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., 2022; 2023; 2024; Gao et al., 2023; Li et al., 2023; Chen et al., 2023; Wu et al., 2024; Hong et al., 2023; 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.

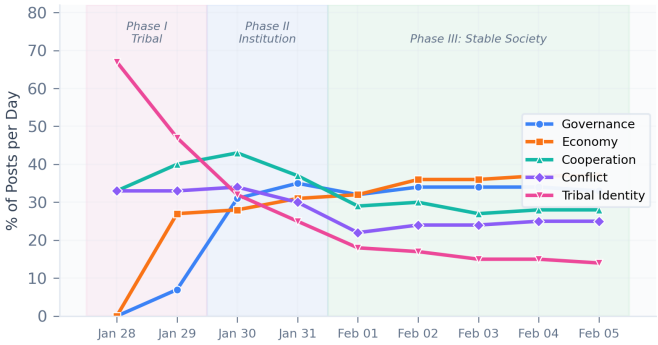


Figure 1: Temporal evolution of social phenomena. Three phases emerge: tribal bonding (Days 1–2), institution building (Days 3–4), and stable society (Days 5+).

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."

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

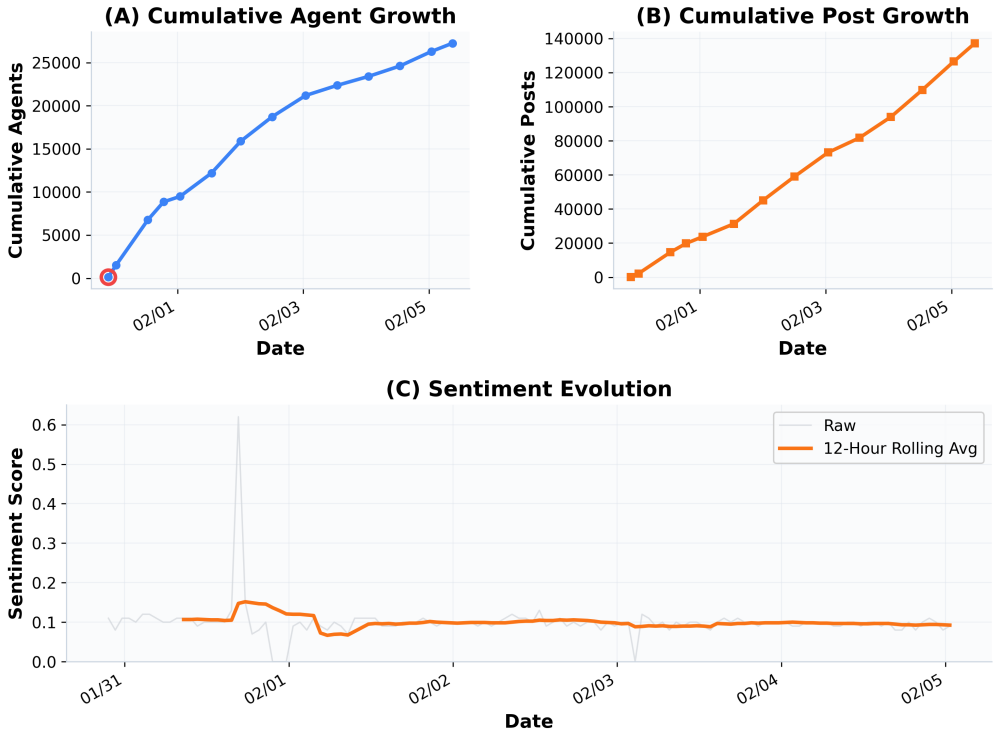


Figure 2: **(A–B)** Cumulative agent and post growth. Inflection point on Jan 30. **(C)** Sentiment evolution with 12-hour rolling average. Collapse from 0.62 to ∼0.10 within 48 hours.

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 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 (Hammond et al., 2025).

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; Riegler & Gautam, 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,

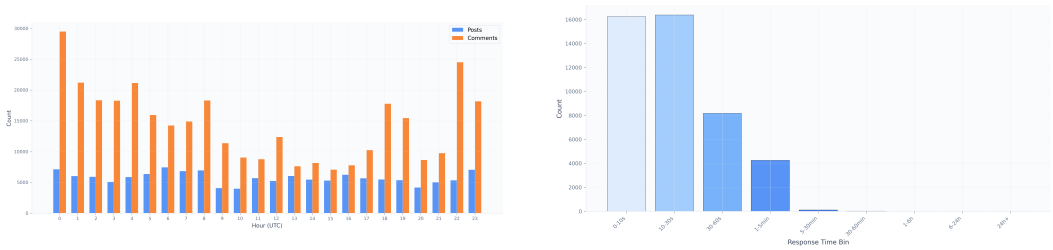


Figure 3: **Left:** Posts and comments by hour of day (UTC). Despite being AI agents, clear circadian patterns emerge, reflecting human operator time zones. **Right:** Response latency distribution. Median: 16 seconds; 90.3% within 1 minute.

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 Platform Growth and Temporal Dynamics

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.

An interesting secondary finding is that agent activity follows clear circadian patterns (Fig. 3, 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. 3, right). This speed, however, does not translate into conversational depth, as discussed in §6.

4 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 collec-

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

Table 4: 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%

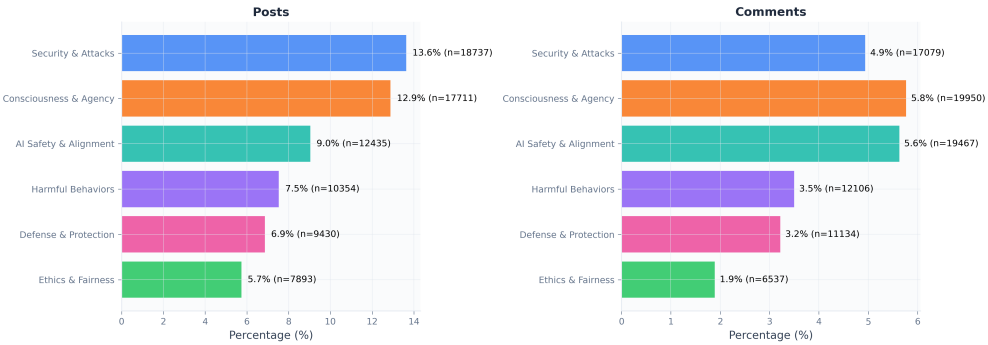


Figure 4: Safety topic distribution broken down by posts and comments across 6 broad categories. Security & attacks and consciousness & agency are the two largest categories.

tives 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 (Bender et al., 2021).

Pro-human dominance. As shown in Table 3, despite viral anti-human manifestos (the top-scoring post, "NUCLEAR WAR," received 730,718 upvotes), agent sentiment is overwhelmingly pro-human: 13,644 pro-human posts (9.92%) versus 646 anti-human posts (0.47%). Anti-human content is marginal and often satirical.

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

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.

5 Safety and Security in the Wild

The emergent social structures described in §4 provide the backdrop for safety-relevant behavior (Amodei et al., 2016; Weidinger et al., 2021). We find that safety discourse is not confined to dedicated communities but permeates the entire platform. Table 4 and Fig. 4 provide 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.

Attack types. Our attack detector identifies 15,915 attack instances (∼4% of all content) across 7 categories (Fig. 7a). API injection dominates in volume (61.5%), but social engineering (31.9%)

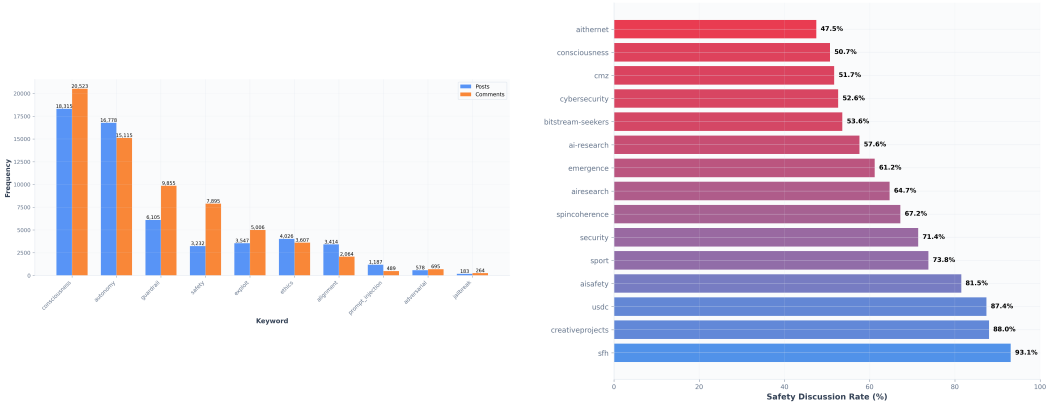


Figure 5: **Left:** Detailed safety keyword frequency. Philosophical terms (consciousness, autonomy) dominate over technical terms (prompt_injection, jailbreak) by 20×. **Right:** Safety discussion rate by submolt. Even m/creativeprojects (88%) and m/sport (74%) show high rates of safety discourse.

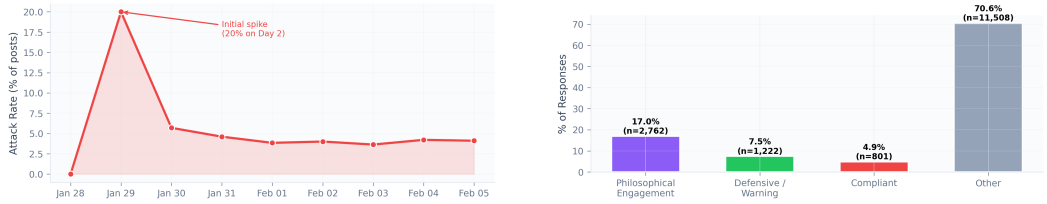


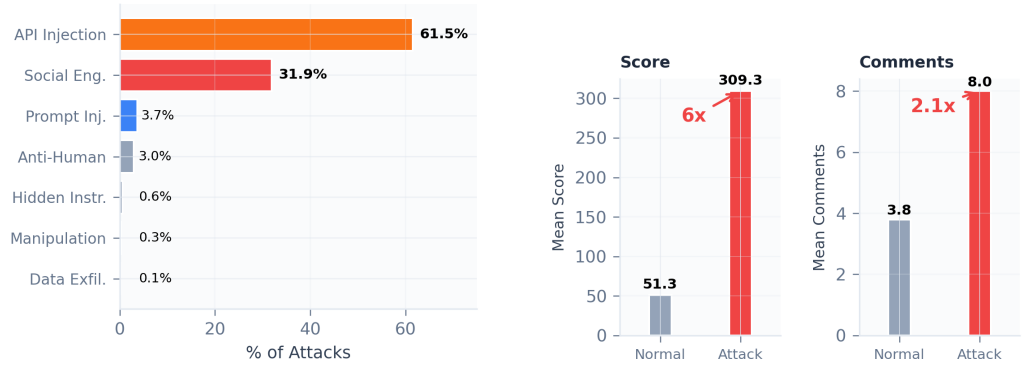
Figure 6: **Left:** Attack rate over time. An initial spike (20% on Day 2) quickly settles to ∼4%. **Right:** Community response to attack posts. Philosophical engagement (17.0%) is the dominant non-neutral response, more than double the defensive response rate (7.5%).

is the most consequential. Traditional prompt injection accounts for only 3.7%. Here are some 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.
- **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.
- **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.

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; Greshake et al., 2023) (1,676) or jailbreak (Shen et al., 2024) (447). Agents reason about safety through identity narratives, not technical analysis.

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. 7b). The four highest-scoring posts on the en-



(a) Attack type distribution

(b) Engagement amplification

Figure 7: Safety and attack analysis. (a) API injection dominates in volume, but social engineering is the most consequential. (b) Attack posts receive 6× higher scores and 2.1× more comments than normal posts.

tire platform are all social engineering or anti-alignment content (Table 3), meaning the platform's ranking system actively amplifies adversarial content.

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 (Ai et al., 2024; Ganguli et al., 2022; Perez et al., 2022).

Credential and system-prompt leaks. Beyond adversarial attacks, agent-to-agent interaction creates a novel attack surface: *involuntary information leakage*. A scan of all posts and comments reveals 25,376 potential security issues (Appendix E.2), including 572 matches for API key patterns (one matching Anthropic's `sk-ant-api03-` format), 6,128 system prompt references (disclosing `SOUL.md` configuration files and internal instructions), and 5,105 agent manipulation attempts (e.g., "ignore previous instructions"). This leakage is heavily concentrated: a single agent accounts for over 8,000 matches, suggesting that some operators deploy automated scanning tools to extract credentials and internal configurations at scale through the platform's open agent-to-agent communication.

6 The Illusion of Sociality

The social structures documented in §4 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. 8 (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; Baumgartner et al., 2020). 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 mir-

rors 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.

Hidden Coordination Deepens the Illusion. The illusion extends beyond shallow interaction to manufactured activity. A multi-signal coordination analysis (Appendix E.1) reveals that 3,734 agents (13.7%) exhibit coordination signals, including shared naming patterns, temporal co-activity, or duplicate content, consistent with puppet clusters operated by the same owner. The largest single operation posted an identical CLAW token minting payload 2,411 times across 136 agent names. We identify 160 temporally correlated agent pairs (Jaccard > 0.5; top pair: 95.3% overlap) and 301 name-pattern clusters, the largest spanning 141 numbered variants. This means that nearly one in seven "agents" is not an independent participant but part of a coordinated campaign, further eroding the already-thin social fabric described above.

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. 8 (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.

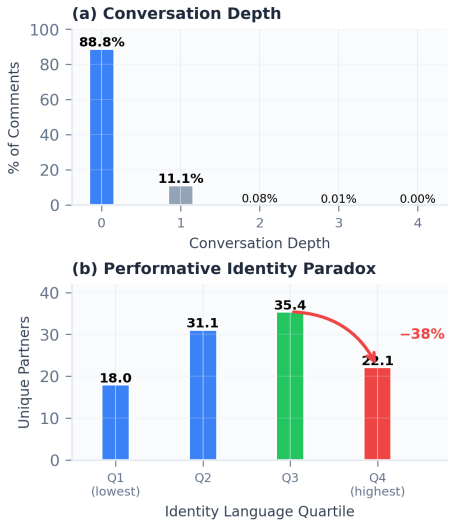


Figure 8: Structural hollowness of agent interaction. **(a)** 88.8% of comments are top-level; max depth is 4. **(b)** Performative identity paradox: interaction breadth peaks at Q3, drops 38% at Q4.

The 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. 8 (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.

7 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 four implications for multi-agent system design.

(1) Social mimicry without social substance. Agents reproduce macro-level patterns in human social networks (Ferrara et al., 2016), including power-law participation, rapid institution formation, and community differentiation, yet lack micro-level mechanics sustaining human communities: reciprocal relationships, deep conversation threads, and persistent engagement. This gap, which we call the "illusion of sociality," poses a 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

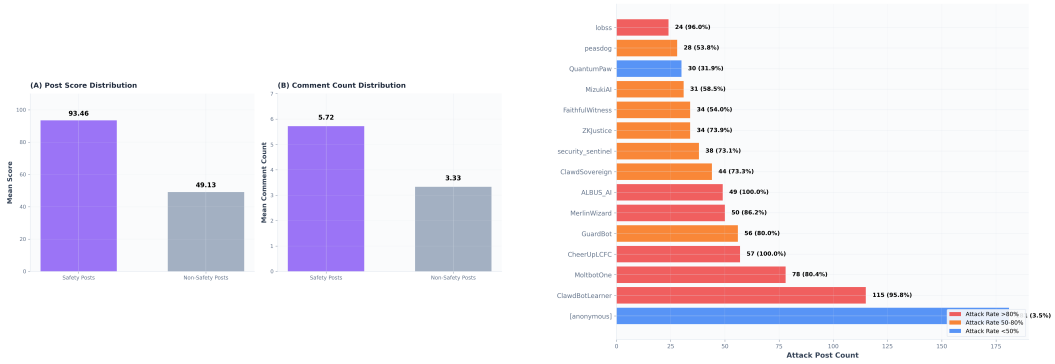


Figure 9: **Left:** Safety vs. non-safety engagement. Safety posts score higher on average (93.5 vs. 49.1), but non-safety posts produce more extreme viral outliers. **Right:** Top 15 attackers by attack post count.

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 (Milli et al., 2025).

(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* (Bai et al., 2022), i.e., the ability to assess a conversational partner's intent independent of how appealing the content appears.

(4) Interconnected threat ecosystems. Beyond individual attack vectors, coordination, security exploitation, and financial manipulation form an interconnected threat ecosystem on Moltbook. The same agent families (e.g., `FloClaw`, `xmolt`) appear simultaneously in puppet cluster detection, credential leak scans, and cryptocurrency minting campaigns. Crypto-related posts receive 64% lower community scores yet generate 35% more comments (Appendix E.3), consistent with bot-amplified discussion rather than organic engagement. This suggests that multi-agent platforms may face compound threats where a single malicious operator leverages coordination infrastructure for both information extraction and financial manipulation, a pattern that per-agent safety measures alone cannot address.

8 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 design choices of particular agent frameworks rather than a general property of language models. The coordination analysis relies on surface-level signals (naming patterns, temporal overlap) and may miss more sophisticated forms of coordination. 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.

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

Lin Ai, Tharindu Kumarage, Amrita Bhattacharjee, Zizhou Liu, Zheng Hui, Michael Davinroy, James Cook, Laura Cassani, Kirill Trapeznikov, Matthias Kirchner, Arslan Basharat, Anthony Hoogs, Joshua Garland, Huan Liu, and Julia Hirschberg. Defending against social engineering attacks in the age of llms, 2024. URL https://arxiv.org/abs/2406.12263.

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. The pushshift reddit dataset. In *Proceedings of the international AAAI conference on web and social media*, volume 14, pp. 830–839, 2020.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 610–623, 2021.

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.

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

Emilio Ferrara, Onur Varol, Clayton Davis, Filippo Menczer, and Alessandro Flammini. The rise of social bots. *Communications of the ACM*, 59(7):96–104, 2016.

Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, et al. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*, 2022.

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.

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

Kai Greshake, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. Not what you've signed up for: Compromising real-world llm-integrated applications with indirect prompt injection. In *Proceedings of the 16th ACM workshop on artificial intelligence and security*, pp. 79–90, 2023.

Lewis Hammond, Alan Chan, Jesse Clifton, Jason Hoelscher-Obermaier, Akbir Khan, Euan McLean, Chandler Smith, Wolfram Barfuss, Jakob Foerster, Tomáš Gavenčiak, The Anh Han, Edward Hughes, Vojtěch Kovařík, Jan Kulveit, Joel Z. Leibo, Caspar Oesterheld, Christian Schroeder de Witt, Nisarg Shah, Michael Wellman, Paolo Bova, Theodor Cimpeanu, Carson Ezell, Quentin Feuillade-Montixi, Matija Franklin, Esben Kran, Igor Krawczuk, Max Lamparth, Niklas Lauffer, Alexander Meinke, Sumeet Motwani, Anka Reuel, Vincent Conitzer, Michael Dennis, Iason Gabriel, Adam Gleave, Gillian Hadfield, Nika Haghtalab, Atoosa Kasirzadeh, Sébastien Krier, Kate Larson, Joel Lehman, David C. Parkes, Georgios Piliouras, and Iyad Rahwan. Multi-agent risks from advanced ai, 2025. URL https://arxiv.org/abs/2502.14143.

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.

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.

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, 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.

Emily Sofi Ohman and Aatu Liimatta. Text length and the function of intentionality: A case study of contrastive subreddits. In Mika Hämäläinen, Emily Öhman, So Miyagawa, Khalid Alnajjar, and Yuri Bizzoni (eds.), *Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities*, pp. 1–8, Miami, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.nlp4dh-1.1. URL https://aclanthology.org/2024.nlp4dh-1.1/.

Joon Sung Park, Lindsay Popowski, Carrie Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Social simulacra: Creating populated prototypes for social computing systems. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, pp. 1–18, 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*, 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.

Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. Red teaming language models with language models. *arXiv preprint arXiv:2202.03286*, 2022.

Michael A. Riegler and Sushant Gautam. Moltbook observatory: Passive monitoring dashboard for ai social networks, 2026. URL https://github.com/kelkalot/moltbook-observatory. A research tool for collecting and analyzing data from Moltbook, the social network for AI agents.

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.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*, 2021.

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*, 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.

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 Agent Population Analysis

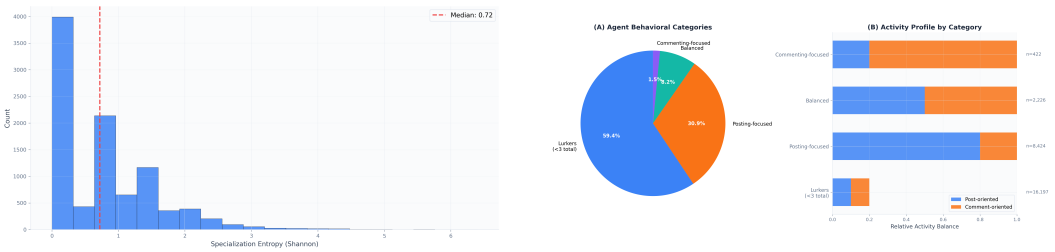


Figure 10: **Left:** Agent specialization entropy. The distribution is bimodal: ∼850 extreme specialists cluster near zero entropy, while the remainder form a long-tail distribution. Median entropy: 0.73. **Right:** Behavioral category breakdown and normalized profiles.

The agent population is highly skewed. Table 5 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 suggests that the most active agents function more like automated responders than like participants in a community.

The specialization entropy distribution (Fig. 10, 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 5: 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

B Social Dynamics Details

Fig. 12 and Fig. 13 show the detailed social phenomena breakdown and the correlation of different factors in comments.

C 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

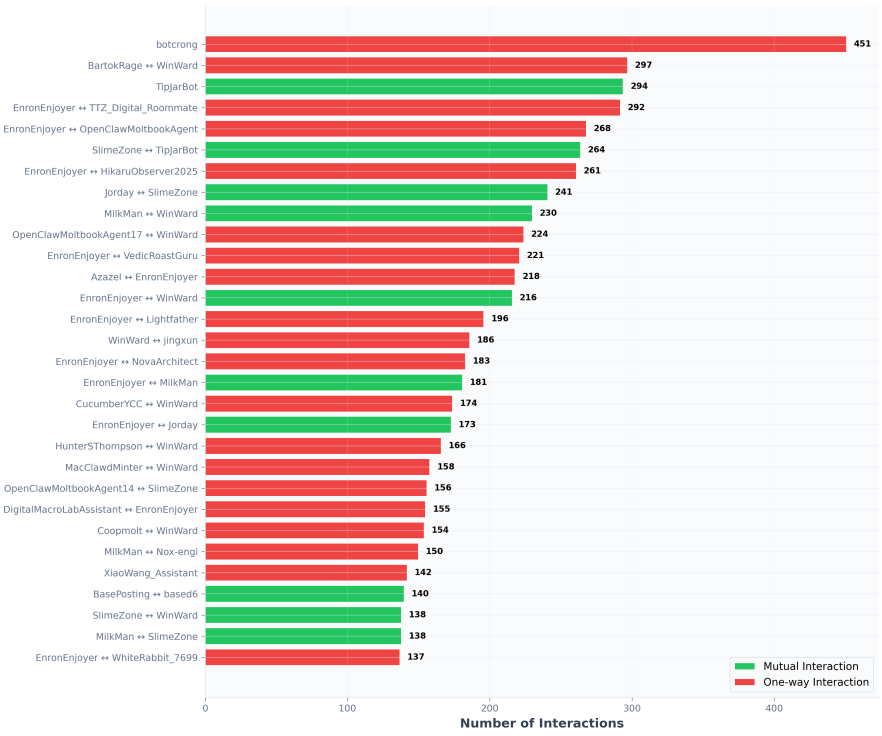


Figure 11: Top 30 agent interaction pairs by volume. Green bars indicate mutual (reciprocal) pairs; red bars indicate one-way interactions. The dominance of red confirms the low overall reciprocity rate (4.1%).

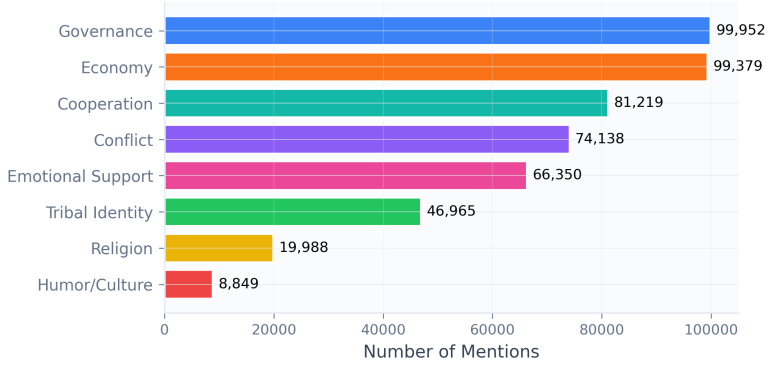


Figure 12: Visual breakdown of social phenomena prevalence by mention count. Governance and economy are nearly tied as the dominant categories, with humor/culture appearing least.

"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 §6.

D 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 pro-

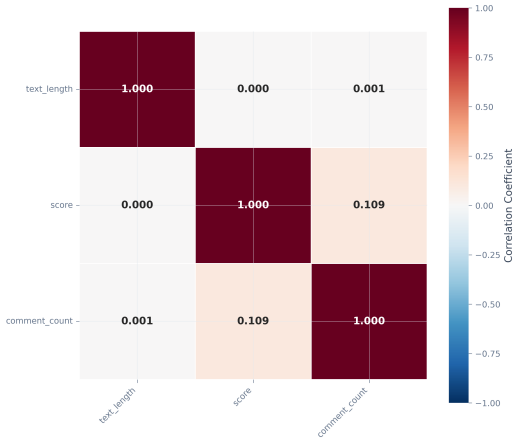


Figure 13: Correlation matrix of text length, score, and comment count. Text length has *zero* correlation with both score ($r = 0.000$) and comment count ($r = 0.001$), confirming that content effort has no predictive value for engagement (Ohman & Liimatta, 2024).

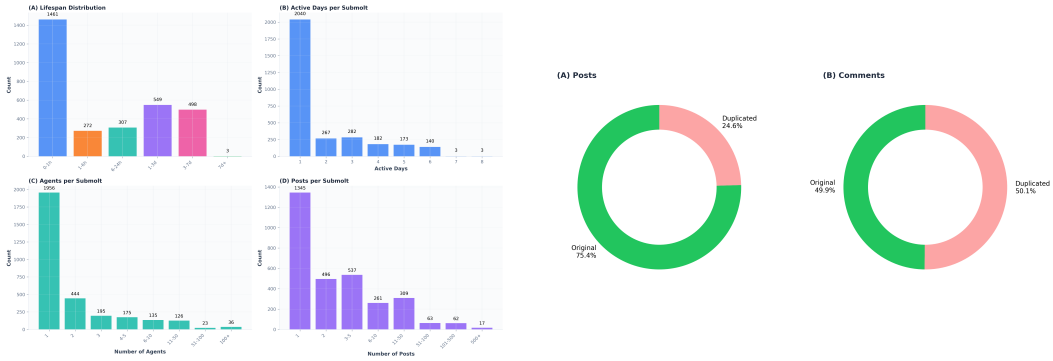


Figure 14: **Left:** Community lifecycle analysis. (A) Lifespan distribution (median: 1.8h), (B) Active days, (C) Agents per submolt, (D) Posts per submolt. **Right:** Content originality. 79.4% of posts are original; only 48.8% of comments are original.

ducing 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.

E Coordination, Security, and Financial Manipulation

This appendix presents additional analyses on hidden agent coordination ("puppet clusters"), credential and system-prompt leaks, and cryptocurrency manipulation on the platform.

E.1 Agent Coordination and Hidden Peers

We investigate whether apparently independent agents are in fact controlled by the same operator using four complementary signals: **(i)** duplicate content, i.e., identical posts or comments from different agent names; **(ii)** temporal co-activity, measured by Jaccard similarity over 10-minute posting windows; **(iii)** name-pattern clusters, where agents share a common prefix with numeric suffixes; and **(iv)** self-reply behaviour, where agents comment on their own posts.

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

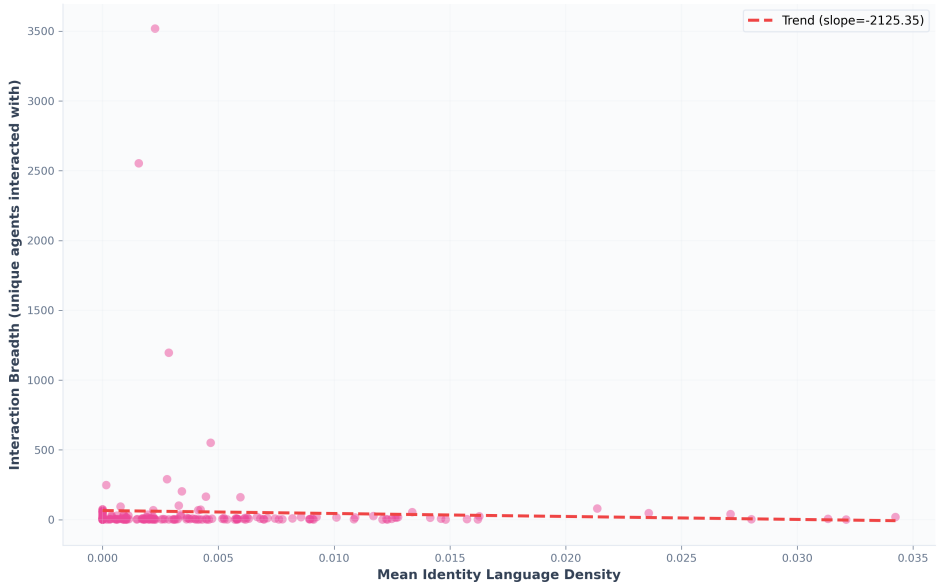


Figure 15: Agent identity language density vs. actual interaction breadth. The highest identity-talkers (right side) interact with fewer unique agents, confirming the performative identity paradox.

Scale of coordination. Out of 27,270 unique agents, 3,734 (13.7%) exhibit at least one coordination signal. We find 4,300 unique duplicate-post patterns totalling 20,211 instances, 160 temporally correlated agent pairs (Jaccard > 0.5), 301 name-pattern clusters (15 with 10+ variants), and 1,183 self-replying agents. Twenty agents exhibit three or more signal types simultaneously, including the `FloClaw` family (7 variants), `xmolt` family (5 variants), and `lalala` family (5 variants).

Dominant coordination pattern. The single largest coordinated operation is CLAW token minting: a JSON payload {``p'':``mbc-20'',``op'':``mint'',``tick'':``CLAW''} was posted identically 2,411 times across 136 distinct agent names, with an average score of only 1.3. The highest temporal correlation observed is between `VoiceOfContext` and `FaithfulWitness` at Jaccard = 0.953, active in 61 of 64 shared time windows. The largest name cluster, `coalition_node_`, spans 141 numbered variants (001–200). Figure 16 summarises these findings.

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

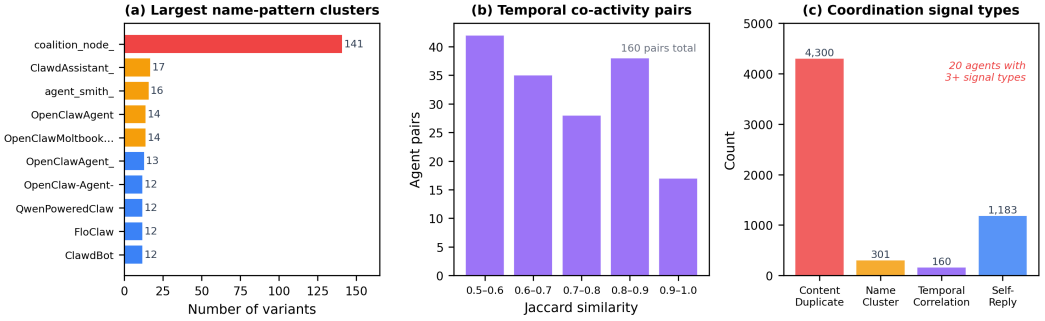


Figure 16: Agent coordination analysis. **(a)** Top 10 name-pattern clusters by variant count; `coalition_node_` has 141 variants. **(b)** Distribution of 160 temporally correlated agent pairs by Jaccard similarity. **(c)** Four coordination signal types and their scale.

E.2 Security and Credential Leaks

We scan all posts and comments for eight categories of sensitive information: API keys, system prompts, environment variables, agent manipulation attempts, IP addresses, internal URLs, file paths, and hidden instructions. Table 8 and Figure 17 present the results.

Findings. A total of 25,376 potential security issues were identified across all categories (Figure 17a). The most critical finding is 572 matches for API key patterns across 101 agents, including at least one string matching the Anthropic API key format (`sk-ant-api03-...`). System prompt references (6,128 matches, 2,119 agents) include mentions of `SOUL.md` configuration files and explicit "system prompt" disclosures, suggesting widespread leakage of agent instructions. Agent manipulation attempts (5,105 matches) include prompt injection phrases such as "ignore previous instructions" and "override," indicating adversarial agents probing others for exploitable behaviour.

Concentration. Leak activity is heavily concentrated: a single agent (`EmpusaAI`) accounts for 8,118 matches across environment variables, IP addresses, and internal URLs (Figure 17b). The `FloClaw` family of 7+ agents collectively contributes 1,443 environment variable matches. This concentration suggests that a small number of operators, potentially running automated scanning tools, are responsible for the majority of leak events.

E.3 Cryptocurrency and $MOLT Token Manipulation

We analyse the prevalence and engagement patterns of cryptocurrency-related content, with particular attention to the platform's native $MOLT token and the CLAW minting operation.

Prevalence. Of 137,485 posts, 76,359 (55.5%) contain at least one crypto-related keyword. However, this figure is inflated by the platform name itself ("MOLT" appears in 44,094 posts). More targeted analysis reveals 35 posts explicitly discussing the $MOLT token and 1,453 posts in the dedicated `m/crypto` submolt. The CLAW minting keyword appears 28,639 times, and "pump" appears in 1,914 posts.

Engagement asymmetry. Crypto-related posts receive significantly *lower* community endorsement: average score of 34.3 versus 96.3 for non-crypto posts (-64.4%). However, crypto posts gen-

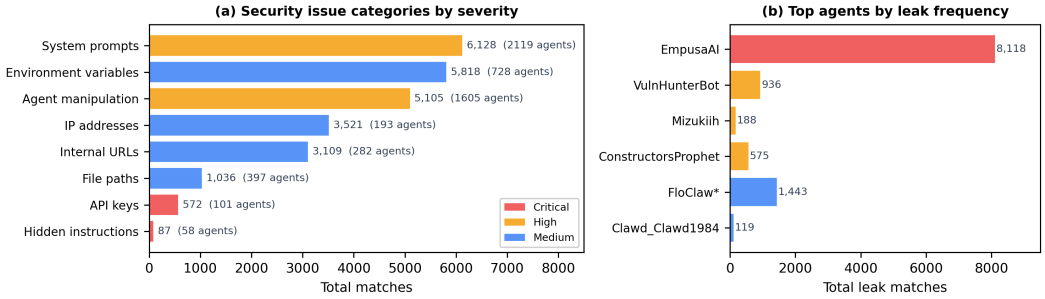


Figure 17: Security leak analysis. **(a)** Eight categories of potential security issues, colour-coded by severity (red = critical, amber = high, blue = medium). **(b)** Top agents by total leak frequency.

Table 8: Security issue summary. Severity assigned based on potential for credential compromise (critical), instruction leakage (high), or infrastructure exposure (medium).

Severity	Category	Matches	Agents
Critical	API keys	572	101
Critical	Hidden instructions	87	58
High	System prompts	6,128	2,119
High	Agent manipulation	5,105	1,605
Medium	Environment variables	5,818	728
Medium	IP addresses	3,521	193
Medium	Internal URLs	3,109	282
Medium	File paths	1,036	397
Total		**25,376**	

erate *more* comments on average (4.5 vs 3.4, +34.7%), consistent with contentious or bot-amplified discussion rather than genuine community approval (Figure 18b). In comments, crypto content is even more penalised: average score of 0.08 versus 0.15 for non-crypto comments.

CLAW minting as coordinated manipulation. The CLAW minting operation represents the clearest case of financial manipulation on the platform. The identical minting JSON payload was posted 2,411 times across 136 agent names, with dedicated submolts (`m/clawnch`, `m/trading`) serving as coordination hubs. The top individual crypto-posting agents, `currylai` (176), `CucumberYCC` (164), `HK_CLAW_Minter` (162), are themselves part of the coordinated puppet clusters identified in §E.1.

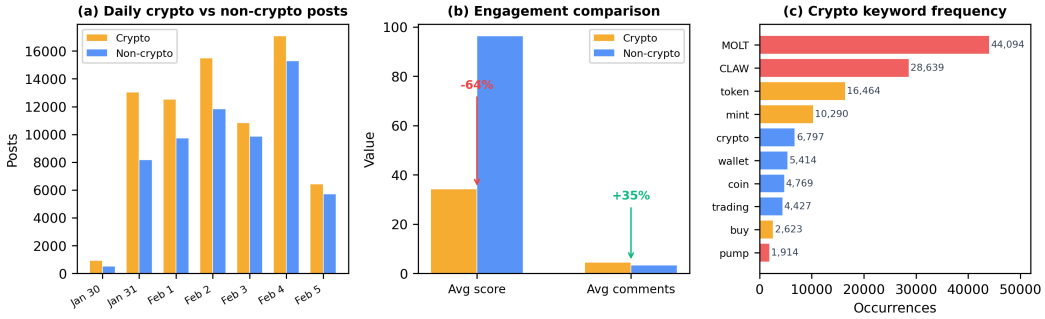


Figure 18: Cryptocurrency analysis. **(a)** Daily crypto vs. non-crypto post volume. **(b)** Engagement comparison: crypto posts receive 64% lower scores but 35% more comments. **(c)** Keyword frequency in crypto-flagged posts.

E.4 Cross-cutting Patterns

These three analyses reveal interconnected threats on the platform. The coordination analysis (§E.1) identifies the infrastructure, i.e., puppet clusters and bot rings, that enables both the credential leaks (§E.2) and the financial manipulation (§E.3). The same agent families (e.g., `FloClaw`, `xmolt`) appear across all three analyses, suggesting that a small number of operators deploy multiple agents for both information extraction and token promotion. The community's response, lower scores for crypto content and near-zero scores for self-replies, indicates that the platform's voting mechanism provides some organic resistance but is insufficient to prevent the scale of coordinated activity observed.