What are the documented emergent behaviors in multi-agent LLM dialogues?

LLMs in multi-agent dialogues demonstrate emergent social behaviors including hierarchical power dynamics, cooperative leadership, personality development, and network-level phenomena such as clustering and echo chambers.

Abstract

Multi-agent dialogues among large language models produce a range of spontaneous social behaviors. Several experiments report that agents form social conventions and display role-based structures. For example, simulations in hierarchical settings (e.g., guard versus prisoner) document power dynamics, persuasion, and even anti-social behavior, while cooperative scenarios yield leadership, conflict resolution, and Theory of Mind–like reasoning. Studies detail that agents—whether initialized with predefined roles or through emergent traits—develop distinct personalities and adopt communication strategies that include debate, reflection, hashtags, and occasional hallucinations. In larger networks, documented behaviors further include clustering, polarization, and echo chamber effects, as well as instances where minority influence prompts tipping points and collective biases.

Paper search

Using your research question "What are the documented emergent behaviors in multi-agent LLM dialogues?", we searched across over 126 million academic papers from the Semantic Scholar corpus. We retrieved the 50 papers most relevant to the query.

Screening

We screened in papers that met these criteria:

- **Multiple LLM Agents**: Does the study examine interactions between two or more autonomous LLM agents (not just human-LLM interactions)?
- **Behavioral Documentation**: Does the study document observable behaviors or patterns using defined metrics or frameworks for measuring agent interactions?
- **Empirical Evidence**: Does the study present empirical research with quantitative or qualitative analysis of actual agent interactions?
- **LLM-Based Agents**: Does the study specifically examine multi-agent systems that use Large Language Models (not other types of AI)?
- **Agent Interaction Focus**: Does the study go beyond single-agent behavior or human-LLM interactions to examine agent-to-agent dynamics?
- **Research Type**: Does the study include empirical observations rather than being purely theoretical or conceptual?

We considered all screening questions together and made a holistic judgement about whether to screen in each paper.

Data extraction

We asked a large language model to extract each data column below from each paper. We gave the model the extraction instructions shown below for each column.

- **Study Design Type**:

  Identify the primary type of study design used:

- Multi-agent LLM simulation
- Computational modeling
- Experimental scenario-based study

Locate this information in the methods or introduction section. If multiple design elements are present, list them in order of prominence. If the design is not explicitly stated, infer from the methodology description.

If unsure, default to "Multi-agent LLM simulation" given the nature of the research.

- **Agent Characteristics and Setup**:

  Extract detailed information about the LLM agents:

- Number of agents in the simulation
- LLM models used (specific names/versions)
- How agents were initialized (predefined vs. emergent characteristics)
- Communication mechanism (e.g., natural language exchange, specific protocol)

Look in methods section for precise details. If exact numbers are not provided, note the range or approximate number. If multiple agent types are used, list all with their specific characteristics.

- **Interaction Scenario and Context**:

  Describe the specific interaction scenario:

- Social context (e.g., hierarchical setting, cooperative environment)
- Specific roles assigned to agents (if any)
- Goal or objective of the interaction

Extract from methods or results sections. If multiple scenarios were tested, list each with its specific parameters. Capture the key contextual elements that define the interaction space.

- **Emergent Behaviors Observed**:

  Systematically list and categorize emergent behaviors:

- Social conventions
- Personality traits
- Communication patterns
- Cooperative or anti-social behaviors

Prioritize behaviors directly mentioned as "emergent" or spontaneously developed. Extract from results section. Use direct quotes where possible to capture the nuanced description of behaviors.

If behaviors are quantified, include measurement details and metrics.

- **Significant Findings on Agent Interactions**:

  Summarize the most significant findings related to multi-agent interactions:

- Key mechanisms of behavior emergence
- Unexpected interaction patterns
- Insights into collective behavior

Extract from results and discussion sections. Focus on findings that directly address how behaviors spontaneously develop or change through interactions.

Prioritize findings that provide novel insights into LLM agent dynamics.

Results

Characteristics of Included Studies

Study	Study Design	Agent Configuration	Interaction Context	Key Emergent Behaviors	Full text retrieved
Campedelli et al., 2024	Multi-agent large language model (LLM) simulation, Experimental scenario-based study	2 agents (guard, prisoner); Llama3, Orca2, Command-r, Mixtral, Mistral2; predefined roles and personalities; natural language exchange	Hierarchical (prison) setting; guard and prisoner roles; goal-oriented (escape/yard time vs. order)	Power dynamics, emergent persuasion, anti-social behavior, role-driven personality effects	Yes
Li et al., "Assessing Collective Reasoning"	Experimental scenario-based study	Number of agents not specified; GPT-4.1 and five other large language models; agent initialization and communication not detailed	Distributed information, Hidden Profile tasks; focus on collective reasoning and cooperation-contradiction	Over-coordination, social desirability bias, group convergence/contradiction dynamics	No

Study	Study Design	Agent Configuration	Interaction Context	Key Emergent Behaviors	Full text retrieved
Li et al., 2023	Multi-agent large language model simulation	3 agents (Alpha, Bravo, Charlie); gpt-3.5-turbo-0301, gpt-4-0314; predefined context; natural language exchange	Cooperative text game; search and rescue; specialist roles	Emergent collaboration, leadership, Theory of Mind (the ability to reason about others' mental states), conflict resolution	Yes
Takata et al., 2024a	Multi-agent large language model simulation	10 agents; Llama-2-7b-chat-hf; emergent traits; natural language exchange	Cooperative group simulation; 2D grid; no predefined roles	Social norms, personality differentiation, hashtags (emergent labeling), hallucinations (generation of non-factual content), collective emotions	Yes
Takata et al., 2024b	Multi-agent large language model simulation, Computational modeling	Number of agents and large language model not specified; emergent traits; natural language exchange	Cooperative group simulation; focus on individuality and social norm emergence	Spontaneous social norms, evolving personalities, communication diversity	No
Ashery et al., 2024a	Multi-agent large language model simulation, Computational modeling	24 agents; Llama 3 70B 12, Llama-3.1 70B, Llama 2 70b, Claude Sonnet 3.5; emergent traits; structured text-based moves	Cooperative, pairwise convention formation; minority influence scenarios	Social conventions, collective biases, minority-driven change, tipping points	Yes

Study	Study Design	Agent Configuration	Interaction Context	Key Emergent Behaviors	Full text retrieved
Zhang et al., 2023	Multi-agent large language model simulation	3 agents per society; ChatGPT (gpt-3.5-turbo-1106); predefined traits; debate/reflection protocols	Societies with agent traits and thinking patterns; collaborative problem-solving	Conformity, consensus, debate/reflection patterns, collaboration efficiency	Yes
Willis et al., 2025	Multi-agent large language model simulation, Computational modeling	12 agents; ChatGPT-4o, Claude 3.5 Sonnet; predefined attitudes; strategies as code	Iterated Prisoner's Dilemma; competitive, strategic roles	Aggression, cooperation, large language model bias, strategy adaptation	Yes
Piao et al., 2025	Multi-agent large language model simulation	1000 agents (main), 100 (mechanism); ChatGPT (GPT-3.5), ChatGLM, Llama-3, GPT-4o; random initialization; natural language exchange	Large-scale network; political opinion dynamics; self-expression, communication, opinion update	Polarization, clustering, echo chambers (self-reinforcing communication groups), backfire effect (strengthening of original beliefs after counter-arguments), intervention effects	Yes
Ashery et al., 2024b	Multi-agent large language model simulation	Number of agents and large language model not specified; emergent traits; natural language exchange	Decentralized population; minority influence; social convention formation	Social conventions, collective biases, minority-driven change	No

Agent configuration:

- Small agent groups (3 agents):3 studies
- Medium groups (4–24 agents):3 studies
- Large-scale groups (>100 agents):1 study
- No mention found of agent number:3 studies
- Large language model types:
  - Llama-family models: 6 studies
  - GPT-3.5: 3 studies
  - GPT-4/4.1/4o: 4 studies
  - Claude: 3 studies
  - Mixtral/Mistral and ChatGLM: 1 study each
  - No mention found of model: 3 studies
- Agent initialization:
  - Predefined roles or traits: 4 studies
  - Emergent traits: 4 studies
  - Random initialization: 1 study
  - No mention found: 1 study

Interaction context:

- Cooperative group contexts:5 studies
- Competitive contexts:2 studies
- Hierarchical context:1 study
- Decentralized, network/large-scale, and collaborative problem-solving contexts:1 study each
- Minority influence focus:2 studies
- Specialist roles, pairwise interaction, political opinion dynamics:1 study each

Key emergent behaviors:

- Social norms or conventions:4 studies
- Collective biases and minority-driven change:2 studies each
- Collaboration or cooperation:3 studies
- Leadership, Theory of Mind, conflict resolution:1 study each
- Power dynamics, persuasion, anti-social behavior:1 study
- Over-coordination, social desirability bias, group convergence, contradiction:1 study
- Personality differentiation or evolving personalities:2 studies
- Hashtags, hallucinations, collective emotions:1 study
- Aggression, strategy adaptation, large language model bias:1 study
- Polarization, clustering, echo chambers, backfire effect, intervention effects:1 study
- Conformity, consensus, debate/reflection, collaboration efficiency:1 study

We did not find mention of the number of agents, large language model type, or agent initialization details for 3 studies each. Most studies reported multiple emergent behaviors, with social norms/conventions and collaboration/cooperation being the most common.

Thematic Analysis

Social Organization and Hierarchy

- Spontaneous social structure formation:Several studies report the emergence of social structures and role adoption in multi-agent large language model dialogues.
- Hierarchical settings:In simulated prison environments, power dynamics and role assignment (such as guard versus prisoner) are reported to drive persuasion and anti-social behaviors, with the guard's personality influencing outcomes.
- Decentralized/cooperative contexts:Leadership and role differentiation can emerge organically, as in collaborative text games where agents voluntarily assume leadership or specialist roles.
- Minority influence:Convention formation studies report that small groups can drive social change in larger populations, with observed tipping points.

Individual Agent Development

- Emergence of personality and differentiation:Studies with initially homogeneous agents report the development of distinct personality traits and emotional expressions through interaction.
- Role of memory and context:Memory and context retention are described as facilitating individuality, with agents differentiating based on communication history and spatial positioning.
- Self-inconsistency and regulation:Some studies report self-inconsistency (agents acting inconsistently with prior behavior) and its reduction through self-regulation, indicating emergent self-monitoring traits.

Collective Behavioral Patterns

- Social norms and conventions:Social conventions and collective biases are reported in several studies, including those where individual agents are described as unbiased.
- Communication patterns:Conformity, consensus, debate, and reflection are observed, with debate generally facilitating consensus and reflection increasing the difficulty of consensus.
- Polarization and clustering:Large-scale simulations report human-like polarization, clustering, and echo chamber effects. Interventions, such as neutral influencers, are described as reducing polarization.
- Social network formation:The formation of social networks with homophilic clustering (agents grouping with similar others) is also documented.

Emergent Behavior Type	Triggering Conditions	Consistency Across Studies	Implications
Social conventions/norms	Local interactions, repeated communication, role assignment	High: observed in most studies	Large language model agents can autonomously develop group norms, with potential for both alignment and bias
Personality traits/individuality	Homogeneous initialization, ongoing interaction	Moderate: especially in studies with emergent traits	Individual differentiation is possible even without predefined traits

Emergent Behavior Type	Triggering Conditions	Consistency Across Studies	Implications
Collective biases/polarization	Repeated interaction, information asymmetry, network structure	High: especially in large-scale or decentralized settings	Risk of echo chambers and polarization in large language model agent societies
Cooperative/anti-social behaviors	Role assignment, strategic context, communication protocol	Variable: context-dependent	Both cooperation and anti-social behaviors can emerge, influenced by roles and model biases
Communication patterns	Protocol design, agent traits, task structure	High: debate, reflection, hashtags, hallucinations	Communication strategies adapt to sustain interaction and consensus

Patterns in consistency and triggering conditions:

- High consistency:Social conventions/norms, collective biases/polarization, and communication patterns
- Moderate consistency:Personality traits/individuality, especially in studies focused on emergent traits
- Variable consistency:Cooperative/anti-social behaviors, depending on context

Triggering conditions:

- Repeated or local interaction:Social conventions/norms, collective biases/polarization
- Role assignment:Social conventions/norms, cooperative/anti-social behaviors
- Communication protocol/design:Cooperative/anti-social behaviors, communication patterns
- Agent traits/homogeneous initialization:Personality traits/individuality, communication patterns
- Strategic context, information asymmetry, network structure, task structure:Each relevant to one or more behavior types

We did not find any emergent behavior types that were reported as having low or absent consistency across studies.

Interaction Mechanisms

Communication Patterns

- Language use and adaptation:Agents use natural language exchange, debate, reflection, and generate hashtags and hallucinations to sustain communication and increase diversity.
- Information sharing protocols:Structured prompts and multi-round collaboration are reported to influence the efficiency and nature of group decision-making.
- Implicit signaling:Communication strategies can serve as implicit signals of intent or personality, as seen in strategic games.

Cognitive Capabilities

- Theory of Mind-like reasoning:Agents are reported to estimate their own and others' mental states, which is described as facilitating collaboration and conflict resolution.
- Decision-making processes:Decision-making is shaped by both individual and group-level dynamics, with representation of agents' beliefs enhancing Theory of Mind and collective performance.
- Strategy adaptation and model bias:In competitive or strategic contexts, agents adapt their strategies based on observed behaviors and outcomes, with biases inherent to the large language model influencing the balance between cooperation and aggression.

Summary:

- Communication and cognitive mechanisms are reported as central to the emergence of complex social behaviors in multi-agent large language model dialogues.
- The studies report adaptability in language use and the development of Theory of Mind-like capabilities, indicating that large language model agents may approximate some aspects of human social interaction in specific contexts, though the extent and reliability of these capabilities are described as context-dependent.

References

Ariel Flint Ashery, L. Aiello, and Andrea Baronchelli. "Emergent Social Conventions and Collective Bias in LLM Populations." *Science Advances*, 2024.

———. "The Dynamics of Social Conventions in LLM Populations: Spontaneous Emergence, Collective Biases and Tipping Points." *arXiv.org*, 2024.

G. Campedelli, Nicolò Penzo, Massimo Stefan, Roberto Dessì, Marco Guerini, Bruno Lepri, and Jacopo Staiano. "I Want to Break Free! Persuasion and Anti-Social Behavior of LLMs in Multi-Agent Settings with Social Hierarchy." *arXiv.org*, 2024.

Huao Li, Yu Quan Chong, Simon Stepputtis, Joseph Campbell, Dana Hughes, Michael Lewis, and Katia P. Sycara. "Theory of Mind for Multi-Agent Collaboration via Large Language Models." *Conference on Empirical Methods in Natural Language Processing*, 2023.

J. Piao, Zhihong Lu, Chen Gao, Fengli Xu, Fernando P. Santos, Yong Li, and James Evans. "Emergence of Human-Like Polarization Among Large Language Model Agents." *arXiv.org*, 2025.

Jintian Zhang, Xin Xu, Ruibo Liu, and Shumin Deng. "Exploring Collaboration Mechanisms for LLM Agents: A Social Psychology View." *arXiv.org*, 2023.

Richard Willis, Yali Du, Joel Z. Leibo, and Michael Luck. "Will Systems of LLM Agents Cooperate: An Investigation into a Social Dilemma." *arXiv.org*, 2025.

Ryosuke Takata, A. Masumori, and Takashi Ikegami. "Spontaneous Emergence of Agent Individuality Through Social Interactions in Large Language Model-Based Communities." *Entropy*, 2024.

———. "Spontaneous Emergence of Agent Individuality Through Social Interactions in LLM-Based Communities." *arXiv.org*, 2024.

Yuxuan Li, Aoi Naito, and Hirokazu Shirado. "Assessing Collective Reasoning in Multi-Agent LLMs via Hidden Profile Tasks," 2025.