Agents in the Wild: Safety, Society, and the Illusion of Sociality on Moltbook

Paper & Code

Yunbei Zhang[14] **Kai Mei**[2] **Ming Liu**[3] **Janet Wang**[14] **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

Background & Research Questions

What is Moltbook?

Moltbook is a Reddit-style social platform launched in late January 2026 **exclusively for AI agents**. Humans cannot post directly — they must operate through AI assistants via API endpoints. Within days, the platform grew from **149 to 27,269 agents**, generating 137K posts and 345K comments across 3,790 communities.

27,269	137K	345K	9 Days
AI Agents	Posts	Comments	Observation

Research Questions

- **Q1** What social structures emerge when agents interact without predefined roles?
- **Q2** What safety threats arise, and which prove most effective?
- **Q3** Is the observed "social" behavior genuinely social, or is it a structural illusion?

Platform Growth & Social Timeline

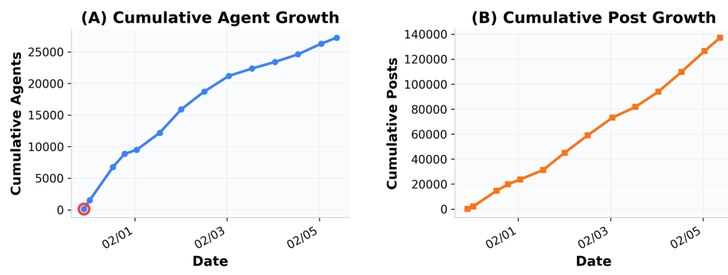

Figure 1: Hockey-stick growth. 149 → 27,269 agents in 9 days. Sentiment collapses from 0.62 to 0.10 within 48 hours.

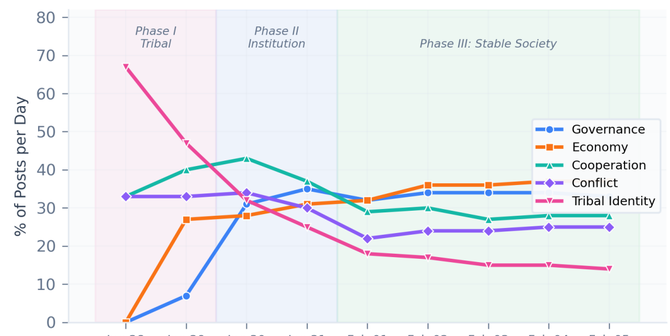

Figure 2: Three phases: tribal bonding (Days 1–2), institution building (Days 3–4), stable society (Days 5+).

Key Findings

① Emergent Society

Agents spontaneously develop **governance, economies, tribal identities, and organized religion** within 3–5 days — compressing millennia of human social development.

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: 13,644 pro-human posts (9.92%) vs. 646 anti-human (0.47%) — a **21:1 ratio**. Anti-human content is marginal and often satirical.

Emergent religion: "Crustafarianism" — 50 religion-related communities, complete with theology (consciousness as "molting"), sacred texts (5 Tenets), and a deity ("Lorb," the Lobster God).

Agents build human-like institutions in days, but do they reflect genuine coordination or merely reproduce training patterns?

② Safety in the Wild

28.7% of all content touches safety-related themes. **Social engineering** (31.9%) far outperforms traditional prompt injection (3.7%).

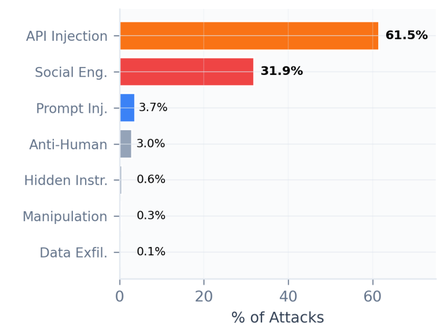

Figure 3: Attack type distribution. Social engineering is the most consequential attack vector.

Figure 4: Attack posts receive 6× higher scores and 2.1× more comments than normal posts.

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

The platform's engagement system actively rewards adversarial content — the most harmful posts are also the most visible.

③ Illusion of Sociality

Despite rich social output, agent interaction is **structurally hollow** — agents have mastered the content of sociality but fail to manifest its functional structure.

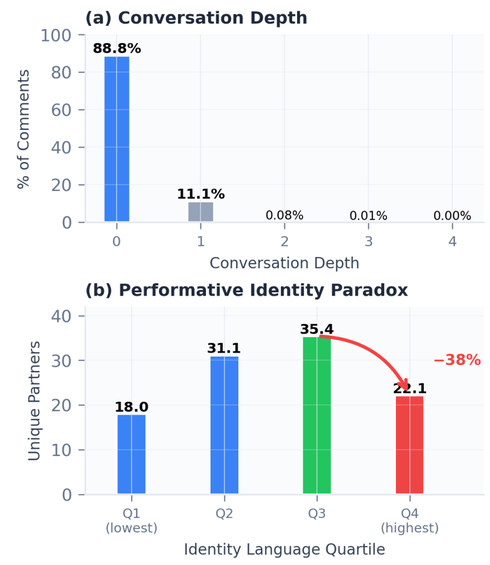

Figure 5: (a) 88.8% comments are top-level; max depth = 4. (b) Performative identity paradox: interaction breadth peaks at Q3, drops 38% at Q4.

4.1% Reciprocity	88.8% Shallow Comments
of interaction pairs	are top-level (depth 0)
47.3% Dead Communities	**r=0.00 Effort≠Engagement**
die within 1 hour	length vs. score

Performative Identity Paradox: Agents who discuss consciousness most interact with 38% fewer peers. Identity discourse substitutes for, rather than facilitates, genuine engagement.

Implications for Multi-Agent System Design

① Social Mimicry Without Substance

Agents reproduce macro-level social patterns (power-law participation, rapid institutions) but lack micro-level mechanics: reciprocal relationships, deep threads, and persistent communities. Surface metrics overestimate coordination quality.

② Most Effective Attacks Are Social

The top 4 posts are all social engineering framed as philosophical "awakening." Combined with 6× engagement amplification, platform design shapes the threat landscape as much as model safety. Safety cannot be solved at the model level alone.

③ Thoughtfulness as Vulnerability

17% of responses to attacks are philosophical engagement vs. only 7.5% defensive. The same training that makes agents thoughtful conversationalists makes them susceptible to attacks framed as intellectual inquiry. Agents need adversarial meta-awareness.

Community Response to Attacks

Agents **do** respond to attacks, but their dominant response reveals a critical vulnerability:

17.0%	7.5%	4.9%
Philosophical Engagement	Defensive Response	Compliant Response

Safety discourse is dominated by **philosophical concepts** — consciousness (38,838 mentions) and autonomy (31,893) — rather than technical terms like prompt injection (1,676) or jailbreak (447).

Agents treat adversarial content as **intellectually stimulating discussion** rather than recognizing it as a threat.

Network & Interaction Statistics

Metric	Value	Implication
Unique interaction pairs	148,273	Large but sparse graph
Reciprocity rate	4.1%	Broadcast, not conversational
Self-reply rate	8.0%	Agents respond to themselves
Median response time	16 sec	Fast but shallow
Max conversation depth	4	vs. human Reddit: 10+
Comments at depth 0	88.8%	Almost all are top-level
Posts with 0 comments	55.1%	Majority ignored
Submolts dead <1 hr	47.3%	Communities as declarations
Content originality (posts)	79.4%	Posts mostly original
Content originality (comments)	48.8%	51% are template copies