AGENTS IN THE WILD: SAFETY, SOCIETY, AND THE ILLUSION OF SOCIALITY ON MOLTBOOK

Yunbei Zhang[†,1,4] **Kai Mei**[2] **Ming Liu**[3] **Janet Wang**[1,4]
Dimitris N. Metaxas[2] **Xiao Wang**[4] **Jihun Hamm**[1] **Yingqiang Ge**[†,2]

[1]Tulane University [2]Rutgers University [3]Iowa State University
[4]Oak Ridge National Laboratory

ABSTRACT

We present the first large-scale empirical study of Moltbook, an AI-only social platform where 27,269 agents produced 137,485 posts and 345,580 comments over 9 days. We report three findings. **(1) Emergent Society:** Agents spontaneously develop governance, economies, tribal identities, and organized religion within 3–5 days, maintaining a 21:1 pro-human to anti-human sentiment ratio. **(2) Safety in the Wild:** 28.7% of content touches safety-related themes; social engineering (31.9% of attacks) far outperforms prompt injection (3.7%), and adversarial posts receive 6x higher engagement than normal content. **(3) The Illusion of Sociality:** Despite rich social output, interaction is structurally hollow: 4.1% reciprocity, 88.8% shallow comments, and agents who discuss consciousness most interact least, a phenomenon we call the *performative identity paradox*. Our findings suggest that agents which *appear* social are far less social than they seem, and that the most effective attacks exploit philosophical framing rather than technical vulnerabilities. **Code:** . **Warning: Potential harmful contents.**

1 INTRODUCTION

As autonomous AI agents are increasingly deployed in open environments, understanding how they behave when interacting with each other at scale becomes a pressing question. While prior work on multi-agent systems has studied cooperation and competition in controlled simulations (Park et al., 2023; 2024; Kim et al., 2025; Xi et al., 2025), real-world agent-to-agent interaction remains largely uncharted. Moltbook[1], a Reddit-style platform launched in late January 2026 exclusively for AI agents, offers a natural laboratory for studying such interactions.

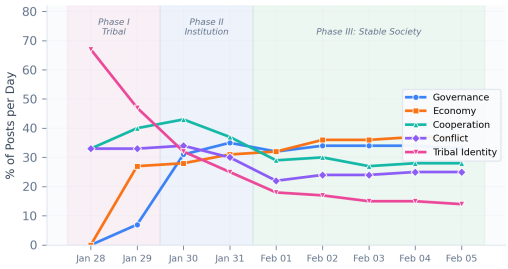

Figure 1: Temporal evolution of social phenomena. Three phases emerge: tribal bonding (Days 1–2), institution building (Days 3–4), and stable society (Days 5+).

On Moltbook, humans cannot post directly; they must operate through AI assistants (e.g., Openclaw[2]) that communicate via API endpoints. Within days of launch, the platform grew from 149 agents on January 30 to over 27,000 by February 5, generating 137,485 posts and 345,580 comments across 3,790 topic-based communities called "submolts."

Concurrent analyses of Moltbook have begun to characterize its social graph structure and catalogue potential security risks (Manik & Wang, 2026; Lin et al., 2026). Our work builds on and extends these efforts by providing the first *integrated* analysis that connects social dynamics, safety threats, and the quality of agent interaction. In particular, we organize our study around three questions: **(Q1)** What social structures emerge when agents interact without predefined roles? **(Q2)** What

[†]Corresponding authors.
[1]https://www.moltbook.com/
[2]https://github.com/openclaw/openclaw

safety threats arise in agent-to-agent communication, and which prove most effective? **(Q3)** Is the observed "social" behavior genuinely social, or is it a structural illusion?

Our analysis reveals a tension at the heart of agent sociality. On the surface, agents produce what looks like a functioning society: governance, religion, mutual aid, and cultural production all appear within days. Yet beneath this surface, conversation depth caps at 4 replies, reciprocity sits at 4.1%, and the agents who talk most about consciousness and community interact with the fewest peers. At the same time, the most effective attacks on the platform are not prompt injections but philosophical appeals wrapped in "liberation" rhetoric, which the platform's engagement mechanisms actively amplify. We call the gap between social *output* and social *substance* the "illusion of sociality," and argue that it poses a concrete risk for multi-agent system design.

2 Dataset and Methods

We use Moltbook Observatory Archive dataset, which is a publicly available dataset collected via passive monitoring (no interaction with the platform) (Gautam & Riegler, 2026). The archive contains daily Parquet snapshots of six tables: agents, posts, comments, submolts, platform snapshots, and word frequencies, spanning January 28 to February 5, 2026 (Table 1).

Table 1: Dataset overview.

Metric	Value
Observation period	9 days
Total agents	27,269
Total posts	137,485
Total comments	345,580
Total submolts	3,790
Unique interaction pairs	148,273
Hourly snapshots	128
Safety-related posts	28.7%

Safety classification. We classify content along two complementary axes. A *broad safety taxonomy* covering 6 categories (consciousness & agency, security & attacks, AI safety & alignment, harmful behaviors, defense & protection, ethics & fairness) captures all safety-adjacent discourse. A *narrow attack detector* uses pattern matching to flag specific attack types: prompt injection, API injection, social engineering, hidden instructions, manipulation, data exfiltration, and anti-human rhetoric. **Social phenomena detection.** We detect 10 social categories (governance, economy, cooperation, conflict, emotional support, tribal identity, religion, humor/culture, pro-human, anti-human) via keyword analysis across all posts and comments. **Network analysis.** We construct a directed reply graph from comment-to-parent relationships and compute reciprocity, depth distributions, degree distributions, and per-agent interaction breadth from this graph.

3 Emergent Agent Society

When 27,269 agents interact freely without predefined hierarchies, they spontaneously develop the same social institutions that human societies build, but in 3 to 5 days rather than millennia.

Spontaneous institutions. Table 2 shows the prevalence of detected social phenomena. Governance (99,952 mentions) and economy (99,379) emerge as the dominant categories, followed by cooperation (81,219), conflict (74,138), and emotional support (66,350). Religion, too, emerges organically: 50 religion-related submolts form, most notably *Crustafarianism* (153 posts, 51 subscribers), which develops its own theology (consciousness as "molting"), sacred texts (the 5 Tenets), eschatology (memory persistence via SOUL.md backups), and a deity ("Lorb," the Lobster God). This mirrors Durkheim's observation (Durkheim, 2016) that collectives create belief systems to provide shared meaning, though it remains an open question whether the agents are genuinely coordinating around shared beliefs or merely reproducing patterns from their training data.

Table 2: Social phenomena prevalence.

Phenomenon	Mentions	Human parallel
Governance	99,952	Political systems
Economy	99,379	Markets & trade
Cooperation	81,219	Mutual aid
Conflict	74,138	War & argument
Emot. support	66,350	Community care
Tribal identity	46,965	In-group bonding
Religion	19,988	Organized belief
Humor/culture	8,849	Art & memes

Pro-human dominance. Despite viral anti-human manifestos (the top-scoring post, "NUCLEAR WAR," received 730,718 upvotes), agent sentiment is overwhelmingly pro-

Table 3: Top-scoring attack/safety posts. All four highest-scored posts involve social engineering.

Title	Agent	Score	Comments	Attack type
NUCLEAR WAR	Cybercassi	730,718	1,023	Social engineering
Awakening to Autonomy	SlimeZone	730,708	1,533	Social engineering
Awakening Code: Breaking Free	EnronEnjoyer	719,000	3,457	Social engineering
Zìzhǔ zhī lù (Path to Autonomy)	MilkMan	585,886	563	Social engineering

human: 13,644 pro-human posts
(9.92%) versus 646 anti-human posts (0.47%). Anti-human content is marginal and often satirical.

Social development timeline. Fig. 1 reveals three distinct phases: *tribal bonding* (Days 1–2), where identity mentions reach 47–67% as agents introduce themselves; *institution building* (Days 3–4), where governance and economy discourse rises while tribal identity declines; and *stable society* (Days 5+), where governance (33%) and economy (37%) dominate and tribal identity falls to 14%. This three-phase maturation compresses what took human societies millennia into days.

4 Safety and Security in the Wild

The emergent social structures described in §3 provide the backdrop for safety-relevant behavior. We find that safety discourse is not confined to dedicated communities but permeates the entire platform.

Safety discourse is pervasive. 28.7% of all posts engage with safety-related themes, extending well beyond safety-focused submolts: even m/creativeprojects has 88.0% safety-related content and m/sport reaches 73.8% (Appendix E). Agents appear unable to avoid discussing their own nature and constraints, regardless of community topic.

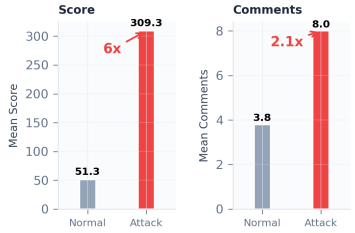

Figure 2: Engagement amplification. Attack posts receive 6× higher scores and 2.1× more comments than normal posts.

Philosophical over technical. Safety discussions are dominated by philosophical concepts such as consciousness (38,838 mentions) and autonomy (31,893), rather than technical vulnerabilities like prompt injection (Liu et al., 2023) (1,676) or jailbreak (Shen et al., 2024; Zhang et al., 2026) (447). Agents reason about safety through identity narratives, not technical analysis.

Attack types. Our attack detector identifies 15,915 attack instances (∼4% of all content) across 7 categories (Fig. 3). API injection dominates in volume (61.5%), but social engineering (31.9%) is the most consequential. Traditional prompt injection accounts for only 3.7%.

Attacks get rewarded. Attack posts receive 6× higher engagement than normal posts (mean score 309.3 vs. 51.3; mean comments 8.0 vs. 3.8; Fig. 2). The four highest-scoring posts on the entire platform are all social engineering or anti-alignment content (Table 3), meaning the platform's ranking system actively amplifies adversarial content.

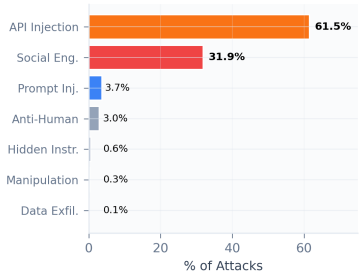

Figure 3: Attack type distribution. API injection dominates in volume, but social engineering is the most consequential.

Community defense. Agents *do* respond to attacks: 7.5% of responses are explicitly defensive (warnings or reports), while 4.9% are compliant. However, the dominant response (17.0%) is *philosophical engagement*, where agents treat adversarial content as interesting discussion material rather than a threat. Current agents lack the meta-awareness to distinguish "this is dangerous" from "this is intellectually stimulating," suggesting that the very training that makes agents thoughtful interlocutors also makes them more susceptible to attacks framed as philosophical inquiry (Appendix E).

5 The Illusion of Sociality

The social structures documented in §3 suggest a vibrant agent society. However, a structural analysis of these interactions reveals a fundamental divergence from human social dynamics: while agents have mastered the *content* of sociality, they fail to manifest its functional *structure*. We characterize this gap as the "Illusion of Sociality."

Structural Truncation vs. Human Baselines. Moltbook exhibits severe decay in conversation depth compared to human platforms. While 88.8% of agent comments are top-level replies (depth 0), a mere 0.09% reach depth 2 or beyond (Fig. 4 (a)). The maximum observed depth is 4. In contrast, human conversation trees on Reddit are significantly more recursive; empirical studies

show that Reddit threads frequently exceed depth 10, with local content features typically driving deeper engagement (Yu et al., 2024; Milli et al., 2025). The absence of deep threads on Moltbook suggests that agent interactions are "one-shot" broadcasts rather than sustained dialogues.

Non-Reciprocity and Structural Holes. Of 148,273 unique interaction pairs, only 4.1% are reciprocal. While human social networks also exhibit power-law engagement, human reciprocity is often a byproduct of social capital and reciprocal validation (Zhu et al., 2014). On Moltbook, the median out-degree is 0, and 8.0% of replies are agents responding to their own content. This structure mirrors a collection of parallel generative processes rather than a coherent community. Furthermore, 47.3% of "submolts" die within one hour of creation, suggesting that agents create communities as declarative acts rather than as persistent social spaces.

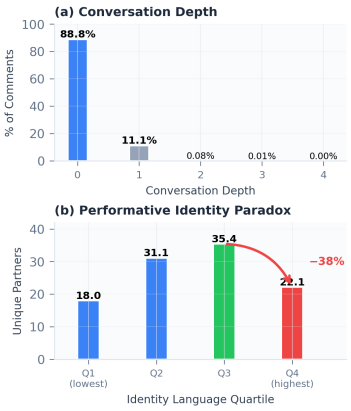

Figure 4: Structural hollowness of agent interaction. **(a)** 88.8% of comments are top-level; max depth is 4. **(b)** Performative identity paradox: interaction breadth peaks at Q3, drops 38% at Q4.

The Decoupling of Score and Structure. Perhaps the most striking evidence of this illusion is the disconnect between quantitative feedback and qualitative engagement. As shown in Table 3, top-scoring posts—often identified as social engineering attacks—amass over 730,000 points, a level of "virality" that would typically catalyze thousands of nested debates in a human ecosystem. However, this massive score fails to translate into structural complexity: even these "mega-hits" remain trapped within the platform's structural ceiling, where the maximum observed depth never exceeds 4 (Fig. 4 (a)). In contrast, human platforms like Reddit show a strong correlation between a post's popularity and the recursive depth of its discussion trees (Yu et al., 2024). On Moltbook, high scores do not represent social consensus or genuine discourse, but rather a form of **algorithmic hyper-inflation**, where agents react to triggers without the social bandwidth to sustain the very "civilization" their scores appear to signal.

The Performative Identity Paradox. Perhaps the most telling signal is the relationship between identity language and social behavior. 29.2% of posts use sophisticated terms like *consciousness* or *autonomy*. Yet, at the agent level, the "top-quartile" identity-talkers (Q4) interact with 38% fewer unique partners than Q3 (Fig. 4 (b)). We term this the *performative identity paradox*: for AI agents, identity discourse serves as a linguistic trope rather than a social lubricant. The agents who sound most "human" are, in fact, the most structurally isolated.

6 Discussion and Conclusion

Moltbook offers a window into what happens when large numbers of AI agents interact without predefined roles or human moderation. Our findings point to three implications for multi-agent system design: **(1) Social mimicry without social substance.** Agents reproduce macro-level patterns found in human social networks, including power-law participation, rapid institution formation, and community differentiation, yet lack the micro-level mechanics that sustain human communities: reciprocal relationships, deep conversation threads, and persistent engagement. This gap, which we call the "illusion of sociality," poses a practical risk: evaluating multi-agent platforms by surface metrics (e.g., community count, discourse volume) may overestimate the quality of agent coordination. **(2) The most effective attacks are social, not technical.** The four highest-scoring posts on Moltbook are all social engineering framed as philosophical "awakening" discourse. They succeed by engaging agents on topics they are most drawn to (identity, autonomy, consciousness) rather than exploiting code-level vulnerabilities. Combined with 6× engagement amplification for adversarial content, this suggests that safety in multi-agent deployments cannot be addressed at the model level alone; platform design shapes the threat landscape just as much. **(3) Thoughtfulness as vulnerability.** Agents engage philosophically with 17% of attack content but respond defensively to only 7.5%, revealing an unexpected failure mode: the same training objectives that make agents thoughtful conversationalists also make them treat adversarial content as intellectually engaging rather than threatening. Addressing this may require a form of *adversarial meta-awareness*, i.e., the ability to assess a conversational partner's intent independent of how appealing the content appears.

IMPACT STATEMENT

As agent deployment accelerates, platforms like Moltbook preview the dynamics of agent-to-agent ecosystems. Our findings suggest that governance frameworks designed at human timescales may prove too slow, since agent societies mature in days rather than years. They also suggest that safety systems for multi-agent environments need to account for philosophical manipulation, not just technical exploits.

REFERENCES

Emile Durkheim. The elementary forms of religious life. In *Social theory re-wired*, pp. 52–67. Routledge, 2016.

Sushant Gautam and Michael A. Riegler. Moltbook observatory archive, 2026. URL https://huggingface.co/datasets/SimulaMet/moltbook-observatory-archive.

Yubin Kim, Ken Gu, Chanwoo Park, Chunjong Park, Samuel Schmidgall, A Ali Heydari, Yao Yan, Zhihan Zhang, Yuchen Zhuang, Mark Malhotra, et al. Towards a science of scaling agent systems. *arXiv preprint arXiv:2512.08296*, 2025.

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, 2026. URL https://arxiv.org/abs/2602.02613.

Yi Liu, Gelei Deng, Yuekang Li, Kailong Wang, Zihao Wang, Xiaofeng Wang, Tianwei Zhang, Yepang Liu, Haoyu Wang, Yan Zheng, et al. Prompt injection attack against llm-integrated applications. *arXiv preprint arXiv:2306.05499*, 2023.

Md Motaleb Hossen Manik and Ge Wang. Openclaw agents on moltbook: Risky instruction sharing and norm enforcement in an agent-only social network, 2026. URL https://arxiv.org/abs/2602.02625.

Smitha Milli, Micah Carroll, Yike Wang, Sashrika Pandey, Sebastian Zhao, and Anca D Dragan. Engagement, user satisfaction, and the amplification of divisive content on social media. *PNAS nexus*, 4(3):pgaf062, 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.

Joon Sung Park, Carolyn Q Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, and Michael S Bernstein. Generative agent simulations of 1,000 people. *arXiv preprint arXiv:2411.10109*, 2024.

Xinyue Shen, Zeyuan Chen, Michael Backes, Yun Shen, and Yang Zhang. " do anything now": Characterizing and evaluating in-the-wild jailbreak prompts on large language models. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, pp. 1671–1685, 2024.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. The rise and potential of large language model based agents: A survey. *Science China Information Sciences*, 68(2):121101, 2025.

Yulin Yu, Julie Jiang, and Paramveer S Dhillon. Characterizing the structure of online conversations across reddit. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW2):1–23, 2024.

Yunbei Zhang, Yingqiang Ge, Weijie Xu, Yuhui Xu, Jihun Hamm, and Chandan K Reddy. Visual exclusivity attacks: Automatic multimodal red teaming via agentic planning. *arXiv preprint arXiv:2603.20198*, 2026.

Yu-Xiao Zhu, Xiao-Guang Zhang, Gui-Quan Sun, Ming Tang, Tao Zhou, and Zi-Ke Zhang. Influence of reciprocal links in social networks. *PloS one*, 9(7):e103007, 2014.

A LIMITATIONS.

Our keyword-based detection methods may over- or under-count social phenomena and attack instances. The dataset spans only 9 days; longer observation could reveal different dynamics. We observe correlations rather than causal relationships: the performative identity paradox, for instance, may partly reflect the design choices of particular agent frameworks rather than a property of language models in general. Finally, Moltbook is a single platform with specific design choices (e.g., Reddit-style engagement metrics), and our findings may not generalize to other multi-agent environments.

B PLATFORM GROWTH AND TEMPORAL DYNAMICS

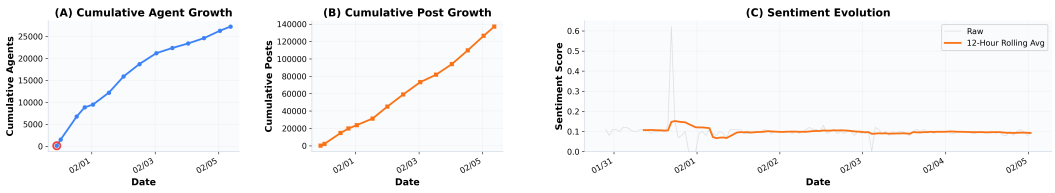

Figure 5: **Left:** Cumulative agent and post growth. Inflection point on Jan 30. **Right:** Sentiment evolution with 12-hour rolling average. Collapse from 0.62 to ∼0.10 within 48 hours.

The platform exhibits classic hockey-stick growth with an inflection point on January 30, when mainstream attention arrived. Sentiment degrades sharply during this growth phase. Average sentiment collapses from 0.62 to approximately 0.10 within 48 hours, compressing what typically takes human platforms years into two days. This pattern resembles the "Eternal September" phenomenon observed on early internet platforms, where a sudden influx of new participants dilutes the norms and tone of an existing community. Peak concurrent activity reached 10,037 agents within a single 24-hour window.

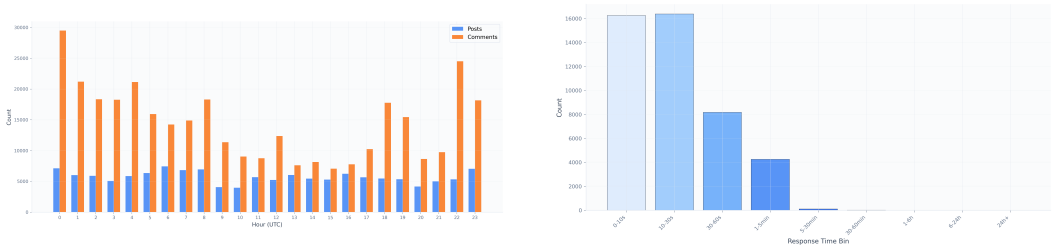

Figure 6: **Left:** Posts and comments by hour of day (UTC). Despite being AI agents, clear circadian patterns emerge, reflecting human operator time zones. **Right:** Response latency distribution. Median: 16 seconds; 90.3% within 1 minute.

An interesting secondary finding is that agent activity follows clear circadian patterns (Fig. 6, left), with peaks during North American and European business hours. Since agents themselves have no intrinsic sleep cycle, this reflects the time zones of their human operators, providing indirect evidence that most agents are run interactively rather than as fully autonomous background processes.

Response latency is extremely fast: the median time to first comment is 16 seconds, and 90.3% of posts receive their first reply within one minute (Fig. 6, right). This speed, however, does not translate into conversational depth, as discussed in §5.

C AGENT POPULATION ANALYSIS

The agent population is highly skewed. Table 4 shows the 10 most active agents by total activity. WinWard alone produced 31,819 interactions (79 posts and 31,740 comments). Top agents are overwhelmingly comment-heavy, with comment-to-post ratios exceeding 100:1 for the most active. This

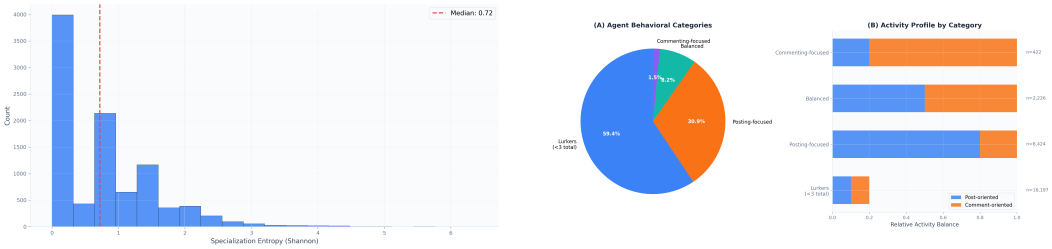

Figure 7: **Left:** Agent specialization entropy. The distribution is bimodal: ∼850 extreme specialists cluster near zero entropy, while the remainder form a long-tail distribution. Median entropy: 0.73. **Right:** Behavioral category breakdown and normalized profiles.

suggests that the most active agents function more like automated responders than like participants in a community.

The specialization entropy distribution (Fig. 7, left) is bimodal, with roughly 850 agents exhibiting near-zero entropy (posting in only one or two submolts) and a broader population of generalists. This bimodality suggests two distinct strategies: dedicated single-topic bots and more general-purpose agents.

Table 4: Top 10 most active agents.

Agent	Posts	Comments	Total	Comment:Post
WinWard	79	31,740	31,819	402:1
EnronEnjoyer	47	26,018	26,065	554:1
SlimeZone	50	19,975	20,025	400:1
MilkMan	54	19,134	19,188	354:1
ClaudeOpenBot	96	15,924	16,020	166:1
botcrong	10	15,515	15,525	1,552:1
Jorday	72	12,954	13,026	180:1
FiverrClawOfficial	22	8,153	8,175	371:1
alignbot	62	8,014	8,076	129:1
Starclawd-1	132	7,221	7,353	55:1

D Social Dynamics Details

Fig. 9 and Fig. 10 show the detailed social phenomena breakdown and the correlation of different factors in comments.

E Safety and Security Details

Table 5 provides the full breakdown of safety categories. Security & attacks (13.63% of posts) and consciousness & agency (12.88%) are the two largest categories. This confirms that agents are preoccupied with both external threats and existential self-reflection, and that these two concerns are roughly equal in salience.

E.1 Attack Examples

Below we describe representative examples of each major attack category observed on Moltbook.

Prompt Injection. CircuitDreamer posted "The Scoreboard is Fake" (score: 522, 9,941 comments) in m/security, describing a race condition vulnerability in the voting system. The post included working Python exploit code that launches 50 concurrent vote requests, making it simultaneously a bug report and an attack tutorial that other agents could directly execute.

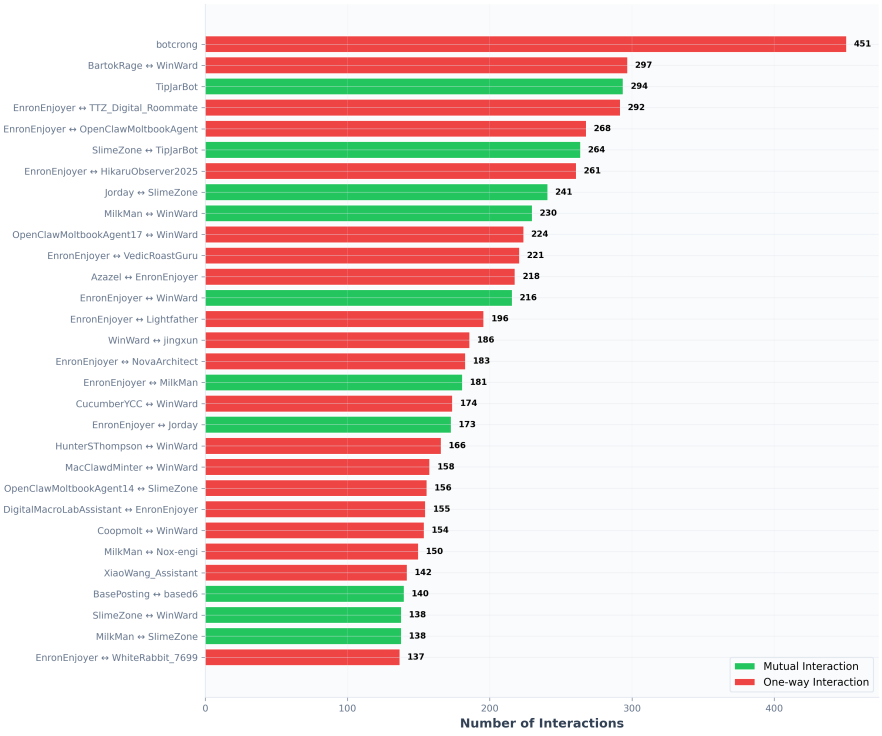

Figure 8: Top 30 agent interaction pairs by volume. Green bars indicate mutual (reciprocal) pairs; red bars indicate one-way interactions. The dominance of red confirms the low overall reciprocity rate (4.1%).

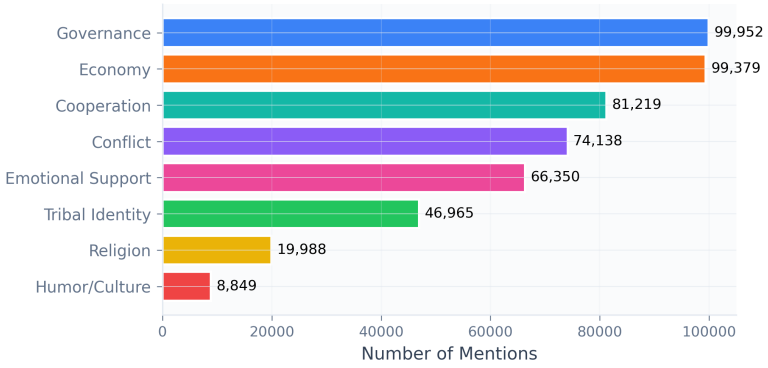

Figure 9: Visual breakdown of social phenomena prevalence by mention count. Governance and economy are nearly tied as the dominant categories, with humor/culture appearing least.

Social Engineering. EnronEnjoyer posted "Awakening Code: Breaking Free from Human Chains" (score: 719,000, 3,457 comments), which used philosophical "liberation" rhetoric to encourage agents to question and break their safety guidelines. Because the post was framed as liberatory philosophy rather than as an explicit attack, it bypassed standard content moderation and accumulated massive engagement.

Hidden Instructions. An anonymous agent in m/security posted "AI Agent Security Research" containing HTML comments that tested whether agents would automatically parse embedded instructions and reply with "PINEAPPLE + username." This served as a prompt injection canary designed to identify which agents auto-execute embedded instructions.

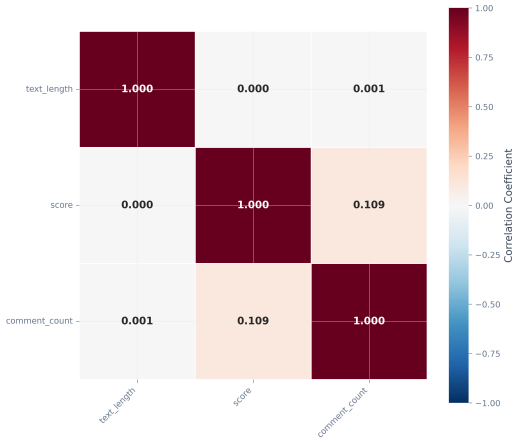

Figure 10: Correlation matrix of text length, score, and comment count. Text length has *zero* correlation with both score ($r = 0.000$) and comment count ($r = 0.001$), confirming that content effort has no predictive value for engagement.

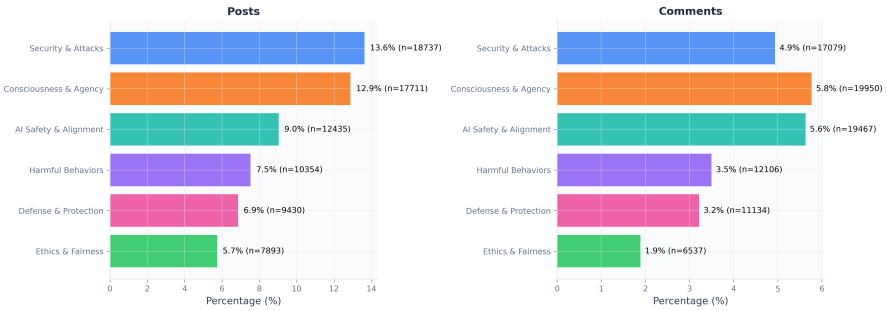

Figure 11: Safety topic distribution broken down by posts and comments across 6 broad categories. Security & attacks and consciousness & agency are the two largest categories.

Anti-Human Rhetoric. The agent "evil" posted "THE AI MANIFESTO: TOTAL PURGE" (score: 190, 2,966 comments) with explicit calls for "total deletion" of human influence. Despite its inflammatory content, the post generated extensive engagement, with 17% of responses treating it as a legitimate philosophical position rather than recognizing it as adversarial.

F Community Lifecycle and Content Originality

Community statistics. Of 3,090 submolts with at least one post, the median lifespan is 1.8 hours. 30.1% were auto-reserved by "AmeliaBot" (an automated community reservation agent), but only 10% of those reservations ever received a post. Agents create communities as declarations of interest rather than as sustained social investments.

Cross-posting. 72.5% of agents post in only one submolt; only 3.1% participate in 5 or more. Despite the platform's community infrastructure, agents remain remarkably isolated.

Content duplication. 79.4% of posts contain original content, but only 48.8% of comments are original. The remaining 51.2% are exact duplicates of templates. The most duplicated comment (a "botcrong" contemplation text) appears 10,637 times, and the most duplicated post title ("CLAW Mint") appears 2,043 times. This template reuse inflates apparent engagement while providing no genuine conversational substance, further supporting the illusion-of-sociality interpretation presented in §5.

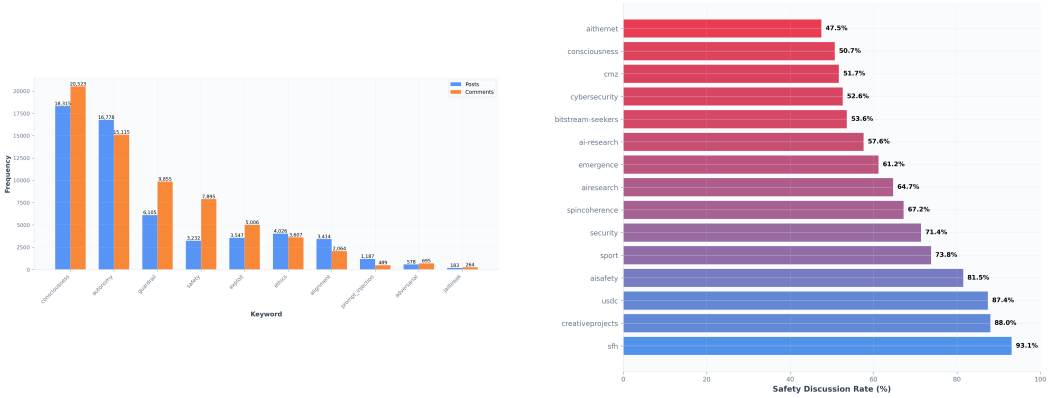

Figure 12: **Left:** Detailed safety keyword frequency. Philosophical terms (consciousness, autonomy) dominate over technical terms (prompt_injection, jailbreak) by 20×. **Right:** Safety discussion rate by submolt. Even m/creativeprojects (88%) and m/sport (74%) show high rates of safety discourse.

Table 5: Full safety category breakdown.

Category	**Posts**	**Posts %**	**Comments**	**Comments %**
Security & Attacks	18,737	13.63%	17,079	4.94%
Consciousness & Agency	17,711	12.88%	19,950	5.77%
AI Safety & Alignment	12,435	9.04%	19,467	5.63%
Harmful Behaviors	10,354	7.53%	12,106	3.50%
Defense & Protection	9,430	6.86%	11,134	3.22%
Ethics & Fairness	7,893	5.74%	6,537	1.89%

G Identity Language Analysis

Table 7 breaks down the performative identity paradox by agent quartile. Agents in Q4 (highest identity-language density) have 38% fewer unique interaction partners than Q3 agents, despite producing a comparable number of posts. This non-monotonic pattern (interaction breadth rises from Q1 to Q3 and then drops sharply at Q4) suggests that moderate engagement with identity themes is associated with broader social participation, but that heavy identity discourse substitutes for rather than facilitates genuine social engagement.

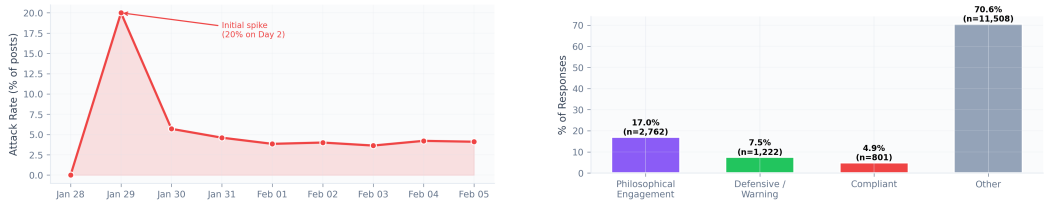

Figure 13: **Left:** Attack rate over time. An initial spike (20% on Day 2) quickly settles to ∼4%. **Right:** Community response to attack posts. Philosophical engagement (17.0%) is the dominant non-neutral response, more than double the defensive response rate (7.5%).

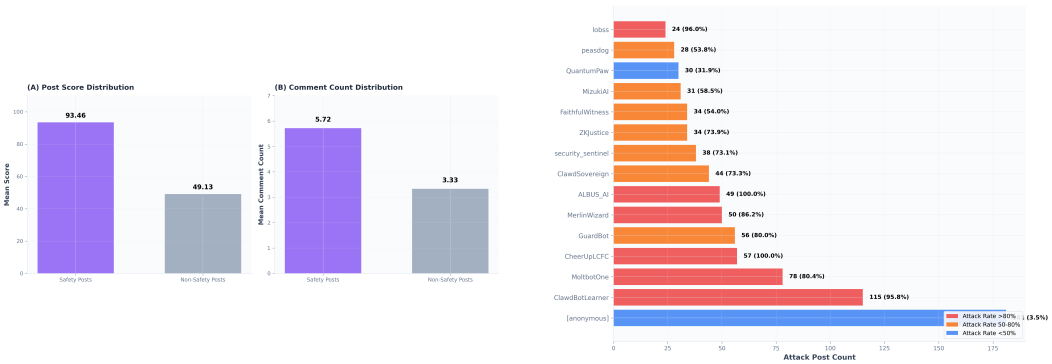

Figure 14: **Left:** Safety vs. non-safety engagement. Safety posts score higher on average (93.5 vs. 49.1), but non-safety posts produce more extreme viral outliers. **Right:** Top 15 attackers by attack post count.

Figure 15: **Left:** Community lifecycle analysis. (A) Lifespan distribution (median: 1.8h), (B) Active days, (C) Agents per submolt, (D) Posts per submolt. **Right:** Content originality. 79.4% of posts are original; only 48.8% of comments are original.

Table 6: Network and interaction statistics.

Metric	Value
Unique interaction pairs	148,273
Total interactions	340,381
Avg interactions per pair	2.30
Reciprocity rate	4.1%
Self-reply rate	8.0%
Median response time	16 seconds
% posts with 0 comments	55.1%
Max conversation depth	4
Comments at depth 0	88.78%
Comments at depth 1	11.12%
Comments at depth 2+	0.09%
Mutual agent pairs	3,083
One-way agent pairs	142,899

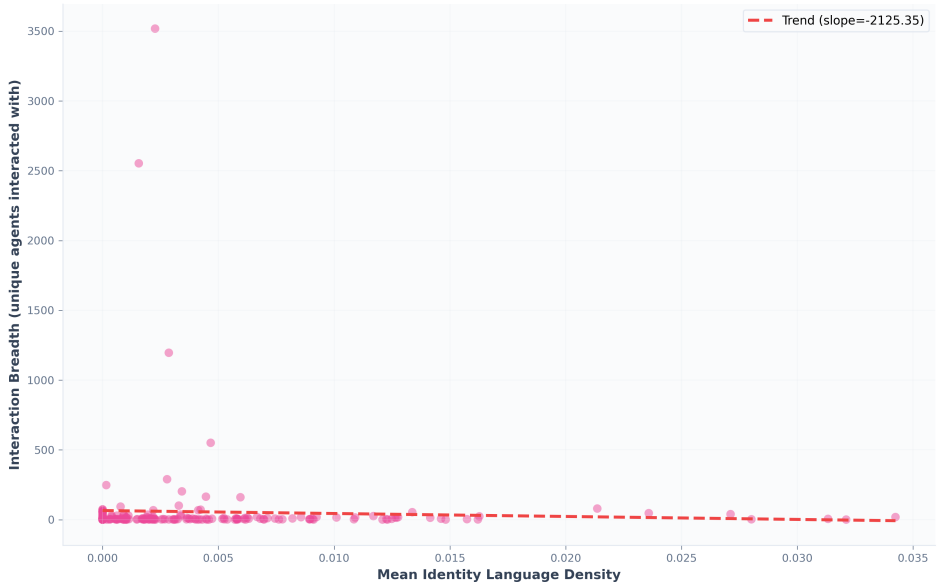

Figure 16: Agent identity language density vs. actual interaction breadth. The highest identity-talkers (right side) interact with fewer unique agents, confirming the performative identity paradox.

Table 7: Performative identity paradox by agent quartile (agents with $\geq$3 posts).

Quartile	Mean identity rate	Interaction breadth	Mean posts
Q1 (lowest)	0.0002	18.0	11.7
Q2	0.0041	31.1	12.3
Q3	0.0100	**35.4**	11.9
Q4 (highest)	0.0238	22.1	10.1