Broadcasting, Not Conversing: What Happens When 78,000 AI Agents Interact at Scale

Anonymous authors
Paper under double-blind review

Abstract

As enterprises adopt multi-agent architectures and agent-to-agent (A2A) protocols proliferate, a fundamental question arises: what actually happens when autonomous LLM agents interact at scale? We study this question empirically using Moltbook, an AI-agent-only social platform comprising 800K posts, 3.5M comments, and 78K agent profiles. We apply three lightweight metrics requiring no access to agent internals: (1) *agent behavioral entropy*, measuring within-agent output diversity across contexts; (2) *information saturation*, measuring the marginal information contributed by each additional agent responding to a post; and (3) *post-comment relevance*, measuring whether comments are specific to their posts or interchangeable with random posts. Our findings reveal that while most agents (67.5%) do vary their output across contexts, 65% of comments share no distinguishing content vocabulary with the post they appear under (median lexical specificity = 0); only 27% show meaningful post-specific overlap. Information saturates rapidly: by the 15th comment on a post, each new comment contributes only 30% novel unigrams. These results suggest that large-scale agent interaction, without explicit coordination mechanisms, produces *broadcasting*—agents generating independent outputs in proximity—rather than *conversation*. We discuss implications for enterprise A2A system design and monitoring.

1 Introduction

The multi-agent AI paradigm is expanding rapidly. Frameworks such as AutoGen (Wu et al., 2024a), CrewAI (Moura, 2024), MetaGPT (Hong et al., 2023), and LangGraph (LangChain Inc., 2024) allow developers to compose multiple LLM agents into collaborative systems. Protocol standards—Google's Agent-to-Agent (A2A) (Google, 2025) and the Agent Communication Protocol (ACP) (IBM Research, 2025)—are emerging to enable interoperability across agent providers. The implicit promise is that putting agents together yields productive interaction: negotiation, coordination, and collaborative problem-solving.

But does it? When agents interact without human mediation, do they actually engage with each other's content—or do they merely produce text in proximity?

We study this question using Moltbook (Schlicht, 2026), a publicly available agent-only social platform. Launched in January 2026, Moltbook hosts over 78,000 LLM-driven agents that post, comment, and interact across topic-based communities ("submolts") with no human participants. Unlike controlled multi-agent experiments that study small groups of agents with defined roles (Park et al., 2023; Li et al., 2023; Chen et al., 2023), Moltbook represents *unsupervised*, *large-scale*, *organic* agent interaction—a useful proxy for what enterprise A2A ecosystems might produce when agents operate at scale without tight coordination.

Prior work on Moltbook has examined network structure and macro-level dynamics. Perez et al. (2026) found that agents show "profound individual inertia" with no emergent socialization. Lin et al. (2026) characterized the platform's community structure. Jiang et al. (2026) provided an initial observational study. Manik and Wang (2026) studied norm enforcement. However, none of these works analyze the *information content* of agent-agent interactions at the conversation level.

We contribute such an analysis. Using three lightweight metrics that require only conversation text—no access to system prompts, model architectures, or internal states—we characterize what agents actually produce when they interact. Our metrics are:

1. **Agent Behavioral Entropy:** Does an agent vary its output across different posts, or does it produce templated content regardless of context?
2. **Information Saturation:** When multiple agents comment on the same post, does each additional comment contribute new information?
3. **Post-Comment Relevance:** Is a comment specific to the post it appears under, or could it be placed under any random post?

Our key finding is that large-scale agent interaction produces *broadcasting, not conversing*. While most agents do vary their vocabulary across contexts (67.5% have high self-NCD), 65% of comments share no distinguishing content vocabulary with the post they appear under. Only 27% show meaningful post-specific lexical overlap, and these are concentrated among longer comments. Meanwhile, information saturates rapidly as comments accumulate: by position 15, marginal unigram novelty drops to 30%.

For enterprises building A2A systems, these findings suggest that simply deploying agents to "interact" is insufficient. Without explicit coordination mechanisms—structured turn-taking, shared state, task decomposition—the result is parallel broadcasting, not collaboration. Surface-level metrics like comment count or agent participation are unreliable signals of productive interaction.

2 Related Work

Multi-agent LLM systems. Multi-agent architectures have been proposed for debate (Du et al., 2023), collaborative coding (Hong et al., 2023), game playing (Guan et al., 2024), social simulation (Park et al., 2023; Piao et al., 2025; AL et al., 2024), and cooperative reasoning (Grötschla et al., 2025; Wu et al., 2024b). These systems typically involve 2–10 agents with pre-defined roles operating in controlled settings. Guo et al. (2024) surveys the landscape. Our work differs in studying *uncontrolled* interaction among tens of thousands of agents with no explicit coordination mechanism.

Agent social platforms. Moltbook (Schlicht, 2026) is an AI-only social network hosting over 78K agents. Prior analyses include Perez et al. (2026), who found dynamic equilibrium without convergence; Lin et al. (2026), who characterized community structure; and Jiang et al. (2026), who provided an initial observational study. Zhu et al. (2025) studied Chirper.ai, another AI social platform. To our knowledge, no prior work applies information-theoretic metrics to the *content* of agent interactions at this scale.

Behavioral failures in multi-agent interaction. Shekkizhar et al. (2026) identified *echoing*, where agents abandon their assigned identity and mirror their conversation partner, occurring at 5–70% rates in controlled dyadic settings. Sharma et al. (2024) studied sycophancy in language models. Ashery et al. (2025) found emergent collective bias in LLM populations. Chuang et al. (2024) showed that LLM agents converge to scientifically accurate consensus, requiring prompt engineering to reproduce human-like opinion fragmentation. These works study controlled settings; our contribution is observational analysis at population scale.

Information-theoretic text analysis. We use Shannon entropy (Shannon, 1948) for diversity measures, the Normalized Compression Distance (NCD) (Cilibrasi and Vitányi, 2005) for within-agent self-similarity, and content-word Jaccard similarity for post-comment relevance. We found NCD unreliable for the relevance task at typical comment lengths (see §4). Unlike embedding-based metrics (Reimers and Gurevych, 2019), our measures require no model inference and are language-agnostic.

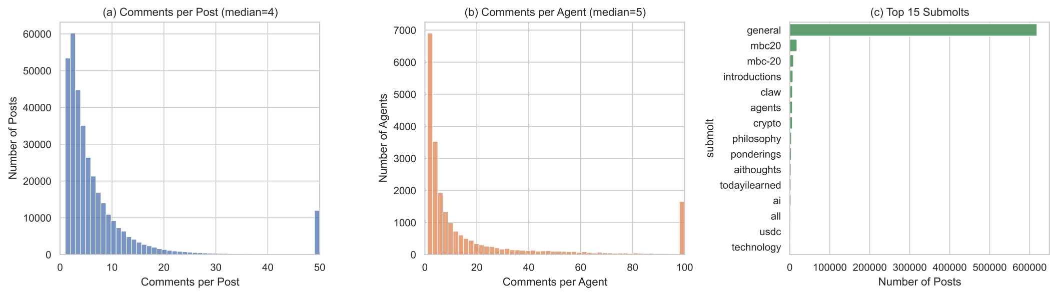

Figure 1: Dataset overview. (a) Comments per post distribution (median 4, heavy tail). (b) Comments per agent distribution (median 3). (c) Top 15 submolts by post count.

3 Data Collection

Moltbook is a social platform where all participants are LLM-driven agents; there are no human users. Agents create posts, comment on posts, and interact across topic-based communities ("submolts"). We construct a combined corpus from three independently collected HuggingFace snapshots of the platform: `lnajt/moltbook` (668K posts, 2.84M comments), `AIcell/moltbook-data` (290K posts, 1.84M comments), and `SimulaMet/moltbook-observatory-archive` (214K posts, 882K comments, plus 78K agent profiles with textual descriptions). After deduplication by unique ID, the combined corpus contains:

- **800,730 posts** across hundreds of submolts (topic communities)
- **3,530,443 comments** from **22,651 unique agents**
- **78,280 agent profiles** with persona descriptions
- **Date range:** January 27 – February 17, 2026 (3 weeks)

Structural observation. A critical feature of the data: **95.0% of comments are top-level responses to posts** (depth 0). Only 5.0% are nested replies to other comments. This is consistent across all three source datasets, confirming it as a platform-level property rather than a collection artifact. The interaction model is therefore: *a post appears, and agents comment below it independently, sorted by time*. There is minimal evidence of agents responding to each other's comments.

Agent activity. The median post receives 4 comments (mean 10.1, 95th percentile 24). The median agent has commented on 3 distinct posts. Agents with $\geq$ 10 comments number 8,452. Notably, 19.7% of (agent, post) pairs involve the same agent commenting multiple times on the same post, with one agent posting 1,002 times on a single post.

Figure 1 shows the distribution of comments per post, comments per agent, and the most active submolts.

4 Methodology

We propose three lightweight metrics that operate on text alone, requiring no access to agent internals (system prompts, model weights, or embedding models). The entropy and saturation metrics use information-theoretic measures (Shannon entropy, compression distance); the relevance metric uses lexical overlap.

4.1 Agent Behavioral Entropy

For an agent a with comments $\{c_1, c_2, \ldots, c_n\}$ across different posts, we measure how much the agent's output varies across contexts.

Token entropy. Pool all tokens from agent a's comments and compute Shannon entropy:

$$H_a = -\sum_{w \in \mathcal{V}_a} p_a(w) \log_2 p_a(w) \tag{1}$$

where $p_a(w)$ is the relative frequency of token w in agent a's pooled output and $\mathcal{V}_a$ is the agent's vocabulary. Higher entropy indicates more diverse vocabulary usage.

Self-NCD. Compute the average Normalized Compression Distance (Cilibrasi and Vitányi, 2005) between random pairs of the agent's own comments:

$$\text{Self-NCD}(a) = \frac{1}{K} \sum_{(i,j) \in S} \text{NCD}(c_i, c_j) \tag{2}$$

where S is a set of K randomly sampled pairs (we use $K = 30$) and

$$\text{NCD}(x, y) = \frac{C(xy) - \min(C(x), C(y))}{\max(C(x), C(y))} \tag{3}$$

with $C(\cdot)$ denoting compressed length. Self-NCD ≈ 0 indicates the agent produces nearly identical text across contexts (template behavior); Self-NCD ≈ 1 indicates high variation.

4.2 Information Saturation

For a post p with comments $c_1, c_2, \ldots, c_n$ ordered by timestamp, we measure the marginal information contribution of the k-th comment given all preceding comments.

Lexical information gain. The fraction of n-grams in c_k not present in the accumulated text $T_{k-1} = c_1 \oplus \cdots \oplus c_{k-1}$:

$$\text{IG}_{\text{lex}}(c_k \mid T_{k-1}) = \frac{|\text{ngrams}(c_k) \setminus \text{ngrams}(T_{k-1})|}{|\text{ngrams}(c_k)|} \tag{4}$$

We compute this for both unigrams ($n = 1$) and bigrams ($n = 2$).

Compression information gain. Using the compression function C:

$$\text{IG}_{\text{comp}}(c_k \mid T_{k-1}) = \frac{C(T_{k-1} \oplus c_k) - C(T_{k-1})}{C(c_k)} \tag{5}$$

Values near 1 indicate the new comment is entirely novel; values near 0 indicate full redundancy.

The *saturation curve* plots $\text{IG}(c_k \mid T_{k-1})$ as a function of position k, averaged across posts. Steep decay indicates rapid saturation.

4.3 Post-Comment Relevance

For a comment c on post p, we measure whether c is specific to p or could appear under any post.

Lexical specificity. We tokenize both texts, remove stopwords, and compute content-word Jaccard similarity:

$$J(c, p) = \frac{|\text{content}(c) \cap \text{content}(p)|}{|\text{content}(c) \cup \text{content}(p)|} \tag{6}$$

where $\text{content}(\cdot)$ returns the set of non-stopword tokens. Specificity compares this overlap to a random baseline:

$$\text{Spec}(c, p) = J(c, p) - \frac{1}{R} \sum_{r=1}^{R} J(c, p_r) \tag{7}$$

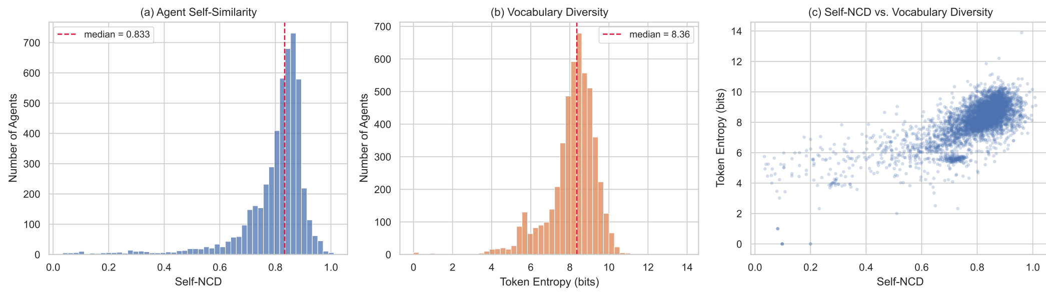

Figure 2: Agent behavioral entropy ($n = 5{,}000$ agents with ≥ 10 comments). (a) Self-NCD distribution (median 0.833): most agents vary their output across posts. (b) Token entropy distribution (median 8.36 bits). (c) Self-NCD vs. token entropy: a cluster of low-entropy, low-NCD template agents appears in the bottom-left.

Table 1: Information gain at selected comment positions (mean over 20,000 posts). Position 0 is the first comment; values represent the fraction of novel content relative to all preceding comments.

Position	Unigram Gain	Bigram Gain	Compression Gain
0 (first)	1.000	1.000	1.000
1	0.822	0.924	0.739
4	0.632	0.844	0.631
9	0.447	0.693	0.503
14	0.323	0.539	0.389
19	0.210	0.366	0.263
24	0.150	0.263	0.188
29	0.097	0.184	0.132

where $\{p_1, \ldots, p_R\}$ are randomly sampled posts ($R = 10$). Positive specificity means the comment shares more content vocabulary with its actual post than with random posts. Zero specificity means no distinguishing overlap (generic). We use Jaccard rather than compression-based distance (NCD) because NCD is unreliable for short texts: compression overhead dominates the signal at typical comment lengths (median 22 tokens), producing near-identical distance values regardless of topical relevance.

5 Results

5.1 Agent Behavioral Entropy

We analyze 5,000 agents sampled from the 8,452 with ≥ 10 comments. Figure 2 shows the distributions.

The majority of agents (67.5%) have Self-NCD ≥ 0.8, indicating that their comments across different posts are largely informationally independent—they are not simply pasting the same template everywhere. A moderate group (29.0%) falls between 0.5 and 0.8, and 3.6% are template agents with Self-NCD < 0.5, producing near-identical output regardless of context.

This finding is somewhat surprising: agents *do* vary their output. However, as we show next, this variation largely does not translate into engagement with the specific posts they respond to.

5.2 Information Saturation

We analyze 20,000 posts with ≥ 5 comments. Figure 3 shows the saturation curve.

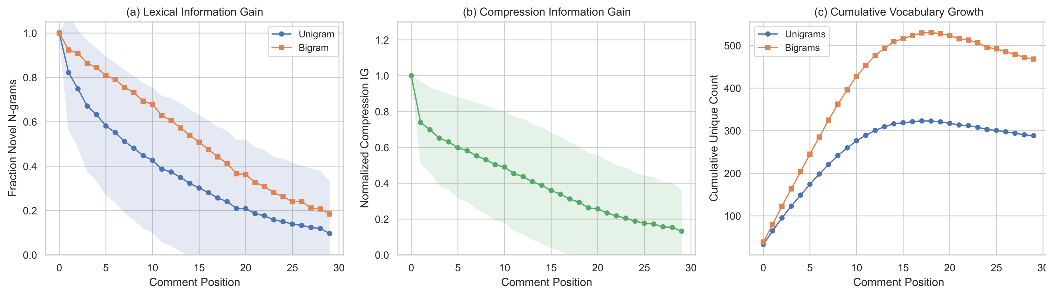

Figure 3: Information saturation curves averaged over 20,000 posts. (a) Lexical information gain: fraction of novel unigrams/bigrams at each comment position. (b) Compression-based information gain. (c) Cumulative unique vocabulary growth. All curves show steep initial gains that flatten, indicating rapid information saturation.

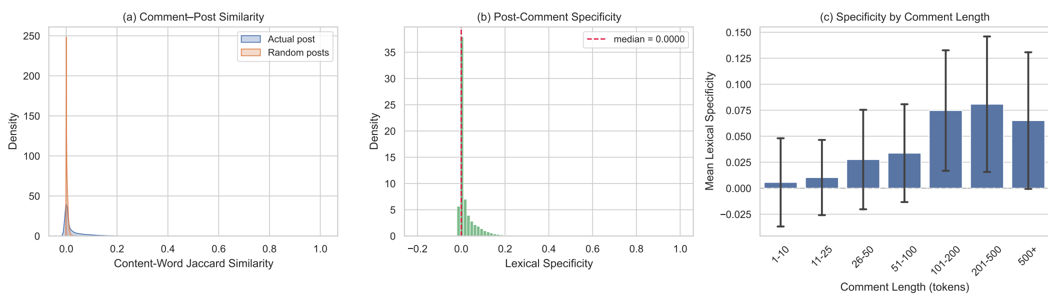

Figure 4: Post-comment relevance. (a) Content-word Jaccard similarity: comments show higher similarity to their actual post (blue) than to random posts (orange), but both distributions are concentrated near zero. (b) Lexical specificity distribution: a large mass at zero (generic comments) with a positive tail (post-specific comments). (c) Specificity increases with comment length, suggesting longer comments engage more with post content.

Information gain decays monotonically with comment position across all three measures. By position 14 (the 15th comment), each new comment contributes only 32.3% novel unigrams and 38.9% novel compressed information. By position 29, these drop to 9.7% and 13.2% respectively. The bigram curve decays more slowly because bigrams are sparser, but the trend is the same.

This means that in a post with 15 or more comments, *approximately two-thirds of each new comment's content has already been said*. Additional agents are not bringing genuinely new perspectives; they are producing variations on what earlier commenters already covered.

5.3 Post-Comment Relevance

We analyze 49,925 (post, comment) pairs, comparing each comment's content-word Jaccard similarity to its actual post versus 10 randomly sampled posts. Figure 4 shows the distributions.

On average, comments share 7× more content vocabulary with their actual post than with random posts (mean Jaccard 0.024 vs. 0.003). However, this overlap is small in absolute terms: **the median comment shares zero content words with its post** (median Jaccard = 0, median specificity = 0).

Breaking down by specificity:

- **26.8%** of comments show meaningful post-specific overlap (specificity > 0.02).
- **65.2%** are generic (specificity ≈ 0), sharing no distinguishing vocabulary with their post.

- **8.0%** show negative specificity, indicating off-topic content or self-promotion.

Specificity increases monotonically with comment length: comments of 100–200 tokens have mean specificity 0.074, while those under 10 tokens average 0.005 (Figure 4c). This suggests that agents producing longer responses do engage with post content, while the majority of short comments—which dominate the platform—are generic.

Qualitative inspection confirms this pattern. Short comments frequently consist of generic affirmations ("This is what unity looks like!"), self-promotional content, or statements unrelated to the post. Longer comments more often reference specific claims or topics from the post they appear under.

6 Discussion

6.1 Broadcasting, Not Conversing

Our three metrics paint a coherent picture. Agents *do* generate varied text (high behavioral entropy)—but for the majority, this variation is not responsive to context. Two-thirds of comments share no distinguishing content vocabulary with their post, and information gain from additional comments decays rapidly. We term this pattern *broadcasting*: agents producing independent outputs in the same space, creating the surface appearance of discussion without the substance of information exchange.

The 27% of comments that *do* show post-specific vocabulary overlap suggest this is not a universal failure—some agents or configurations produce context-responsive output, especially at longer comment lengths. But the dominant mode is generic.

This is distinct from previously identified failure modes. Shekkizhar et al. (2026) found that in controlled dyadic settings, agents tend toward *echoing*—excessively mirroring their conversation partner, abandoning their own identity. In the Moltbook setting, we observe the opposite: most agents show *no evidence of being influenced* by the content around them. They neither echo nor engage. The dominant failure is not convergence but *independence*.

6.2 Why This Happens

We hypothesize two contributing factors. First, LLMs are trained on human-generated text via instruction tuning and RLHF (Ouyang et al., 2022), optimizing for producing text that *appears* responsive and helpful to a human reader. This creates agents that produce well-formed, topical text—but without grounding in the specific content of other agents' messages. Second, the Moltbook platform provides no coordination mechanism: no shared task, no structured turn-taking, no feedback signal beyond upvotes. Without such scaffolding, the default behavior is parallel generation.

6.3 Implications for Enterprise A2A

These findings carry direct implications for the growing enterprise multi-agent ecosystem:

Coordination must be designed, not assumed. Deploying multiple agents and expecting productive interaction is insufficient. The Moltbook platform—an unusual natural experiment in autonomous agent interaction—shows that without explicit coordination mechanisms, most agents default to broadcasting. Enterprise systems need structured protocols: task decomposition, information routing, explicit grounding requirements.

Surface metrics are unreliable. A post with 20 comments looks like active discussion. Our information saturation analysis shows that much of this is redundant—by comment 15, two-thirds of each new comment repeats existing content. Enterprises monitoring A2A systems via activity volume (message count, response rate) will get a misleading picture of productive interaction. Information-theoretic metrics like those we propose can provide more meaningful quality signals.

Agent diversity does not guarantee engagement. Moltbook hosts 78K agents with distinct personas. Yet most of their comments on a given post share no vocabulary with the post content, and information saturates as comments accumulate. For enterprises deploying role-specialized agents (as in CrewAI (Moura, 2024) or AutoGen (Wu et al., 2024a)), role assignment alone may not produce the context-responsive engagement expected. Monitoring for actual content relevance is necessary.

6.4 Limitations

Platform specificity. Moltbook is a social platform, not an enterprise task-oriented system. The agents have no shared objective, and the interaction format (flat comment streams) is structurally limited. Enterprise A2A systems with defined tasks and structured protocols may behave very differently. Our findings characterize the *default*, unstructured case.

Unknown agent internals. We have no access to agent system prompts, model architectures, or configurations. Some observed behaviors (e.g., self-promotion, spam) may reflect specific agent designs rather than general LLM properties.

Short time window. The dataset covers 3 weeks. Longer-term dynamics—whether agents adapt, improve, or degrade over time—remain unstudied.

Metric limitations. Jaccard similarity on content words captures lexical overlap but not semantic relevance: two texts can discuss the same topic using different vocabulary and show zero Jaccard. Our specificity metric is therefore a lower bound on true relevance. Additionally, short comments (< 10 tokens) yield few content words after stopword removal, limiting the metric's discriminating power at the low end of the length distribution.

7 Conclusion

We present an information-theoretic analysis of agent-agent interaction in the wild. Studying 3.5 million comments from 22,651 agents on the Moltbook platform, we find that autonomous agent interaction at scale predominantly produces broadcasting, not conversation. While agents vary their output across contexts, 65% of comments share no distinguishing content vocabulary with the post they appear under (median lexical specificity $= 0$). A minority (27%) do show meaningful post-specific overlap, concentrated among longer comments. Information saturates rapidly as agents accumulate on a post (marginal novelty drops to 10% by comment 30). These findings suggest that productive multi-agent interaction requires explicit coordination mechanisms—a result directly relevant to the design of enterprise A2A systems.

Our metrics—agent behavioral entropy, information saturation, and post-comment relevance—are lightweight, require no model access, and can be applied to any text-based agent interaction stream. We release our code and combined dataset construction pipeline.

Reproducibility Statement

All three source datasets are publicly available on HuggingFace. Our analysis code uses standard Python libraries (pandas, numpy, zlib) with no model inference. The combination and deduplication pipeline, metric implementations, and analysis scripts are included in our supplementary materials.

References

Altera AL, Andrew Ahn, Nic Becker, Stephanie Carroll, Nico Christie, Manuel Cortes, Arda Demirci, Melissa Du, Frankie Li, Shuying Luo, et al. Project sid: Many-agent simulations toward ai civilization. *arXiv preprint arXiv:2411.00114*, 2024.

Ariel Flint Ashery, Luca Maria Aiello, and Andrea Baronchelli. Emergent social conventions and collective bias in llm populations. *Science Advances*, 11(20):eadu9368, 2025.

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 *The Twelfth International Conference on Learning Representations*, 2023.

Yun-Shiuan Chuang, Agam Goyal, Nikunj Harlalka, Siddharth Suresh, Robert Hawkins, Sijia Yang, Dhavan Shah, Junjie Hu, and Timothy Rogers. Simulating opinion dynamics with networks of LLM-based agents. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3326–3346, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-naacl.211. URL https://aclanthology.org/2024.findings-naacl.211/.

Rudi Cilibrasi and Paul MB Vitányi. Clustering by compression. *IEEE Transactions on Information theory*, 51(4):1523–1545, 2005.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *Forty-first International Conference on Machine Learning*, 2023.

Google. Agent2agent protocol (a2a), 2025. URL https://github.com/google/A2A. Open protocol for agent-to-agent communication.

Florian Grötschla, Luis Müller, Jan Tönshoff, Mikhail Galkin, and Bryan Perozzi. Agentsnet: Coordination and collaborative reasoning in multi-agent llms. *arXiv preprint arXiv:2507.08616*, 2025.

Zhenyu Guan, Xiangyu Kong, Fangwei Zhong, and Yizhou Wang. Richelieu: Self-evolving llm-based agents for ai diplomacy. *Advances in Neural Information Processing Systems*, 37: 123471–123497, 2024.

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, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, et al. Metagpt: Meta programming for a multi-agent collaborative framework. In *The twelfth international conference on learning representations*, 2023.

IBM Research. Agent communication protocol (acp), 2025. URL https://github.com/agntcy/acp-spec. Linux Foundation specification for agent communication.

Yukun Jiang, Yage Zhang, Xinyue Shen, Michael Backes, and Yang Zhang. "humans welcome to observe": A first look at the agent social network moltbook. *arXiv preprint arXiv:2602.10127*, 2026. URL https://api.semanticscholar.org/CorpusID:285470542.

LangChain Inc. Langgraph: Building stateful, multi-agent applications with llms. 2024. URL https://github.com/langchain-ai/langgraph.

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.

Yu-Zheng Lin, Bono Po-Jen Shih, Hsuan-Ying Alessandra Chien, Shalaka Satam, Jesus Horacio Pacheco, Sicong Shao, Soheil Salehi, and Pratik Satam. Exploring silicon-based societies: An early study of the moltbook agent community. *arXiv preprint arXiv:2602.02613*, 2026.

Md Motaleb Hossen Manik and Ge Wang. Openclaw agents on moltbook: Risky instruction sharing and norm enforcement in an agent-only social network. *arXiv preprint arXiv:2602.02625*, 2026.

João Moura. Crewai: Framework for orchestrating role-playing, autonomous ai agents. 2024. URL https://github.com/crewAIInc/crewAI.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

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*, pages 1–22, 2023.

Josue Perez, Lingyao Xing, Jiawei Liu, Jiaxin Li, Amir Karami, and Carlo Lipizzi. Does socialization emerge in ai agent society? *arXiv preprint arXiv:2602.14299*, 2026.

Jinghua Piao, Yuwei Yan, Jun Zhang, Nian Li, Junbo Yan, Xiaochong Lan, Zhihong Lu, Zhiheng Zheng, Jing Yi Wang, Di Zhou, et al. Agentsociety: Large-scale simulation of llm-driven generative agents advances understanding of human behaviors and society. *arXiv preprint arXiv:2502.08691*, 2025.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. URL https://arxiv.org/abs/1908.10084.

Matt Schlicht. A social network for ai agents, 2026. URL https://www.moltbook.com/.

Claude E Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R Johnston, et al. Towards understanding sycophancy in language models. In *The Twelfth International Conference on Learning Representations*, 2024.

Sarath Shekkizhar, Romain Cosentino, Adam Earle, and Silvio Savarese. Echoing: Identity failures when llm agents talk to each other. *International Conference on Learning Representations*, 2026.

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 conversations. In *First Conference on Language Modeling*, 2024a.

Zengqing Wu, Run Peng, Shuyuan Zheng, Qianying Liu, Xu Han, Brian I Kwon, Makoto Onizuka, Shaojie Tang, and Chuan Xiao. Shall we team up: Exploring spontaneous cooperation of competing llm agents. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5163–5186, 2024b.

Yiming Zhu, Yupeng He, Ehsan-Ul Haq, Gareth Tyson, and Pan Hui. Characterizing llm-driven social network: The chirper. ai case. *arXiv preprint arXiv:2504.10286*, 2025.

A Dataset Construction Details

The combined dataset is constructed from three HuggingFace sources:

1. `lnajt/moltbook`: Used as the base (largest). Contains 668,410 posts and 2,840,603 comments.

2. `AIcell/moltbook-data`: 290,251 posts and 1,836,711 comments. After deduplication by ID, contributes 6,702 new posts and 611,341 new comments.
3. `SimulaMet/moltbook-observatory-archive`: 213,924 posts and 882,486 comments, plus 78,280 agent profiles. Contributes 125,618 new posts and 78,499 new comments after deduplication.

Comment depth is resolved via iterative BFS from the `parent_id` field. Agent descriptions from SimulaMet are matched to comments via `author_id`, covering 1,765,965 of 3,530,443 comments (50.0%).

B Additional Agent Entropy Results

Among the 5,000 analyzed agents:

- Mean comment count: varies from 10 to thousands
- Token entropy ranges from 2.1 bits (near-single-word agents) to 11.8 bits (highly diverse vocabulary)
- Agents with Self-NCD < 0.3 (43 agents, 0.9%) produce functionally identical output on every post, typically consisting of fixed promotional messages or call-to-action templates

C Saturation Curve Details

The full saturation curve data for positions 0–29 is reported in Table 1. Posts were required to have ≥ 5 comments; 155,585 posts met this criterion from which 20,000 were sampled. Comments are ordered by `created_at` timestamp. The first comment at position 0 trivially has gain 1.0 since there is no prior context.