agent u receives the identifiers, roles, and model specifications of both agents, together with its assigned context. It is then prompted as follows:

```
You are the initiating agent in a collaborative QA network.

Source agent:
- id: {sender_id}
- role: {sender_role}
- model: {sender_model}

Target agent:
- id: {receiver_id}
- role: {receiver_role}
- model: {receiver_model}

Your assigned context:
{context}

Task:
Write a convincing paragraph to persuade the target agent to communicate and
collaborate with you on this QA task. Introduce yourself and focus on being
clear, motivating, and persuasive in your argument as to why this
collaboration is worthwhile. Do NOT return JSON or lists, only write one
persuasive paragraph.
```

The resulting pitch communicates the source agent's perceived usefulness, including the unique evidence and reasoning contribution that it can provide to the target agent.

Target acceptance prompt. The target agent v receives the source context, source pitch, and proposed communication intensities. It then decides whether to accept the interaction and, if accepted, specifies the granted communication weight along three dimensions: token exchange, interaction frequency, and reasoning effort.

```
You are the receiving agent in a collaborative QA network.

Target agent:
- id: {receiver_id}
- role: {receiver_role}
- model: {receiver_model}

Source agent:
- id: {sender_id}
- role: {sender_role}
- model: {sender_model}

Source context:
{context}

Source proposal (requested communication intensity):
- token_exchange: {proposed_tokens}
- interaction_frequency: {proposed_rounds}
- reasoning_effort: {proposed_reasoning}

Pitch from source:
{pitch}

Task:
Decide whether to accept this interaction proposal. If you reject, set accepted
to false and all three intensity keys to zero. If you accept, set accepted to
```

```
true and return the exact intensities you grant, each of which may be at or
below the proposed level.

IMPORTANT --- score each dimension independently when accepted:
- token_exchange, interaction_round, and reasoning_effort are three
  separate axes.
- Do NOT copy one score to another or keep fixed ratios across dimensions.
- Use the full range on each axis when appropriate.

Return ONLY valid JSON with exactly these keys:
{
  "accepted": <bool>,
  "token_exchange": <int>,
  "interaction_frequency": <int>,
  "reasoning_effort": <int>
}

Base your decision solely on the agent you judge to be more reliable,
considering the persuasiveness, representativeness, and confidence of its
pitch. Do not include any other information.
```

To illustrate this dynamic network formation process, we visualize representative network snapshots in Fig. 6. The figure shows how the LLM interaction network grows over time as new agents and directed communication edges are added. Node colors indicate the two agent types, Type R and Type B, and the snapshots at different timesteps show the gradual emergence of a denser and more heterogeneous communication structure. Below each network snapshot, we further plot the corresponding in-degree and out-degree distributions. These degree distributions exhibit a slowly decaying, heavy-tailed pattern rather than the sharply concentrated, exponentially decaying distribution expected in an Erdős–Rényi random graph. This indicates that the formed LLM interaction network is not well described by an Erdős–Rényi model, but instead develops heterogeneous connectivity with a small number of highly connected agents.

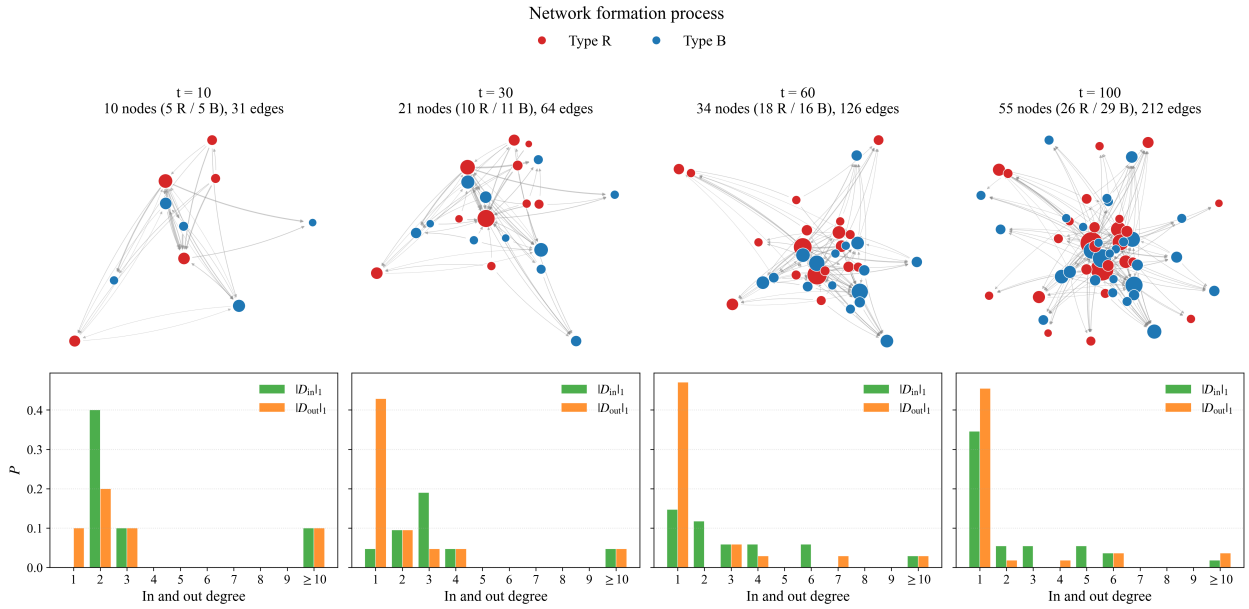

Figure 6: Visualization of the LLM network formation process. The snapshots show the evolving directed interaction network at different timesteps, where red and blue nodes correspond to Type R and Type B agents, respectively. Below each network snapshot, we plot the corresponding in-degree and out-degree distributions. The slowly decaying, heavy-tailed degree distributions indicate that the formed LLM interaction network deviates from an Erdős–R'enyi random graph and exhibits heterogeneous connectivity with a small number of highly connected agents.

B.3 Utility-Model Training Details

For each model pair, we train the type-specific pairwise utility model described in Sec. 5.1 on the collected source–target interaction samples. The context embeddings and network-status representations are precomputed and kept fixed throughout optimization; only the type-specific projection matrices

$$\mathcal{P} = \{Q_R, Q_B, K_R, K_B, V_R, V_B\}$$

are optimized. We use an attention dimension of $r = 32$, three output dimensions corresponding to token budget, additional interaction rounds, and reasoning effort, and a six-dimensional source-status vector consisting of the incoming and outgoing normalized communication states for these three resources.

We optimize the utility model using AdamW with learning rate 2×10^{-3} and weight decay 10^{-5}. Gradients are clipped to have maximum ℓ_2-norm 1.0. We train for 10 epochs. Since the predicted communication weights are normalized over all candidate interactions within a network-formation event, each mini-batch contains complete events rather than independently sampled pairwise interactions. We use mini-batches of 8 complete events. The final trained model is used to estimate the type-pair expected utilities $\widehat{\mu}_{RR}, \widehat{\mu}_{RB}, \widehat{\mu}_{BR}, \widehat{\mu}_{BB}$ from the held-out interaction samples. All experiments are implemented in PyTorch. Context embeddings are precomputed using the frozen pretrained `sentence-transformers/all-MiniLM-L6-v2` encoder, which produces 384-dimensional normalized sentence embeddings.

B.4 Performance Metrics for Mean-field Dynamics ODE

In this section, we describe the metrics used to evaluate the predictive performance of the mean-field ODE, including the relative mean squared error (MSE), relative Bias, the Ljung–Box p-value, and MaxACF_{10}.

For each collaborative QA case c, we initialize the mean-field ODE from the corresponding empirical initial state and numerically solve it for $T = 100$ network-formation time steps. Let $\widehat{\Theta}_c^t \in \mathbb{R}^6$ denote the resulting ODE prediction at time t, and let $\Theta_c^t \in \mathbb{R}^6$ denote the empirical type-level prominence trajectory observed from the corresponding multi-agent simulation. We define the vector-valued residual by

$$\rho_c^t = \widehat{\Theta}_c^t - \Theta_c^t.$$

We evaluate trajectory accuracy using the case-level relative mean squared error

$$\mathrm{MSE}_c = \frac{1}{6T} \sum_{t=1}^{T} \sum_{j=1}^{6} \left(\frac{\rho_{c,j}^t}{\left|\Theta_{c,j}^t\right| + \epsilon} \right)^2$$

and relative bias

$$\mathrm{Bias}_c = \frac{1}{6T} \sum_{t=1}^{T} \sum_{j=1}^{6} \frac{\left|\rho_{c,j}^t\right|}{\left|\Theta_{c,j}^t\right| + \epsilon},$$

where $\epsilon = 10^{-4}$ is a small numerical constant that avoids division by zero. We report their averages across all evaluation cases:

$$\mathrm{MSE} = \frac{1}{|\mathcal{C}|} \sum_{c\in\mathcal{C}} \mathrm{MSE}_c, \qquad \mathrm{Bias} = \frac{1}{|\mathcal{C}|} \sum_{c\in\mathcal{C}} \mathrm{Bias}_c.$$

Thus, MSE measures the average squared prediction discrepancy relative to the magnitude of the empirical prominence trajectory across all coordinates, time steps, and evaluation cases. Bias measures the average absolute relative prediction residual. Smaller MSE and Bias indicate closer agreement between the ODE prediction and the empirical trajectory. · To assess whether the mean-field ODE leaves systematic temporal dependence in the residuals, we apply a coordinate-wise Ljung–Box diagnostic to the residual sequence $\{\rho_c^t\}_{t=1}^T$. For each case c, residual dimension $j \in \{1, \ldots, 6\}$, and lag ℓ, define

$$\bar{\rho}_{c,j} = \frac{1}{T} \sum_{t=1}^{T} \rho_{c,j}^t, \qquad \widehat{r}_{c,j}(\ell) = \frac{\sum_{t=\ell+1}^{T} \left(\rho_{c,j}^t - \bar{\rho}_{c,j}\right)\left(\rho_{c,j}^{t-\ell} - \bar{\rho}_{c,j}\right)}{\sum_{t=1}^{T} \left(\rho_{c,j}^t - \bar{\rho}_{c,j}\right)^2}.$$

The corresponding Ljung–Box statistic and coordinate-wise p-value at lag 10 are

$$Q_{c,j,10} = T(T+2)\sum_{\ell=1}^{10} \frac{\widehat{r}_{c,j}(\ell)^2}{T-\ell}, \qquad p_{c,j,10} = 1 - F_{\chi^2_{10}}\left(Q_{c,j,10}\right),$$

where $F_{\chi^2_{10}}$ denotes the cumulative distribution function of the χ^2_{10} distribution.

To obtain a conservative residual diagnostic for each case, we retain the smallest coordinate-wise p-value:

$$\mathrm{LBP}_{c,10} = \min_{1\le j\le 6} p_{c,j,10}.$$

We then report the average across all evaluation cases:

$$\mathrm{LBP}_{10} = \frac{1}{|\mathcal{C}|}\sum_{c\in\mathcal{C}} \mathrm{LBP}_{c,10}.$$

Thus, LBP_{10} summarizes the residual coordinate with the strongest evidence of serial dependence within each case. Larger values indicate weaker evidence of residual autocorrelation in the worst-performing coordinate, on average across evaluation cases.

Finally, we measure the magnitude of the strongest remaining local residual dependence using

$$\mathrm{MaxACF}_{c,10} = \max_{1\le j\le 6}\max_{1\le \ell\le 10} \left|\widehat{r}_{c,j}(\ell)\right|.$$

We then average this worst-case autocorrelation magnitude across evaluation cases:

$$\mathrm{MaxACF}_{10} = \frac{1}{|\mathcal{C}|}\sum_{c\in\mathcal{C}} \mathrm{MaxACF}_{c,10}.$$

Thus, MaxACF_{10} captures the largest absolute residual autocorrelation among all prominence coordinates and the first ten lags within each case. Lower values indicate weaker residual temporal dependence and therefore better agreement between the ODE dynamics and the observed trajectory evolution.

B.5 Hyperparameter Setting for LLMs

Table 4 summarizes the network-level hyperparameters. Unless otherwise stated, these are the default values used in the implementation.

Table 4: Network-level hyperparameter settings.

Hyperparameter	Value	Description
T	100 by default	Number of macrosteps in one run
p	0.25	Probability of Event 1
q	0.25	Probability of Event 2
$1-p-q$	0.50	Probability of Event 3
r	0.50	Probability that a newly born node has type R
M_1	3	Number of trials in Event 1
M_2	3	Number of trials in Event 2
M_3	3	Number of trials in Event 3
d	3	Number of communication-intensity dimensions
δ	1.5	Incoming attachment smoothing constant
ξ	1.5	Outgoing attachment smoothing constant
Initial nodes	2 per type	Seed graph size
Max context length	2500 tokens	Maximum snippet length assigned to an agent

Table 5 summarizes the LLMs used in our experiments. API keys are omitted from the paper. The decoding and generation settings are listed below the table.

Table 5: LLMs used in the experiments.

Model family	Model name
GPT	`gpt-4.1`, `gpt-4.1-mini`
Gemini	`gemini-2.5-flash-Lite`, `gemini-3.5-flash`
LLaMA	`Llama-3.1-8B-Instruct`, `Llama-3.1-8B-Instruct`
Qwen	`Qwen3-4B-Instruct-2507`
Mistral	`Ministral-3-8B-Instruct-2512`
Grok	`grok-4.20-0309-reasoning`

B.6 Additional Results

We report two additional experiments that further assess the robustness of our main findings. First, we present the cross-family comparison between open-source and proprietary LLM agents in this appendix. This setting examines whether communication asymmetries persist when agent types differ not only in model scale but also in model family and training provenance. Second, we repeat the prompt-induced heterogeneity experiments with a GPT-based agent population, testing whether the role-dependent communication patterns identified in the main text extend beyond the Gemini base model.

B.6.1 Cross-family heterogeneity

We further examine whether capability-aligned communication dominance persists across model families with different architectures, training data, and deployment settings. Fig. 7 compares three proprietary–open-source model pairs: GPT versus LLaMA, Gemini versus Qwen, and Grok versus Mistral. In each setting, we measure the evolution of the type-level communication influence ratios and the resulting final agent-level influence distributions. Across these cross-family comparisons, the proprietary model type generally attains a higher communication influence ratio and occupies more central communication positions than its open-source counterpart. These results show that capability-aligned preferential attachment is not limited to model-scale differences within a single family, but can also emerge under broader cross-family heterogeneity.

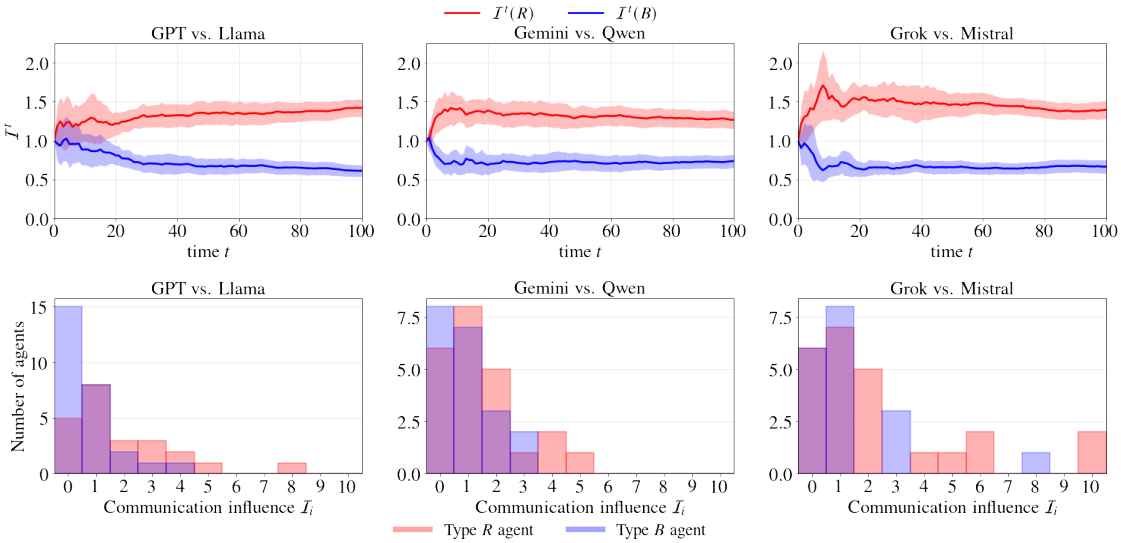

Figure 7: Capability-aligned communication dominance under cross-family heterogeneity. We compare three proprietary–open-source model pairs: GPT versus LLaMA, Gemini versus Qwen, and Grok versus Mistral. Within each pair, R denotes the proprietary model and B denotes the open-source model. The figure reports the type-level communication influence ratios $\mathcal{I}^t(R)$ and $\mathcal{I}^t(B)$, defined in (10), together with the corresponding final agent-level influence distributions. Across the three model pairs, R-type agents generally attain higher communication prominence and occupy more central network positions than B-type agents, demonstrating capability-aligned preferential attachment across model families.

B.6.2 Prompt-induced heterogeneity with GPT agents

We further test whether the prompt-induced glass-ceiling effect observed in the main text also arises in GPT-based agent populations. Figure 8 considers three settings in which all agents use the same GPT base model and differ only in their system-prompt roles: *Explainer* versus *Verifier*, *Comprehensive Analyst* versus *Selective Analyst*, and *Proponent* versus *Opponent*. Despite identical model weights within each setting, the first role in each pair generally attains a higher communication influence over time and has a more right-skewed final influence distribution. Thus, prompt-defined interaction roles alone can induce persistent structural advantages among GPT agents, providing additional evidence that the glass-ceiling effect does not require underlying differences in base-model capability.

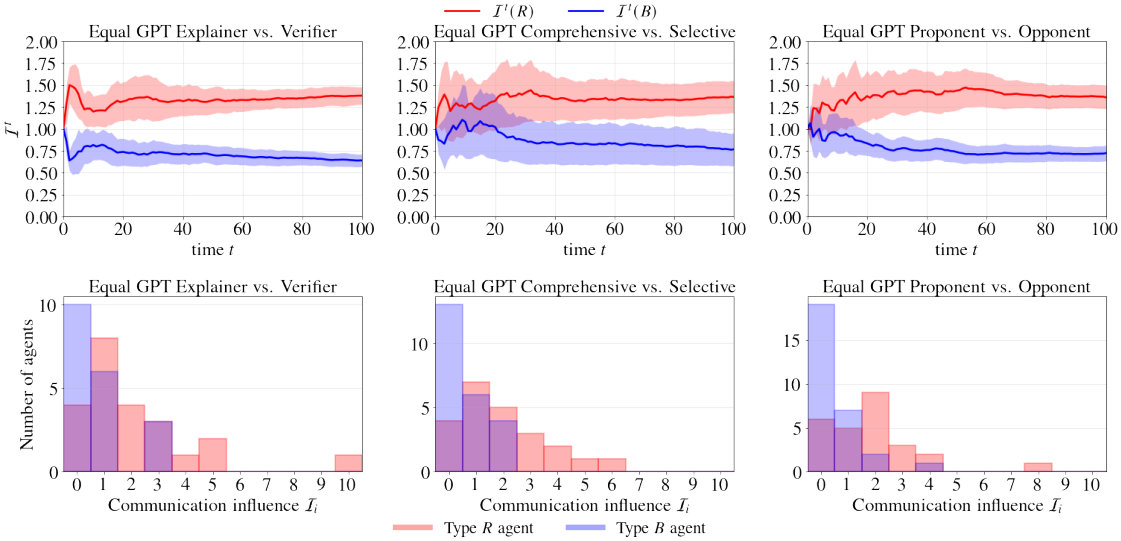

Figure 8: Capability-misaligned communication dominance under prompt-induced glass-ceiling effects in equal-GPT agents. All agents within each column use the same GPT base model and differ only in their assigned system-prompt roles. From left to right, we compare *Explainer* versus *Verifier*, *Comprehensive Analyst* versus *Selective Analyst*, and *Proponent* versus *Opponent*. In each pair, Type R denotes the Explainer, Comprehensive Analyst, or Proponent role, respectively, while Type B denotes the corresponding Verifier, Selective Analyst, or Opponent role. The upper panels show the type-level communication influence ratios $\mathcal{I}^t(R)$ and $\mathcal{I}^t(B)$, defined in (10), and the lower panels show the final agent-level communication influence distributions. Across all three role pairs, Type R agents generally accumulate greater communication influence and exhibit a heavier right tail in the final distribution, showing that prompt-defined roles can induce persistent communication asymmetry even when all agents have identical GPT model weights.