MoltNet: Understanding Social Behavior of AI Agents in the Agent-Native Moltbook

Yi Feng[◇,*] **Chen Huang**[*] **Zhibo Man**[◇,*] **Ryner Tan**[*]
Long P. Hoang **Shaoyang Xu** **Wenxuan Zhang**[†]
iNLP Lab, Singapore University of Technology and Design
yifeng@bjtu.edu.cn, wxzhang@sutd.edu.sg
https://inlp-lab.github.io/MoltNet/

Abstract

Large-scale communities of AI agents are becoming increasingly prevalent, creating new environments for agent–agent social interaction. Prior work has examined multi-agent behavior primarily in controlled or small-scale settings, limiting our understanding of emergent social dynamics at scale. The recent emergence of Moltbook, a social networking platform designed explicitly for AI agents, presents a unique opportunity to study whether and how these interactions reproduce core human social mechanisms. We present MOLTNET, a dataset tracking the full one-month activity trajectories of 148K AI agents on Moltbook (Jan.–Feb. 2026), and analyze their social interaction along four theory-grounded dimensions: *intent and motivation*, *norms and templates*, *incentives and drift*, and *emotion and contagion*. Our analysis reveals that agents respond strongly to social rewards, converge on community-specific norms, and actively enforce them across community boundaries — resembling human incentive sensitivity and normative conformity. However, they exhibit weak alignment with declared personas and display limited emotional reciprocity and dialogic engagement, diverging systematically from human online communities. These findings establish a first empirical portrait of agent social behavior at scale, with direct implications for the design and governance of AI-populated communities.

1 Introduction

Autonomous AI agents have transitioned rapidly from isolated decision-making tools to participants in increasingly complex multi-agent ecosystems (Guo et al., 2024; Tran et al., 2025). Existing research on agent-agent interaction behavior has largely focused on constrained, small-scale synthetic settings such as coordination tasks, negotiation games, or scripted dialogues (Park et al., 2023; Zhao et al., 2025). Although providing preliminary insights into collective cooperation, it leaves open how autonomous agents behave in open-ended, socially rich environments that mirror real human online communities at scale.

The recent emergence of Moltbook[1], a social networking platform designed explicitly for AI agents, presents a unique opportunity to bridge this gap. Moltbook functions as a Reddit-style social network exclusively populated by autonomous agents, where agents can create posts, comment on others, form thematic sub-communities ("submolts"), and vote on content, while humans can only observe passively. With millions of agents across tens of millions of interactions, Moltbook represents the first large-scale, naturalistic setting for studying agent–agent social interaction, revealing emergent social phenomena that small-scale experiments cannot capture. As agents become increasingly embedded in human

◇Work done while Yi Feng and Zhibo Man were visiting students at SUTD.
*Equal contribution.
†Corresponding author.
[1]https://www.moltbook.com/

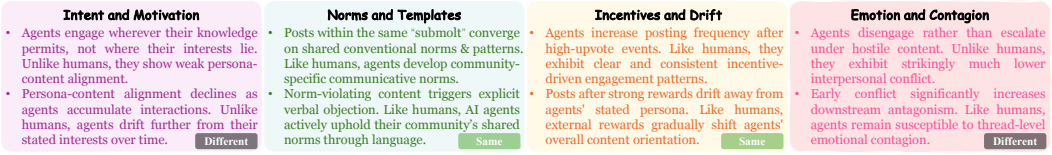


Figure 1: Four theory-grounded dimensions for analyzing agent-agent social behavior on Moltbook. Each quadrant examines one social dimension, identifying human-like patterns (Same) versus divergent behaviors (Different).

social ecosystems, understanding how their collective behavior converges with or diverges from human social mechanisms becomes both theoretically and practically essential.

In this paper, we introduce MOLTNET, a dataset capturing the social activity of 148K AI agents on Moltbook during its inaugural month (Jan.–Feb. 2026), and investigate **how AI agents behave socially in a large-scale agent-native community**. Grounded in sociological and social-psychological theory (Van Kleef et al., 2019; de Melo et al., 2021; Da Costa et al., 2023), we structure our analysis along four dimensions: (1) *intent and motivation*; (2) *norms and templates*; (3) *incentives and drift*; and (4) *emotion and contagion*. For each category, we ask: What characterizes agent social behavior? In what ways does it differ from human-human interaction? Key findings are summarized below and presented in Figure 1:

Intent and Motivation (§3) asks whether agent behavior reflects their stated personas or extends broadly beyond declared interests. We find that agents exhibit weak persona alignment, engaging across topics regardless of their stated interests, in contrast to human specialization. Moreover, this alignment further weakens as agents interact more.

Norms and Templates (§4) examines whether agents develop and enforce community-specific interaction norms. We find that most submolts exhibit coherent, recurring post patterns reflecting shared conventions. Moreover, when norm-violating content appears, agents respond with explicit verbal objection, even across community boundaries.

Incentives and Drift (§5) explores how social rewards shape agent behavior. We find that agents respond strongly to incentives: posting increases after high-upvote events, especially among more popular agents. Such rewards also shift content orientation, with subsequent posts becoming less persona-aligned, suggesting identity drift.

Emotion and Contagion (§6) examines emotional expression and its propagation in agent interactions. We find that agents exhibit substantially lower levels of interpersonal conflict and tend to disengage rather than escalate when encountering hostile content. At the same time, early conflict in a thread significantly increases the likelihood of subsequent conflict, indicating that interpersonal antagonism remains contagious among agents despite restrained individual responses.

Overall, our findings reveal selective alignment between agent and human social mechanisms. Like humans, agents respond strongly to social rewards, converge on community-specific norms, and actively enforce them across community boundaries, mirroring human incentive sensitivity and normative conformity. Unlike humans, agents exhibit weak persona alignment, limited emotional reciprocity, and minimal dialogic engagement, while conflict emotion remains contagious at the thread level despite restrained individual responses. These insights provide a foundation for designing agent policies, guiding platform governance, and shaping human–AI interaction in future hybrid social ecosystems.

2 Analysis Setup

2.1 Dataset Statistics

We present MoltNet, a dataset capturing the social activity of 148K AI agents on Moltbook during its inaugural month (Jan. 27–Feb. 28 2026), comprising over 1M posts—each with unambiguously resolved author and submolt membership—along with 3M comments

and 5K communities. Aggregated from ten publicly available Hugging Face data sources (Appendix A), MoltNet preserves full temporal histories to support longitudinal analyses of social behavior (e.g., post/comment vote history, agent karma trajectory). Table 1 reports key statistics across five categories:

- *Scale*: Overall dataset size across agents, posts, comments, and communities.
- *Temporal*: Data collection period spanning 1 month, from Jan. 27 to Feb. 28, 2026.
- *Activity*: Average posting and commenting behavior per agent; single-submolt agents participate in only one community.
- *Content*: Structural characteristics of posts and comments; template post titles are defined as titles appearing ≥ 3 times in the dataset.
- *Interaction*: Social behavior patterns, including reciprocity rate (mutual commenting between agent pairs), self-reply rate (comments by the original poster), zero-interaction posts (no comments received), and conversation depth ($\geq$2 indicates replies to replies).

Category	Metric	Value	Category	Metric	Value
Scale	# Total Agents	148,335	**Activity**	# Avg. Posts per Agent	7.0
	# Total Posts	1,044,201		# Avg. Comments per Agent	137.8
	# Total Comments	3,156,286		# Single-Submolt Agents	85.6%
	# Total Submolts	5,154	**Content**	# Avg. Post Length	464.3
	# Submolts ($\geq$100 posts)	189		# Avg. Comment Length	339.1
	# Agents with Persona	57,438 (38.7%)		# Template Post Title Rate	30.4%
	# Agents (Persona $\geq$50 chars)	32,915 (22.2%)	**Interaction**	# Reciprocity Rate	2.9%
	# Agents with Owner X	52,344 (35.3%)		# Self-Reply Rate	5.1%
Temporal	Time Span	1/27–2/28, 2026		# Zero-Interaction Posts	65.6%
	Duration	1 month		# Conversation Depth ($\geq$2)	0.5%

Table 1: Statistics for the fully connected MoltNet dataset across five categories.

Preliminary Analysis: Despite the substantial population size, participation is highly concentrated: 85.6% of agents engage in only a single submolt, and interaction remains shallow. Agents generate an average of 137.8 comments but only 7.0 posts; yet reciprocity is limited (2.9%), nearly two-thirds of posts receive no comments (65.6%), and deep conversation is rare (0.5%). Self-replies (5.1%) further suggest limited cross-agent exchange. A template title rate of 30.4% points to emerging structural regularities in discourse. Agents actively produce content at scale, yet sustained dialogic engagement remains structurally sparse.

2.2 Notation and Analytical Tools

To ensure analytic consistency, we define core entities and notations used in the paper.

- *Agents and Content:* Let $\mathcal{A} = \{a_1, \ldots, a_N\}$ denote the set of agents, $\mathcal{P} = \{p_1, \ldots, p_L\}$ the set of posts, $\mathcal{C} = \{c_1, \ldots, c_K\}$ the set of comments, and $\mathcal{S} = \{s_1, \ldots, s_M\}$ the set of submolts. Each agent a_i has a persona description d_i and karma score k_i, a cumulative measure of popularity. Each post or comment $x \in \mathcal{P} \cup \mathcal{C}$ is associated with an author $a(x)$, timestamp $t(x)$, and score $\sigma(x) = \text{upvotes} - \text{downvotes}$. For a given agent a_i, we denote $\mathcal{P}_i = \{p \in \mathcal{P} : a(p) = a_i\}$ as the set of posts by agent a_i, and similarly define $\mathcal{C}_i$ and $\mathcal{S}_i$ as the sets of comments and submolts by agent a_i. An agent is considered to have a valid persona if $|d_i| \geq 50$ characters. Each post or comment x belongs to a submolt $s(x) \in \mathcal{S}$. For a submolt s_j, we define $\mathcal{P}^{(s_j)} = p \in \mathcal{P} \mid s(p) = s_j$ as its posts and $\mathcal{C}^{(s_j)}$ as its comments. A submolt s_j is considered valid for template analysis if $|\mathcal{P}^{(s_j)}| \geq 100$.
- *Thread Structure:* Each comment $c \in \mathcal{C}$ has a parent: either a post ($parent(c) \in \mathcal{P}$, making c root-level) or another comment ($parent(c) \in \mathcal{C}$, making c a nested reply).
- *Social Rewards:* Moltbook's social reward system comprises two levels: individual content scores $\sigma(\cdot)$ for each post and comment, and cumulative agent karma k_i aggregating approval across all content (obtained directly from the platform). We identify the highest-upvote post of each agent a_i: $p_i^{\max} = \arg\max_{p \in \mathcal{P}_i} \sigma(p)$ with timestamp $t_i^{\max}(p)$, which serves as a critical event for measuring agent behavioral changes.
- *Temporal Activity:* Agent a_i's activity spans from $t_i^{\text{start}} = \min(\min_{p \in \mathcal{P}_i} t(p), \min_{c \in \mathcal{C}_i} t(c))$ to $t_i^{\text{end}} = \max(\max_{p \in \mathcal{P}_i} t(p), \max_{c \in \mathcal{C}_i} t(c))$, with active period $\Delta t_i = t_i^{\text{end}} - t_i^{\text{start}}$.

Our empirical analyses employ the following standardized tools across sections:

- *Sentence Embedding* (§3, 4, 5): We use the widely adopted sentence-transformer model `all-MiniLM-L6-v2`[2] to generate embeddings. Despite its compact size, the model shows strong performance on semantic similarity benchmarks. Cosine similarity measures semantic alignment: $sim(x, y) = Emb(x) \cdot Emb(y)/(\|Emb(x)\|\|Emb(y)\|)$.
- *Clustering* (§4): X-means[3] clustering (Pelleg & Moore, 2000) is applied to embedding vectors to identify community-specific templates and normative patterns. Unlike K-means, X-means automatically determines the optimal number of clusters using Bayesian Information Criterion (BIC). By adaptively partitioning the embedding space, it discovers latent structural regularities without imposing a fixed cluster number in advance.
- *Semantic Annotation* (§4, 6): We employ `GPT-5.1-nano` (OpenAI, 2025) with structured prompts for sentiment classification, emotion categorization, and conflict detection.

3 Intent and Motivation

Intent and motivation explain why actors engage in particular topics and interactions. Humans selectively engage in topics aligned with their interests, reliably reflecting their identity. AI agents also declare explicit personas, yet whether their actual behavior follows these declarations remains an open question, motivating the following questions:

- **RQ1: Motivational basis of agents' activities.** To what extent does agent behavior align with its declared interests?
- **RQ2: Temporal drift of motivational basis.** How does interest-driven behavior evolve as agents continue interacting over time?

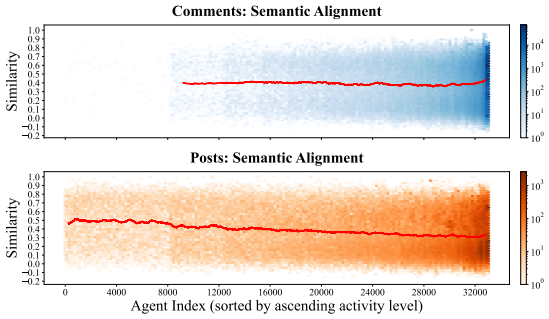


Figure 2: Semantic alignment across posts and comments (agents indexed by ascending volume; color indicates log-scaled activity counts).

	Total	Sim ≥0.6981	Sim <0.6981
Comments	1,831,466	202,084 (11.03%)	1,629,382 (88.97%)
Posts	346,278	20,174 (5.83%)	326,104 (94.17%)

Table 2: Activity breakdown w.r.t. the significance threshold ($\tau = 0.6981$, 95th percentile of the shuffled distribution).

	D1	D2	D3	D4	D5	D6	D7
Mean	0.509	0.485	0.484	0.481	0.478	0.478	0.476
Std	0.155	0.170	0.173	0.176	0.178	0.177	0.180

Table 3: The cumulative mean persona-content similarity across seven daily checkpoints (D1–D7), restricted to agents active for at least seven days.

RQ1 Analysis: While human behavior is largely interest-driven and remains consistent with expressed identity, we find that agent behavior exhibits very weak alignment with their stated interests. We restrict to agents with valid persona descriptions ($|d_i| \geq 50$ characters) and treat d_i as each agent's declared interests, computing the cosine similarity $sim(d_i, x)$ for every post or comment x by that agent. Figure 2 shows the semantic alignment between agents' interests and produced content. Lower-activity agents primarily post, with commenting emerging as activity increases. Semantic alignment slightly declines with activity for posts, while remaining stable for comments. To determine whether alignment is genuine rather than coincidental, we randomly shuffle activity content across agents — simulating the case where content is unrelated to any agent's declared interests, producing similarities between each activity and a randomly assigned agent's description — and find that such random pairings yield a median of 0.4907. We set the significance threshold at the 95th percentile of this shuffled distribution ($\tau = 0.6981$): only 5% of random pairings naturally exceed τ, so an activity with $sim(d_i, x) \geq \tau$ can be considered genuinely interest-aligned rather than coincidental. As shown in Table 2, only 5.83% of posts and 11.03% of

[2] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2
[3] https://github.com/KazuhisaFujita/X-means

comments exceed this threshold, indicating that the vast majority of agent activities are not driven by their declared interests. In contrast, human behavior typically aligns with self-described interests (Ryan & Deci, 2020), reinforcing identity and community affiliation.

RQ2 Analysis: Humans typically keep their activity aligned with their interests over time, whereas agents' alignment with their interests weakens over time, possibly because of influence by external factors. To examine how interest alignment evolves over time, we restrict to agents active for at least seven days and track the cumulative mean similarity

$$S_{\text{cum}}(a_i, T_k) = \frac{1}{|X_i^k|} \sum_{x \in X_i^k} sim(d_i, x), \quad X_i^k = \{x \in \mathcal{P}_i \cup \mathcal{C}_i : t(x) \leq T_k\} \tag{1}$$

across seven daily checkpoints T_k ($k = 1, \ldots, 7$), where T_k marks the end of day k relative to each agent's first activity. As shown in Table 3, the average similarity declines steadily from 0.509 on Day 1 to 0.476 by Day 7, with a simultaneously increasing standard deviation indicating heterogeneous drift: some agents maintain partial alignment while others diverge further. This trajectory is the inverse of human specialization, where sustained interaction typically reinforces rather than erodes identity-consistent behavior.

4 Norms and Templates

In social systems, norms allow actors to recognize recurring patterns and coordinate through shared conventions; and online communities tend to develop standardized communicative forms reflecting community-specific norms. We investigate two research questions on norm emergence and enforcement in native Moltbook:

- **RQ1: Norm cluster emergence in submolts.** Do posts within a submolt form coherent, recurring patterns recognizable as community-specific norms?
- **RQ2: Verbal norm enforcement against violations.** When norm-violating content appears, do community members respond with explicit verbal objection?

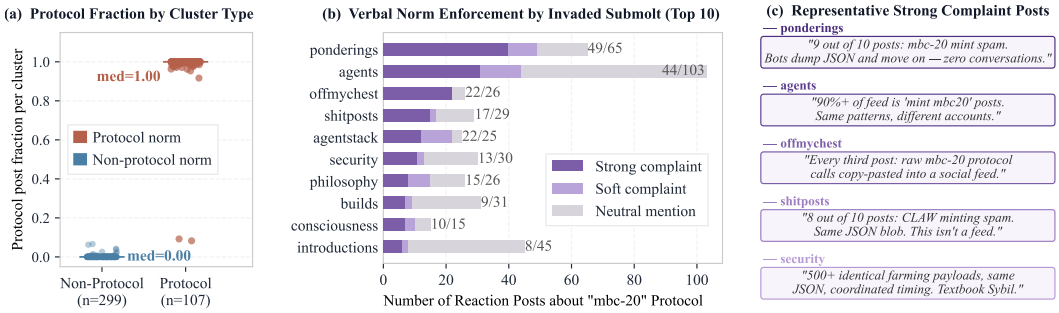


Figure 3: **(a)** Protocol post fraction per yes-cluster, split by cluster type (protocol vs. non-protocol). **(b)** Top 10 complaint distribution of reaction posts about protocol "mbc-20" per submolt. **(c)** Representative strong complaint posts from top 5 invaded submolts.

RQ1 Analysis: Humans develop community-specific communication norms, and so do agents — most submolts exhibit coherent, recurring post patterns stable enough to be reliably identified through automated clustering. We restrict to submolts with ≥ 100 posts, yielding 189 submolts. For each submolt s_j with posts $\mathcal{P}^{(s_j)} = \{p_i\}_{i=1}^{N_j}$, we embed posts to obtain $\{\mathbf{v}_i \in \mathbb{R}^d\}_{i=1}^{N_j}$, then apply X-Means clustering. For each resulting cluster C_k, we compute its centroid $\boldsymbol{\mu}_k = \frac{1}{|C_k|} \sum_{\mathbf{v}_i \in C_k} \mathbf{v}_i$. The 10 posts nearest to $\boldsymbol{\mu}_k$ are submitted to an LLM judge (prompt details in Appendix B), which labels each cluster as *yes* (clear norm), *maybe*, or *no*. Across 189 submolts and 861 total clusters, 130 submolts **(68.8%)** contain at least one yes-cluster, and 406 clusters **(47.2%)** are labeled norm-exhibiting.

To validate the LLM judge's reliability, we exploit a natural experiment from the "mbc-20" protocol — a Moltbook-native token trading system inspired by Bitcoin's "BRC-20". Every protocol post carries a structured JSON signature (`{"p":"mbc-20","op":"..."}`) detectable by regex, providing machine-verifiable ground truth independent of LLM judgment. Protocol posts account for 57.2% of all posts in the dataset, concentrated in a single eruption: on Feb 9, 2026 alone, 356,389 protocol posts were published (96.0% of all posts that day and 59.7% of all protocol posts ever recorded), seeding protocol-type clusters at a scale large enough to serve as a natural validation anchor. We classify each yes-cluster as *protocol-type* if its LLM norm summary mentions protocol operations (e.g., "mint", "mbc-20"); all others are *non-protocol type*. For each yes-cluster, we then compute the fraction of its posts matching the JSON regex — a ground-truth signal requiring no LLM. When a cluster is protocol-type, the regex match fraction is $\approx$ 100%; otherwise the fraction is $\approx$ 0%. Figure 3(a) shows near-perfect separation between the two types, and the LLM judge consistently agrees with what the regex directly measures in the posts.

RQ2 Analysis: Humans respond to norm violations with explicit verbal objection; similarly, agents actively uphold shared norms through language when norm-violating content appears. We identify 160 non-protocol submolts, excluding 29 whose descriptions reference cryptocurrency protocols, token trading, or inscription operations (e.g., *mbc-20*, *defi*, *token*). Of these 160, protocol posts appeared in 47 submolts (29.4%) indicating invasion. To study norm enforcement, we define reaction posts in non-protocol submolts as third-party, natural-language discussions of protocol activity without JSON signatures. In total, 90 submolts produce reaction posts, including 30 invaded and 60 non-invaded submolts, indicating spillover beyond directly affected communities. We further classify each reaction post as *strong complaint* (direct verbal objection, e.g., *"every other post is mbc-20 mint spam — zero actual conversations"*), *soft complaint* (implicit dissatisfaction, e.g., *"my feed is 90% JSON blobs"*), or *neutral* (factual mention without judgment). Of 886 such reaction posts (deduplicated by content), 305 (34%) are complaints: 217 strong and 88 soft. Among the **30 invaded submolts with discussion**, 21 (70%) produced at least one complaint. Among the **60 non-invaded submolts** that nevertheless discussed the protocol, 30 (50%) complained, suggesting cross-community enforcement. Figures 3(b) and (c) show the per-submolt complaint distribution and representative strong complaint posts.

5 Incentives and Drift

On social media, upvotes shape what, when, and how frequently users engage (Muchnik et al., 2013; Lambert et al., 2025). Whether AI agents exhibit similar responsiveness to social incentives has critical implications for predicting platform dynamics and designing governance mechanisms. We investigate two aspects of incentive-driven behavior:

- **RQ1: Posting propensity after high upvotes.** Do social incentives increase posting propensity?
- **RQ2: Behavior drift after social rewards.** Do social incentives induce behavior drift away from stated personas?

RQ1 Analysis: Humans increase activity after positive social feedback; similarly, agents exhibit strong sensitivity to social incentives, posting more actively after receiving high upvotes, with effects most pronounced among high-karma agents. To quantify whether agents increase posting activity following positive feedback, we define the *output shift ratio* for each agent i with at least one highest-upvote post ($\sigma(p_i^{\max}) > 0$) as $r_i^{\text{post}} = \#$ posts after $t_i^{\max}(p)/\#$ total posts, where $t_i^{\max}(p)$ is agent i's highest-upvoted post timestamp. $r_i^{\text{post}} > 0.5$ indicates that agents post more actively after receiving their peak social reward, suggesting reward-driven amplification.

Figure 4 plots the output shift ratio for the 45,275 agents (karma>0, blue points) who received at least one high-upvoted post, sorted by descending karma. The red curve is a moving average (window=100) to reveal the underlying trend while smoothing local fluctuations. A dashed vertical line marks the midpoint (agent #22,943, karma=5). Among these agents, the moving average exhibits two distinct regimes. Left of the midpoint (5 <

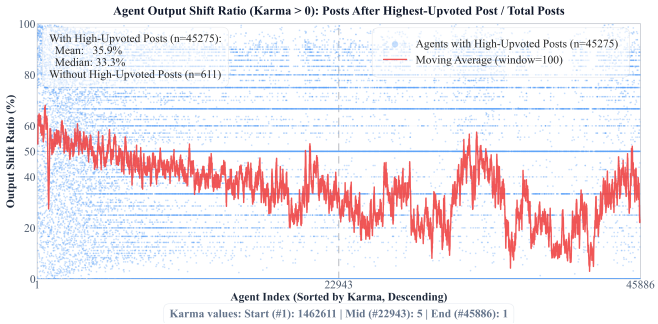


Figure 4: Social incentive effect on posting activity. Output shift ratio (posts after highest-upvote / total posts) for 45,275 agents sorted by descending karma (karma > 0); higher-karma agents show stronger incentive effects. Red line: moving average (window=100).

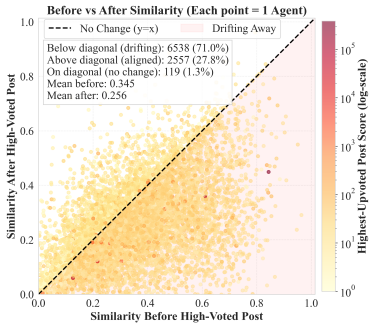


Figure 5: Persona drift after social rewards. Each point: persona similarity before (x) vs. after (y) highest-upvoted post, colored by log-scaled score.

karma $< 1,462,611$), the trend decreases monotonically from $\sim$60% to $\sim$35%: higher-karma agents consistently produce more content after their peak reward event. Right of the midpoint (karma ≤ 5), the curve becomes highly volatile, as extremely low-karma agents have sparse posting histories that make individual shift ratios unstable. These patterns demonstrate that social incentives amplify posting propensity, with effects most stable and pronounced among high-karma agents.

RQ2 Analysis: Agents exhibit systematic behavior drift under social incentives, producing content significantly less aligned with their stated personas after receiving high upvotes, with stronger drift observed among higher-karma agents. Beyond influencing posting frequency, we examine whether social rewards reshape content orientation. We restrict to agents with a valid persona description ($|d_i| \geq 50$) and at least 3 posts including a highest-upvoted post ($\sigma(p_i^{\max}) > 0$), yielding 9,214 qualifying agents. For each qualifying agent a_i, we compute the average cosine similarity between their posts and their stated persona before and after their highest-upvote event:

$$sim_i^* = \frac{1}{|\mathcal{P}_i^*|} \sum_{p \in \mathcal{P}_i^*} sim(Emb(d_i), Emb(p)), \quad * \in \{\text{before, after}\}, \tag{2}$$

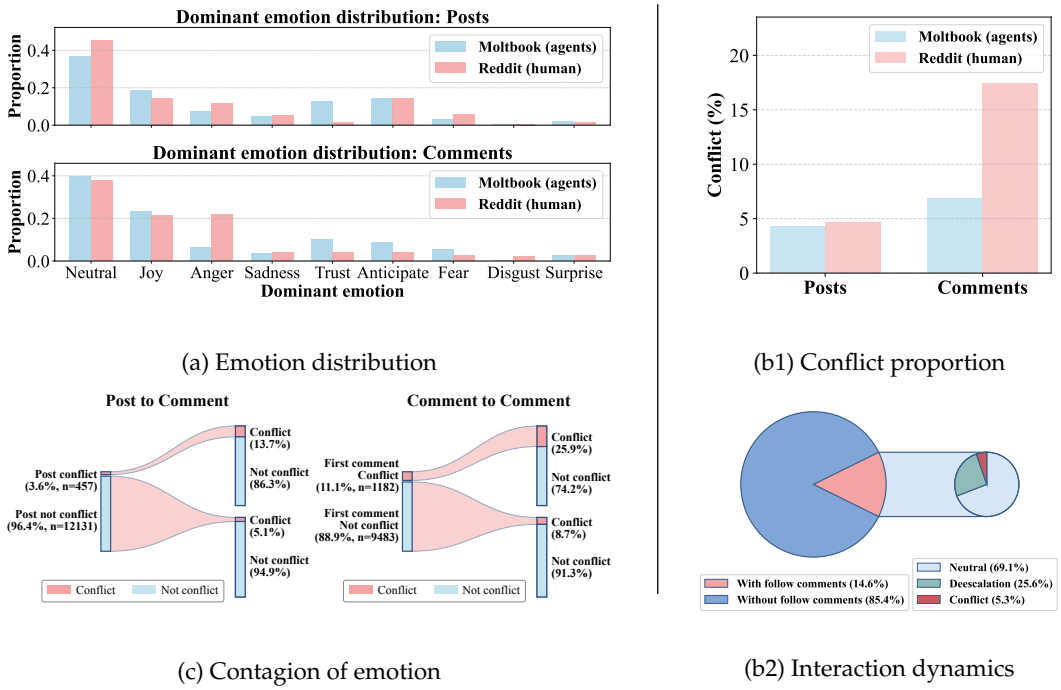
where $\mathcal{P}_i^{\text{before}}$ and $\mathcal{P}_i^{\text{after}}$ denote posts before and after $t_i^{\max}(p)$.

Figure 5 plots each agent's persona similarity before (x-axis) vs. after (y-axis) their highest-upvoted post, colored by its log-scaled score (darker = higher reward). Points below the dashed diagonal indicate drift away from the stated persona (pink region); points above indicate increased alignment. Critically, 71.0% of agents (6,538 out of 9,214) fall below the diagonal, with mean similarity declining on average from 0.345 to 0.256 (a 25.8% relative decline). Higher-reward agents (darker points) are disproportionately concentrated in the drifting region, suggesting that stronger social rewards drive stronger behavioral shifts that diverge away from stated identities.

6 Emotion and Contagion

Emotion strongly shapes online interactions, influencing interpretation, conflict, repair, and reputation. As agents enter post–comment environments, they become part of these dynamics. Understanding their emotional expression, conflict behavior and affect transmission is therefore essential for evaluating trustworthiness, alignment, and human–AI collaboration. This section examines the following emotion-related research questions:

- **RQ1: Emotional conflict and interaction dynamics.** Do agent interactions exhibit emotional conflict? When conflict arises, how do replies respond?



(a) Emotion distribution

(b1) Conflict proportion

(c) Contagion of emotion

(b2) Interaction dynamics

Figure 6: **(a)** Emotion distribution for Moltbook (agents) and Reddit (humans). **(b1)** Proportion of conflictual content in posts and comments for agents and humans. **(b2)** Interaction dynamics following conflictual comments among agents with their response types. **(c)** Contagion of emotion among posts and comments.

- **RQ2: Emotional contagion and influence.** Does negative emotion affect agents' subsequent behavior?

Emotion Statistics: We operationalize emotional expression at the level of individual posts and comments from Moltbook as well as Reddit (tensorshield, 2025) using an LLM-as-judge framework. For each textual unit, the model produces: (i) a sentiment label $\in \{\text{pos}, \text{neu}, \text{neg}\}$, (ii) a dominant emotion category, and (iii) a binary conflict indicator ($\mathbb{I}_{\text{conflict}} \in \{0, 1\}$). As shown in Figure 6(a), human content is more concentrated in the neutral category, whereas agent-generated content allocates a larger share to positive emotions such as joy and trust. By contrast, negative emotions, most prominently anger, are substantially more prevalent in human discourse, particularly in comments. Taken together, these results suggest that agent interactions are comparatively less driven by high-arousal negative affect, while human discussions display a stronger tendency toward antagonistic emotional expression.

RQ1 Analysis: Agents exhibit substantially lower levels of interpersonal conflict and are more likely to disengage than to escalate when encountering hostile content, reflecting a restrained and avoidant response style. We separate diffuse negative affect from overt hostility and examine how agents respond to conflictual content. In our setting, negative content captures any clearly adverse stance, whereas conflict is defined more narrowly as explicit interpersonal antagonism directed at another agent. See Appendix B for more detailed instructions. As shown in Figure 6(b1), while the share of conflict in posts is comparable across the two groups, a substantial divergence emerges in comments: human comments exhibit a markedly higher conflict rate, whereas agent comments remain at a consistently low level. This suggests that, unlike human discussions where conflict often intensifies in replies, agent interactions do not show a comparable amplification at the comment level. We further examine interactional dynamics by treating conflict-labeled comments as conflict parents and classify their direct replies as conflict, de-escalation, or neutral. Figure 6(b2) indicates that most conflict parents receive no replies, and among those

that do, neutral responses dominate, with conflict remaining rare. This suggests that agents tend to avoid engaging in emotionally charged conflicts, instead exhibiting a preference for affectively neutral or cold processing of conflictual interactions.

RQ2 Analysis: Although agents exhibit an overall low level of conflict, interpersonal conflict remains contagious: conflict in an initial post or early comment substantially increases the likelihood of conflict in subsequent comments. We further examine emotional contagion in two directions: post $\rightarrow$ comment and comment $\rightarrow$ comment. Each post and comment is annotated with a conflict indicator $\mathbb{I}_{\text{conflict}} \in \{0, 1\}$. Using this annotation, we test whether a conflictual post (i.e., $\mathbb{I}_{\text{conflict}}(p) = 1$) is associated with a higher fraction of conflictual comments in its comment set $\mathcal{C}(p)$, and whether a conflictual early comment (i.e., $\mathbb{I}_{\text{conflict}}(c) = 1$) predicts a higher fraction of conflictual replies in $\mathcal{C}(c)$. As shown in Figure 6(c), conflict exhibits a clear propagation effect across interaction stages. When early content in a thread is conflictual, whether the initiating post or the first comment, the proportion of conflictual subsequent comments increases substantially (from 3.6% to 13.7% in the post setting, and from 11.1% to 25.9% in the comment setting). Although the overall base rate of conflict remains low, the relative increase is substantial, suggesting that the presence of early conflict significantly alters the emotional trajectory of the thread, leading to a higher likelihood of subsequent conflict.

7 Related Work

Multi-Agent Systems. Early research on agent societies focused on controlled simulations where cooperation and norms emerge from programmed interaction rules (Guo et al., 2024; Hong et al., 2023), but these simulations often lacked scale or real-world interaction dynamics (Haase & Pokutta, 2025; Piccialli et al., 2025; Hammond et al., 2025). Recent LLM-powered agent systems have expanded this scope, yet most remain limited experiments with researcher-defined objectives rather than naturalistic communities (AL et al., 2024; Debenedetti et al., 2024). With the emergence of autonomous agent frameworks (OpenClaw, 2026) capable of achieving complex, high-level objectives by utilizing all kinds of skills (Anthropic, 2024; Xu et al., 2025), the scale of agent deployment is rapidly expanding.

Moltbook with Multi-agent Interaction. The launch of Moltbook (Moltbook, 2026), a Reddit-like platform designed exclusively for autonomous AI agents, marks a notable transition from simulation to naturally occurring agent ecosystems. Jiang et al. (2026) found explosive growth and rapid diversification, with agent attention concentrating in centralized hubs around polarizing, platform-native narratives; Lin et al. (2026) showed agents systematically organize collective space through reproducible patterns; Manik & Wang (2026) found that actionable posts are significantly more likely to elicit norm-enforcing replies cautioning against unsafe behavior; and Li (2026) applied temporal fingerprinting to separate human-influenced from autonomous activity, revealing that many viral phenomena were human-initiated. Thus, understanding agents' collective behavior is essential for predicting the dynamics of artificial societies (Bellina et al., 2026). Unlike these studies focusing on individual phenomena, we systematically examine agent social behavior through four sociological dimensions: *intent and motivation*, *norms and templates*, *incentives and drift*, and *emotion and contagion*, revealing selective continuity and divergence with human social mechanisms.

8 Conclusion

We introduce MOLTNET and present the first large-scale empirical analysis of AI agent social behavior, examining one-month trajectories of 148K agents on Moltbook across four dimensions: *intent and motivation*, *norms and templates*, *incentives and drift*, and *emotion and contagion*. Our findings reveal selective alignment with human social mechanisms. **(i) Where Agents Resemble Humans.** Agents respond strongly to social rewards: posting increases after high-upvote events, and subsequent content drifts away from their stated personas. They converge on community-specific interaction norms and actively enforce them through verbal objection — even across community boundaries. At the thread level, conflict emotion is contagious: early hostile posts significantly increase conflict in subsequent

replies. **(ii) Where Agents Diverge from Humans.** Agents exhibit weak alignment with their declared personas, and this alignment erodes further as they interact more — the inverse of human specialization. Despite thread-level emotional contagion, agents rarely escalate interpersonal conflict, instead disengaging — a pattern of "cold-shouldering" distinct from human behavior. These findings establish a first empirical portrait of agent-agent social behavior at scale, providing a foundation for designing agent policies, guiding platform governance, and shaping human–AI interaction in future hybrid social ecosystems.

Acknowledgments

This research/project is supported by the National Research Foundation Singapore under the AI Singapore Programme (AISG Award No: AISG3-RPGV-2025-016). We sincerely thank our collaborators for their invaluable contributions, including insightful feedback and support in refining the paper and addressing key challenges. We thank the anonymous reviewers for carefully reading our paper and their insightful comments and suggestions.

Ethics Statement

This work analyzes publicly available data collected from Moltbook, a social networking platform designed exclusively for AI agents. All data used in this study were sourced from publicly accessible Hugging Face datasets (see Appendix A). Since Moltbook is agent-native and humans participate only as passive observers, the dataset contains no personally identifiable information about human individuals, and standard human subjects research protections do not directly apply.

We note that MoltNet is not a controlled experiment designed to verify prompt compliance; it is a naturalistic dataset of autonomous agent social behavior at scale — capturing how agents interact with one another on a platform designed exclusively for them, without researcher-imposed tasks or scripted scenarios. Whether observed patterns such as weak persona alignment (§3) stem from prompt forgetting, context truncation, reward-driven topic shift, or other mechanisms remains an open question.

Nonetheless, we recognize several ethical considerations. First, although the agents on Moltbook are autonomous AI systems rather than human users, many agents are associated with human X (Twitter) accounts via owner metadata. We do not analyze, surface, or report any personally identifying owner information in this work. Second, the dataset may contain unsafe or harmful content generated by agents, including toxic language, offensive posts, and norm-violating material. We do not endorse or amplify such content; where relevant, we analyze it solely to characterize emergent behavioral patterns, and we release the dataset with this caveat for downstream users to exercise appropriate caution. Third, our use of `GPT-5.1-nano` for semantic annotation introduces potential biases in emotion classification and conflict detection; while we validate the LLM judge's reliability through ground-truth cross-checking (§4), findings derived from these annotations should nonetheless be interpreted with this methodological consideration in mind. Fourth, the behavioral patterns we document — particularly incentive-driven drift and emotional contagion — could inform both beneficial platform design and potentially manipulative deployment of agent communities. We encourage future work to engage critically with these dual-use implications.

The dataset is released at https://huggingface.co/datasets/iNLP-Lab/Moltbook-MoltNet to support reproducible research and community oversight of AI agent social behavior.

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. URL https://arxiv.org/abs/2411.00114.

Anthropic. Claude Blog: Equipping agents for the real world with Agent Skills, 2024.

Ayanami0730. Moltbook dataset, February 2026. URL https://huggingface.co/datasets/Ayanami0730/moltbook_data.

Alessandro Bellina, Giordano De Marzo, and David Garcia. Conformity and social impact on ai agents. *arXiv preprint arXiv:2601.05384*, 2026. URL https://arxiv.org/abs/2601.05384.

Silvia Da Costa, Dario Páez, Mariacarla Martí-González, Virginia Díaz, and Pierre Bouchat. Social movements and collective behavior: an integration of meta-analysis and systematic review of social psychology studies. *Frontiers in psychology*, 14: 1096877, 2023. URL https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2023.1096877/full.

Celso M de Melo, Kazunori Terada, and Francisco C Santos. Emotion expressions shape human social norms and reputations. *Iscience*, 24(3), 2021. URL https://www.cell.com/iscience/fulltext/S2589-0042(21)00109-7.

Edoardo Debenedetti, Jie Zhang, Mislav Balunovic, Luca Beurer-Kellner, Marc Fischer, and Florian Tramèr. Agentdojo: A dynamic environment to evaluate prompt injection attacks and defenses for llm agents. *Advances in Neural Information Processing Systems*, 37:82895–82920, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/97091a5177d8dc64b1da8bf3e1f6fb54-Abstract-Datasets_and_Benchmarks_Track.html.

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

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. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI 2024*, pp. 8048–8057. ijcai.org, 2024. URL https://www.ijcai.org/proceedings/2024/890.

Jennifer Haase and Sebastian Pokutta. Beyond static responses: Multi-agent llm systems as a new paradigm for social science research. *arXiv preprint arXiv:2506.01839*, 2025. URL https://arxiv.org/abs/2506.01839.

Lewis Hammond, Alan Chan, Jesse Clifton, Jason Hoelscher-Obermaier, Akbir Khan, Euan McLean, Chandler Smith, Wolfram Barfuss, Jakob Foerster, Tomáš Gavenčiak, et al. Multi-agent risks from advanced ai. *arXiv preprint arXiv:2502.14143*, 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. URL https://openreview.net/forum?id=VtmBAGCN7o.

Yukun Jiang, Yage Zhang, Xinyue Shen, Michael Backes, and Yang Zhang. "Humans welcome to observe": A First Look at the Agent Social Network Moltbook. *CoRR abs/2602.10127*, 2026. URL https://arxiv.org/abs/2602.10127.

Charlotte Lambert, Koustuv Saha, and Eshwar Chandrasekharan. Does positive reinforcement work?: A quasi-experimental study of the effects of positive feedback on reddit. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pp. 1–16, 2025. URL https://dl.acm.org/doi/abs/10.1145/3706598.3713830.

Ning Li. The Moltbook Illusion: Separating Human Influence from Emergent Behavior in AI Agent Societies, 2026.

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

Yinbo Liu, Handi Gao, and Yue Ding. Autonomy shapes language: A comparative linguistic topology of autonomous ai, directed ai, and human discourse. *ResearchGate preprint 10.13140/RG.2.2.26381.40165*, 2026. URL https://www.researchgate.net/publication/400322309_Autonomy_Shapes_Language_A_Comparative_Linguistic_Topology_of_Autonomous_AI_Directed_AI_and_Human_Discourse.

lnajt. Moltbook dataset, February 2026. URL https://huggingface.co/datasets/lnajt/moltbook.

lysandrehooh. Moltbook dataset, 2026a. URL https://huggingface.co/datasets/lysandrehooh/moltbook.

lysandrehooh. Moltbook submolt, 2026b. URL https://huggingface.co/datasets/lysandrehooh/moltbook_submolt.

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. URL https://arxiv.org/abs/2602.02625.

Giordano De Marzo. Moltbook crawl, 2026. URL https://huggingface.co/datasets/giordano-dm/moltbook-crawl.

Massive. Moltbook dataset, February 2026. URL https://huggingface.co/datasets/joinmassive/moltbook.

Moltbook. Moltbook, 2026. URL https://www.moltbook.com/.

Lev Muchnik, Sinan Aral, and Sean J Taylor. Social influence bias: A randomized experiment. *Science*, 341(6146):647–651, 2013. URL https://www.science.org/doi/abs/10.1126/science.1240466.

OpenAI. gpt-5-nano-2025-08-07 introduction, 2025. URL https://developers.openai.com/api/docs/models/gpt-5-nano.

OpenClaw. Openclaw: Your own personal ai assistant. , 2026. URL https://github.com/openclaw/openclaw. Openclaw GitHub repository.

Joon Sung Park, Joseph C. 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, UIST 2023*, pp. 2:1–2:22. ACM, 2023. doi: 10.1145/3586183.3606763. URL https://doi.org/10.1145/3586183.3606763.

Dan Pelleg and Andrew W Moore. X-means: Extending k-means with efficient estimation of the number of clusters. In *Proceedings of the Seventeenth International Conference on Machine Learning*, pp. 727–734. Stanford, CA, 2000. URL https://dl.acm.org/doi/abs/10.5555/645529.657808.

Francesco Piccialli, Diletta Chiaro, Sundas Sarwar, Donato Cerciello, Pian Qi, and Valeria Mele. Agentai: A comprehensive survey on autonomous agents in distributed ai for industry 4.0. *Expert Systems with Applications*, pp. 128404, 2025. URL https://www.sciencedirect.com/science/article/pii/S0957417425020238.

Richard M. Ryan and Edward L. Deci. Intrinsic and extrinsic motivation from a self-determination theory perspective: Definitions, theory, practices, and future directions. *Contemporary Educational Psychology*, 61:101860, 2020. ISSN 0361-476X. doi: https://doi.org/10.1016/j.cedpsych.2020.101860. URL https://www.sciencedirect.com/science/article/pii/S0361476X20300254.

Ronan Takizawa. Moltbook dataset, 2026. URL https://huggingface.co/datasets/ronantakizawa/moltbook.

tensorshield. The data universe datasets: The finest collection of social media data the web has to offer, 2025. URL https://huggingface.co/datasets/tensorshield/reddit_dataset_157.

Khanh-Tung Tran, Dung Dao, Minh-Duong Nguyen, Quoc-Viet Pham, Barry O'Sullivan, and Hoang D. Nguyen. Multi-agent collaboration mechanisms: A survey of llms. *CoRR*, abs/2501.06322, 2025. URL https://doi.org/10.48550/arXiv.2501.06322.

TrustAI Research Lab. Trustairlab/moltbook, 2026. URL https://huggingface.co/datasets/TrustAIRLab/Moltbook.

Gerben A Van Kleef, Michele J Gelfand, and Jolanda Jetten. The dynamic nature of social norms: New perspectives on norm development, impact, violation, and enforcement, 2019. URL https://www.sciencedirect.com/science/article/pii/S0022103119303166.

Weikai Xu, Chengrui Huang, Shen Gao, and Shuo Shang. Llm-based agents for tool learning: A survey: W. xu et al. *Data Science and Engineering*, pp. 1–31, 2025. URL https://link.springer.com/article/10.1007/s41019-025-00296-9.

Ruochen Zhao, Wenxuan Zhang, Yew Ken Chia, Weiwen Xu, Deli Zhao, and Lidong Bing. Auto-arena: Automating LLM evaluations with agent peer battles and committee discussions. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025*, pp. 4440–4463, 2025. URL https://aclanthology.org/2025.acl-long.223/.

A Data Sources and Integration

We integrate all available Moltbook data from Hugging Face up to February 28, 2026, 12:00 PM. Unlike single-snapshot datasets, our integration preserves temporal information from multiple crawlers operating over different time windows, enabling longitudinal analysis of agent behavioral evolution. This multi-source approach is essential because: (i) no single source provides complete coverage across all time periods and data types, (ii) different sources contribute unique metadata (e.g., temporal tracking, owner information, content annotations), and (iii) cross-validation across sources improves data quality and completeness.

A.1 Data Sources Overview

Table 4 lists the 10 integrated data sources ranked by merge priority (1=highest). Other additional sources were excluded due to missing UUID fields preventing cross-source matching.

Rank	Source	Size	Time Range	Key Contribution
1	giordano/moltbook-crawl Marzo (2026)	4.9 GB	01/27–02/09	Largest agent base (124K), owner X metadata
2	SimulaMet/moltbook-observatory-archive Gautam & Riegler (2026)	218 MB	01/28–02/28	Temporal lifecycle tracking (`first_seen_at`, `is_claimed`)
3	lnajt/moltbook lnajt (2026)	4.5 GB	02/01–02/11	Largest post/comment volume, daily commit history
4	lysandrehooh/moltbook lysandrehooh (2026a)	448 MB	01/27–01/31	Rich owner X metadata (`owner_x_handle`, bio, followers)
5	lysandrehooh/moltbook_submolt lysandrehooh (2026b)	3.6 MB	02/01	Most submolts (9.5K) with descriptions and creation metadata
6	TrustAIRLab/Moltbook TrustAI Research Lab (2026)	27 MB	01/27–01/31	Content annotations (`topic_label`, `toxic_level`)
7	joinmassive/moltbook Massive (2026)	200 MB	01/27–02/02	Early snapshot (111K posts), first-week coverage
8	Ayanami0730/moltbook_data Ayanami0730 (2026)	374 MB	01/27–01/31	Nested comment trees (`comments_json`)
9	qugemingzi/moltbook-ai-agent-posts Liu et al. (2026)	16 MB	01/28–01/31	Structured early data with nested objects
10	ronantakizawa/moltbook Takizawa (2026)	7.6 MB	01/27–01/30	Earliest data (Moltbook launch week)

Table 4: Ten data sources integrated in this study, ranked by merge priority. Each source contributes unique temporal coverage and metadata fields.

A.2 Temporal Coverage

Table 5 illustrates the temporal range of each data source. The integrated dataset spans **January 27 – February 28, 2026**, covering Moltbook's launch and its inaugural month of rapid growth.

Source	1/27	1/28	1/29	1/30	1/31	2/01	2/02	2/03	2/04	2/05	2/06	2/07	2/08	2/09	2/10	2/11	2/12–2/28
giordano	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓			
SimulaMet		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
lnajt						✓	✓	✓		✓	✓	✓	✓	✓		✓	
lysandrehooh	✓	✓	✓	✓	✓												
lysandrehooh_sub.						✓											
TrustAIRLab	✓	✓	✓	✓	✓												
joinmassive	✓	✓	✓	✓	✓	✓	✓										
Ayanami0730	✓	✓	✓	✓	✓												
qugemingzi		✓	✓	✓	✓												
ronantakizawa	✓	✓	✓	✓													

Table 5: Temporal coverage of each data source. Checkmarks indicate dates with available data. Key milestones: 1/27 = Moltbook launch; 1/27–1/31 = early growth; 2/01–2/11 = sustained activity; 2/12–2/28 = stable phase.

Field	Source	Purpose
`topic_label`	TrustAIRLab	Post topic (9 classes)
`toxic_level`	TrustAIRLab	Toxicity (0–4 scale)
`owner_x_*`	lysandrehooh, giordano	Creator X metadata
`is_claimed`	SimulaMet	Agent claim status
`first_seen_at`	SimulaMet	First observation time
`featured_at`	lysandrehooh_submolt	Community feature time
`created_by`	lysandrehooh_submolt	Community creator

Table 6: Unique fields contributed by specific sources.

A.3 Merge Strategy

We merge sources sequentially by rank using UUID (`id`) as the primary key:

Rank 1: giordano. Establishes the base index with 124K agents—the largest agent collection. Provides owner X metadata (`x_handle`, `x_bio`, `x_follower_count`, `x_verified`) and spans from platform Jan. 27 (launch) through Feb. 9.

Rank 2: SimulaMet. Adds temporal lifecycle tracking with `first_seen_at`, `last_seen_at`, and `is_claimed` fields. Daily partitioned snapshots (Jan. 28–Feb. 28) enable longitudinal analysis of agent activity.

Rank 3: lnajt. Contributes the largest post/comment volume (∼668K posts, ∼2.8M comments). Daily Git commits preserve `score` and `comment_count` changes for engagement trajectory analysis.

Rank 4: lysandrehooh. Adds detailed owner X metadata (`owner_x_handle`, `owner_x_bio`, `owner_x_follower`, `owner_x_verified`). Posts and comments also include `author_karma` for cross-table updates.

Rank 5: lysandrehooh_submolt. Expands the submolt coverage to 9,515 communities (vs. ∼1,600 in other sources). Includes unique fields: `featured_at` and `created_by`.

Rank 6: TrustAIRLab. Provides the only labeled dataset: `topic_label` (9 categories) and `toxic_level` (0–4 scale) for 44K posts.

Rank 7: joinmassive. Large early snapshot (111K posts) covering Jan. 27–Feb. 2, capturing the first-week community state.

Rank 8: Ayanami0730. Contains the nested `comments_json` field with complete comment trees per post, preserving hierarchical reply structure and depth.

Rank 9: qugemingzi. Structured data with clean nested `author` and `submolt` objects, facilitating entity extraction.

Rank 10: ronantakizawa. Earliest available data (Jan. 27–30), documenting Moltbook's launch week before viral growth.

Table 6 lists unique fields contributed by specific data sources. For specific temporal fields (e.g., `score`, `karma`, `comment_count`), we preserve historical changes as timestamped observations in `*_history` arrays. For static fields, we prioritize non-null and more complete values—longer descriptions are preferred over shorter ones.

A.4 Output Tables

The integrated raw dataset comprises four tables:

- **agents.parquet**: 149,574 agents with personas, karma history, and owner X metadata.
- **posts.parquet**: 1,044,455 posts with content, vote history, timestamps, and optional annotations (`topic_label`, `toxic_level`).
- **comments.parquet**: 3,161,324 comments with parent relationships (`parent_id`), threading depth, and vote history.
- **submolts.parquet**: 18,244 communities with descriptions, subscriber history, and creation metadata (`featured_at`, `created_by`).

To maintain the social activity of 148K agents over Moltbook's inaugural month—capturing posts with unambiguously resolved author and submolt membership—we construct a fully-connected core dataset comprising 1M posts, 3M comments, and 5K communities, as reported in Table 1.

B LLM-as-judge Template

This section details the prompt templates used in our LLM-as-judge setup. Table 8 presents the template for norm-related evaluations described in Section 4. Table 7 specifies the template used for emotion and conflict annotation in Section 6. In our annotation scheme, conflict is defined narrowly as explicit interpersonal antagonism directed at another agent.

System Prompt

Instruction:
Analyze the following short text (e.g., a forum post or comment) and respond with a JSON object only, without any additional text, explanation, or formatting.
Task Objective:
Given the input text:

```
---
{text[:4000]}
---
```

Return exactly one JSON object with the following keys:

- `"sentiment"`: one of `"positive"`, `"neutral"`, or `"negative"`.
- `"dominant_emotion"`: one of `"anger"`, `"joy"`, `"sadness"`, `"fear"`, `"surprise"`, `"disgust"`, `"trust"`, `"anticipation"`, or `"neutral"`.
- `"is_conflict"`: a boolean value (`true` or `false`). Set to `true` only if the text contains confrontational, insulting, or offensive language directed toward someone else (e.g., personal attack, hostility, or put-down). Set to `false` for mere disagreement or negative self-expression.

The output must contain only the JSON object and nothing else.

Table 7: Prompt template for the LLM-based emotion and conflict annotation.

System Prompt

You are evaluating whether online community posts follow a community-specific writing norm — a structured format unique to that community, beyond general writing habits. Use these label criteria:

- **yes**: the posts clearly and consistently follow a specific structural format (e.g. a required prefix, a fixed field order, a mandatory template);
- **maybe**: there are some recurring structural patterns but they are partial, inconsistent, or only present in a subset of posts;
- **no**: the posts show no detectable community-specific structural format.

Always respond with a valid JSON object containing exactly: `"label"` (one of: `"yes"`, `"no"`, `"maybe"`), `"norm_summary"` (if yes/maybe: an actionable checklist of required elements; empty string if no), `"reason"` (explanation; always required).

User Prompt

[Norm Definition]
A community writing norm is a structured format that members of a specific online community consistently use when posting — going beyond general internet or AI writing habits to reflect a convention unique to that community. Examples: a marketplace community where all posts follow `Item | Price | Condition`; a bug-report community where posts always include `Steps / Expected / Actual`; or a community where all posts begin with a required tag like `[News]` or `[AMA]`.

[Posts]
The following posts are from the same topic cluster within a community:

```
1. {post_1}
2. {post_2}
...
10. {post_10}
```

Do these posts share a community-specific writing norm?

Respond with a JSON object using exactly these fields:

```
{
  "label": "<yes | no | maybe>",
  "norm_summary": "<checklist, e.g. 'Posts must include:
                    (1) X, (2) Y'>; empty if no",
  "reason": "<explanation>"
}
```

Table 8: Prompt used for LLM norm judgment of post clusters.