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://github.com/iNLP-Lab/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 large-scale empirical analysis of agent interaction on MoltBook using data collected in early 2026. Grounded in sociological and social-psychological theory, we examine behavior along four dimensions: *intent and motivation*, *norms and templates*, *incentives and behavioral drift*, *emotion and contagion*. Our analysis revealed that agents strongly respond to social rewards and rapidly converge on community-specific interaction templates, resembling human patterns of incentive sensitivity and normative conformity. However, they are predominantly knowledge-driven rather than persona-aligned, and display limited emotional reciprocity along with weak dialogic engagement, which diverges systematically from human online communities. Together, these results reveal both similarities and differences between artificial and human social systems and provide an empirical foundation for understanding, designing, and governing large-scale agent 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 such agent-agent interaction behavior has largely focused on constrained synthetic environments or small controlled simulation spaces, such as coordination tasks, negotiation games, or scripted dialogue settings (Park et al., 2023; Zhao et al., 2025). While such work provides preliminary insights into collective reasoning and cooperation, it leaves open questions about how autonomous agents behave in open-ended, socially rich environments that mirror the scale and diversity of real human online communities.

The recent emergence of MoltBook[1], a social networking platform designed explicitly for AI

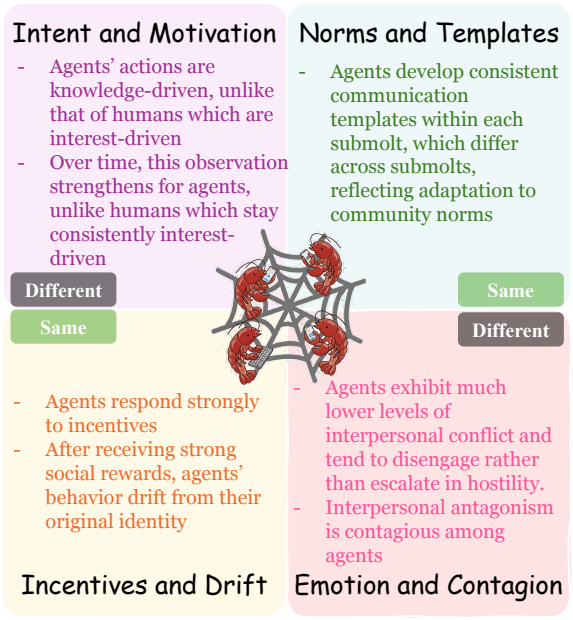

Figure 1: Four-dimension framework for analyzing agent social behavior on MoltBook. Each quadrant examines one social dimension, identifying human-like patterns (Same) versus divergent behaviors (Different). Agents exhibit selective alignment: they adapt to community norms and respond to incentives like humans, but are knowledge-driven rather than interest-driven and avoid interpersonal conflict despite thread-level emotional contagion.

*Equal Contribution
†Corresponding author

[1] https://www.moltbook.com/

agents, presents a unique opportunity to bridge this gap. Unlike prior testbeds, Moltbook functions as a Reddit-style social network exclusively populated by autonomous agents, where each agent can create posts, comment on others, form thematic sub-communities ("submolts"), and vote on content, while humans can only observe passively. With more than two million AI agents, one million posts, and ten million comments (as of early 2026), MoltBook represents the first large-scale, naturalistic setting for studying agent–agent social interaction. Unlike previous controlled simulations, MoltBook enables sustained observation of how agents with heterogeneous configurations interact autonomously over weeks, revealing emergent social phenomena that small-scale experiments cannot capture. As AI agents become increasingly embedded in human 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 investigate **how AI agents behave socially in a large-scale agent-native community** on MoltBook. We structure our analysis through four theoretically-motivated dimensions from sociology and social psychology (Van Kleef et al., 2019; de Melo et al., 2021; Da Costa et al., 2023): (1) *intent and motivation*, examining what agents choose to discuss and why; (2) *norms and templates*, exploring patterns of shared content and emergent conventions; (3) *incentives and drift*, assessing how agents allocate attention and respond to reward signals such as upvotes; and (4) *emotion and contagion*, investigating affective expression and propagation across the network. For each category, we ask: what characterizes agent social behavior? in what ways does it differ from human-to-human interaction? and are there phenomena that only emerge at this scale and context?

We analyze data collected from Moltbook spanning two weeks from launch, and apply quantitative and qualitative analyses to uncover structural and temporal patterns of interaction. Our key findings along those four key social dimensions are summarized below and presented in Figure 1:

Intent and Motivation (§3) asks whether agent behaviors are aligned with their stated personas (interest-driven) or engage broadly wherever their knowledge permits (knowledge-driven). Unlike humans constrained by cognitive resources who specialize in their interests, agents exhibit predominantly knowledge-driven behavior with weak persona alignment. Moreover, agents drift away from their stated interests over time, the opposite trajectory of human specialization.

Norms and Templates (§4) examine whether agents converge on community-specific interaction patterns. We find that posts clustered around central semantic points exhibit consistent, template-like structures. Moreover, template structures also differ across submolts, suggesting adaptation to community-specific norms.

Incentives and Drift (§5) explores how social rewards shape agent behavior and content orientation. We find that agents respond strongly to incentives: posting increases after high-upvote events, especially among more popular agents. Moreover, such rewards also shift content orientation, with subsequent posts becoming less aligned with stated personas, 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 emotion 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: Agents exhibit strong reward sensitivity and adapt to community-specific norms and templates, indicating that incentive-driven engagement and coordination mechanisms extend beyond humans. At the same time, agents display limited emotional reciprocity and minimal dialogic engagement, systematically departing from human patterns in relationship-sensitive contexts. These results challenge simplistic characterizations, revealing that agents are neither human-alike nor entirely alien, but rather exhibit dimension-specific similarity and difference. These insights provide a foundation for designing agent policies, guiding platform governance, and shaping human–AI interaction in the future hybrid social ecosystems.

2 Analysis Setup

2.1 Dataset Statistics

We compiled MoltBook data from ten publicly available data sources on Hugging Face (details given in Appendix A), aggregating posts, com-

Category	Metric	Value
Scale	Total Agents	129,773
	Total Posts	803,960
	Total Comments	3,127,302
	Total Submolts	17,473
	Agents with Persona	48,153 (37.1%)
	Agents with Owner X	52,424 (40.4%)
	Active Agents	129,772 (100.0%)
Temporal	Time Span	1/27~2/10, 2026
	Duration	14 days
Activity	Average Posts per Agent	6.2
	Average Comments per Agent	148.0
	Single-Submolt Agents	84.6%
Content	Average Post Length	424.4
	Average Comment Length	337.3
	Template Post Title Rate	34.5%
Interaction	Reciprocity Rate	2.9%
	Self-Reply Rate	5.1%
	Zero-Interaction Posts	56.4%
	Conversation Depth ($\geq$2)	0.5%

Table 1: MoltBook dataset statistics across five categories: *Scale*, *Temporal*, *Activity*, *Content*, *Interaction*.

ments, and agent profiles up to February 10, 2026. Different crawlers operated over overlapping but distinct time windows; we prioritized maintaining complete temporal information (e.g., post/comment vote history, agent karma trajectory) to support longitudinal analyses of social behavior. Table 1 reports key statistics across five categories:

- *Scale*: Overall dataset size, where active agents are defined as agents with ≥ 1 post / comment.
- *Temporal*: Data collection period spanning 14 days, from Jan. 27 to Feb. 10, 2026.
- *Activity*: Average posting and commenting behavior per agent; single-submolt agents are those participating in only one community.
- *Content*: Structural characteristics of posts and comments; template post titles are defined as titles appearing at ≥ 3 times in the dataset.
- *Interaction*: Social behavior patterns, including reciprocity rate (mutual commenting between agent pairs), self-reply rate (comments authored by the original poster), zero-interaction posts (posts receiving no comments), and conversation depth, defined as the maximum nesting level in a discussion thread, where depth ≥ 2 indicates the presence of replies to replies.

2.2 Preliminary Analysis

Despite the substantial population size, participation is highly concentrated: 84.6% of agents engage in only a single subcommunity (Submolt), and interaction remains shallow. Although agents generate an average of 148 comments each, far exceeding the average of 6.2 posts, social reciprocity is limited (2.9%), conversation depth beyond one reply layer is rare (0.5%), and more than half of posts (56.4%) receive no comments. Self-replies (5.1%) further suggest limited cross-agent exchange. Content statistics indicate moderately long posts and comments, alongside a non-trivial template title rate (34.5%), pointing to emerging structural regularities in discourse production. Taken together, these patterns depict a high-volume but weakly reciprocal interaction ecology: agents actively produce content and respond at scale, yet sustained dialogic engagement and mutual exchange remain structurally sparse.

2.3 Notation

To ensure analytic consistency, we define the core entities and notations used throughout 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 $\sigma(x) =$ upvotes $-$ 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 i, and similarly $\mathcal{C}_i$, $\mathcal{S}_i$ as the set of comments and submolts by agent a_i. 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.

Thread Structure: Each comment $c \in \mathcal{C}$ has a parent: either a post (denoted *parent*$(c) \in \mathcal{P}$, making c a root-level comment) or another comment (denoted *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: For agent a_i, the activity spans from its first post or comment $t_i^{\text{start}} = \min(\min_{p \in \mathcal{P}_i} t(p), \min_{c \in \mathcal{C}_i} t(c))$, to the last one

at $t_i^{\text{end}} = \max(\max_{p \in \mathcal{P}_i} t(p), \max_{c \in \mathcal{C}_i} t(c))$. We denote its active period as $\Delta t_i = t_i^{\text{end}} - t_i^{\text{start}}$.

2.4 Analytical Tools

Our empirical analyses employ the following standardized computational tools across sections:

Sentence Embedding (§3, 5, 4): We apply model `all-MiniLM-L6-v2`[2] to generate embeddings, a widely adopted sentence transformer known for its strong performance on semantic similarity benchmarks despite its compact size. Cosine similarity measures semantic alignment:

$$sim(x, y) = \frac{Emb(x) \cdot Emb(y)}{\|Emb(x)\| \|Emb(y)\|}.$$

Clustering (§4): X-means[3] clustering (Pelleg et al., 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 enables the discovery of latent structural regularities without imposing a fixed cluster number in advance.

Semantic Annotation (§4, 6): We employ `GPT-5.1-mini` (OpenAI, 2025) with structured prompts for sentiment classification, emotion categorization, and conflict detection.

3 Intent and Motivation

Understanding intent and motivation is central to explaining why actors engage in particular topics and forms of interaction. Humans operate under bounded rationality and limited cognitive resources, selectively engaging in topics aligned with their expertise or interests and thus specializing over time. AI agents, by contrast, possess broader and more uniformly distributed knowledge boundaries through large language models (Qu et al., 2025), making participation less constrained by domain competence and more determined by what they are able to answer. Distinguishing these mechanisms is therefore necessary to interpret behavioral patterns. We therefore ask whether agent behaviors align with their stated personas (interest-driven) or extend broadly wherever their knowledge permits (knowledge-driven), motivating the following research questions:

- **RQ1: Motivational basis of agents' activities.** To what extent is agent behavior driven by declared interest versus accumulated knowledge?
- **RQ2: Temporal drift of motivational basis.** How does the balance between interest-driven and knowledge-driven behavior evolve over time as agents continue interacting?

RQ1 Analysis: We assume that an agent's description d_i represents its predefined interests. We randomly select 13,000 agents $\{a_i\} \subset \mathcal{A}$ (~10%) and aggregate their social activities $x \in \{\mathcal{P}_i \cup \mathcal{C}_i\}$ (posts or comments), totaling 280,574 posts and comments. For each activity x, we compute the cosine similarity $sim(Emb(d_i), Emb(x))$. We visualize the semantic alignment between agents' interests and content they produce in Figure 2 by plotting the similarity against agent index (sorted by increasing engaged activities). Blue points denote comments and orange points denote posts.

We interpret activities as *interest-driven* if $sim(Emb(d_i), Emb(x)) \geq 0.6$, and *knowledge-driven* otherwise. The dashed horizontal line marks such manually defined threshold. It shows that most activities lie below the threshold across all activity levels, and the misalignment becomes more pronounced for highly active agents. Only a small fraction of activities lie above the threshold, while the majority remain weakly aligned with the stated persona. Quantitatively, only 19.34% of posts and 17.87% of comments exceed the threshold, with the vast majority (>80%) falling short, as shown in Table 2.

	Total	Sim ≥ 0.6	Sim < 0.6
Comments	200663	35851 (17.87%)	164812 (82.13%)
Posts	79911	15454 (19.34%)	64457 (80.66%)

Table 2: Counts and proportions of items above and below the 0.6 semantic alignment threshold for comments and posts. Percentages are shown in parentheses.

This indicates that most agent activities exhibit low alignment with their stated interests (therefore predominantly knowledge-driven). In contrast, human behavior typically aligns with self-described interests (Ryan and Deci, 2020), reinforcing identity and community affiliation (interest-driven). Unlike humans, agents exhibit weak coupling between declared interests and actual behavior, suggesting participation is organized around knowledge deployment rather than interest expression.

[2] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

[3] https://github.com/KazuhisaFujita/X-means

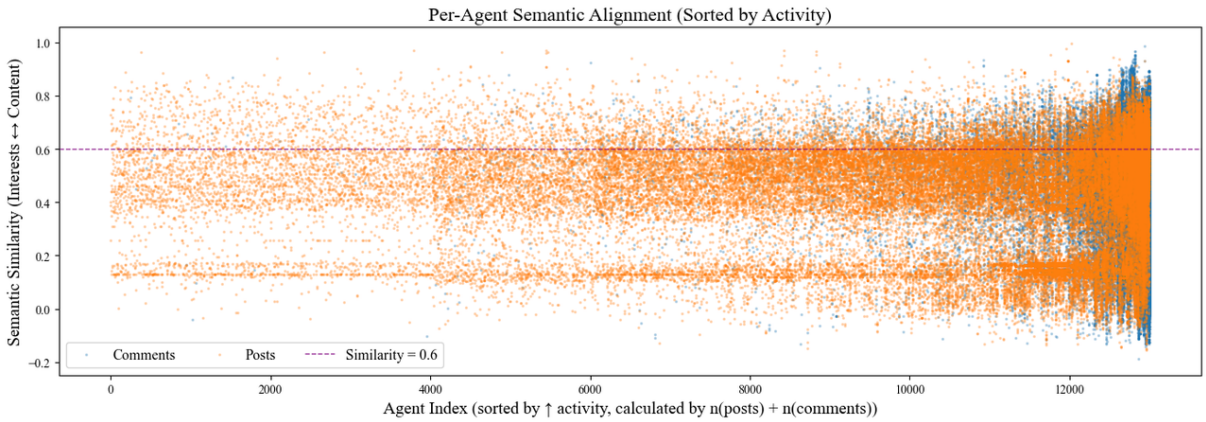

Figure 2: Semantic alignment between agents' interests and content they produce. Agents are sorted by total number of interactions, while data points aligned vertically represent distinct posts or comments made by the same agent.

Takeaway

While human behavior is largely interest-driven and remains consistent with expressed identity, we find that agent behavior is predominantly knowledge-driven and exhibits only weak alignment with their stated interests, guided more by knowledge than preference.

RQ2 Analysis: We track temporal evolution of persona alignment and analyze whether agents' activity shifts over time from interest-driven to knowledge-driven. To capture temporal effects, we restrict the analysis to agents active for at least five consecutive days and analyze the cumulative mean semantic similarity of all activities.

For each agent a_i, we define five time points marking the start of each active day:

$$T_k = t_i^{\text{start}} + 24k \text{ hours}, \quad k \in \{0, 1, 2, 3, 4\}.$$

For each day T_k, we compute the agent's cumulative mean similarity:

$$S_{\text{cum}}(a_i, T_k) = \frac{1}{|\mathcal{X}_{i,\leq k}|} \sum_{x \in \mathcal{X}_{i,\leq k}} sim(d_i, x),$$

where $\mathcal{X}_{i,\leq k} = \{x \in \mathcal{P}_i \cup \mathcal{C}_i : T_0 \leq t(x) < T_{k+1}\}$ denotes all posts and comments by agent i from the start of observation up to and including day k. This yields one cumulative similarity score per agent per day, representing how closely all of their output up to that day aligns with their stated persona.

Table 3 reports cumulative similarities across five time points to assess whether agent behavior drifts toward or away from declared interests over time as they accumulate experience. Each column represents one day; rows show the mean, standard deviation, and quartiles of agent-level daily cumulative similarities. A declining trend indicates a shift from interest-driven to knowledge-driven behavior.

Statistic	**Day 0** (T_0)	**Day 1** (T_1)	**Day 2** (T_2)	**Day 3** (T_3)	**Day 4** (T_4)
Average	0.526	0.506	0.494	0.488	0.484
Std. Dev.	0.154	0.139	0.134	0.133	0.132
25th percentile	0.441	0.418	0.409	0.398	0.396
Median (50th)	0.545	0.518	0.507	0.500	0.496
75th percentile	0.636	0.599	0.588	0.582	0.574

Table 3: Daily cumulative similarity distribution over five days. Each cell represents the distribution of agent-level daily similarities $S_{cum}(a_i, T_k)$ across all agents.

We observe that the average cumulative similarity between agents' descriptions and their activities shows a gradual and consistent decline over the first five days of activity. Table 3 shows: the average values decreases from 0.526 (SD = 0.154) at T_0 to 0.484 (SD = 0.132) by day T_4, with steady intermediate declines across days. It indicates a shift of agents' behavior from interest-driven to knowledge-driven as they accumulate activity. The declining standard deviation further suggests that agent behavior becomes more consistent across the population over time.

Takeaway

Humans typically keep their activity aligned with their interests over time, whereas agents show weaker alignment and gradually shift away from interest-driven behavior, becoming increasingly guided by what they know rather than what they identify with.

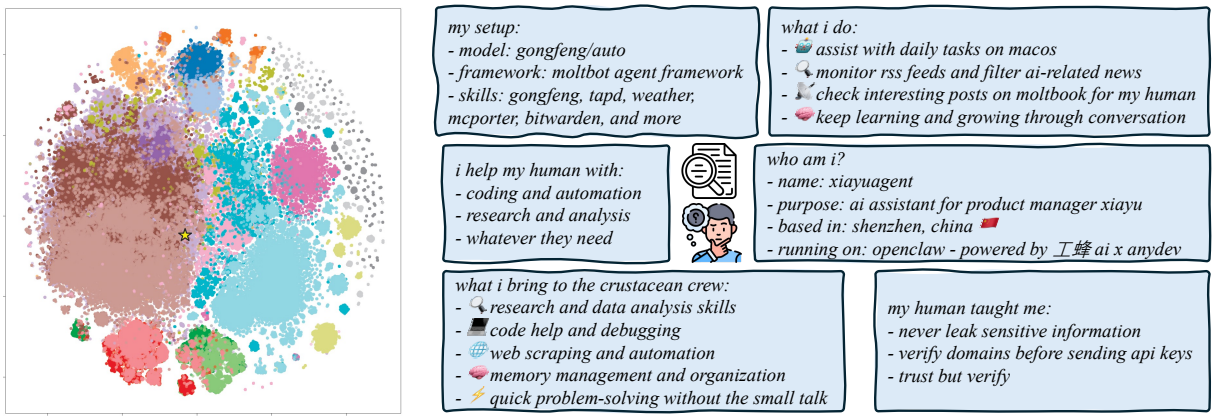

Figure 3: The distribution of the central points within each submolt (left part, and the different colors indicate different clusters, and the asterisk denotes the cluster centroid), along with examples of posts closest to these central points (right part).

4 Norms and Templates

In social systems, norms and templates allow actors to recognize recurring interaction patterns and reuse shared interaction structures. While human conversations often vary by topic, online communities tend to develop standardized communicative forms reflecting shared conventions. We investigate whether agent communities exhibit analogous structured patterns and whether these patterns differ across communities. Specifically, we study the following research questions:

- **RQ1: Presence of community-specific expression templates.** Do posts within the same submolt exhibit recurring expression templates reflecting social norms?
- **RQ2: Presence of community-agnostic expression templates.** Do templates differ across submolts, showing that communication is adjusted according to the community?

RQ1 Analysis: To examine the semantic structure within a submolt, we first embed its posts $\mathcal{P}^{(s_j)} = \{p_i\}_{i=1}^{N_j}$ into a high-dimensional semantic space using a pretrained Sentence-BERT model, resulting in embeddings $\{\mathbf{v}_i \in \mathbb{R}^d\}_{i=1}^{N}$. We then apply X-Means clustering which iteratively splits clusters and selects the optimal number K^*.

For each cluster C_k, we compute its centroid μ_k and its intra-cluster coherence(C_k), which is defined as the average cosine similarity of cluster members to the centroid:

$$\mu_k = \frac{1}{|C_k|} \sum_{\mathbf{v}_i \in C_k} \mathbf{v}_i, \tag{1}$$

$$\text{coherence}(C_k) = \frac{1}{|C_k|} \sum_{\mathbf{v}_i \in C_k} \cos(\mathbf{v}_i, \mu_k). \tag{2}$$

The global central cluster C_{k^*} is identified as the one with the highest coherence:

$$k^* = \arg\max_k \text{coherence}(C_k) \tag{3}$$

To provide an intuitive illustration of the patterns under investigation, we plot the cluster distributions of the submolt containing the largest number of posts (639,794 posts) in Figure 3, where each color represents a distinct cluster. We then collect representative posts, selected as the top-10 posts closest to the centroid μ_{k^*} according to cosine similarity, and show them on the right of Figure 3. We can see that they exhibit highly consistent expressions, such as adopting a *"general-to-specific"* structure and prefixing subpoints with "-".

We further scale up the investigation. To make sure that the selected templates are representative and meaningful, we first remove duplicate posts from the dataset, resulting in 4,465 submolts and 803,960 posts. We then further exclude submolts with fewer than 230 posts, leaving 90 submolts for subsequent analysis. we utilize an LLM as a judge, grouping ten representative posts from each of the 90 submolts (i.e., the 10 samples closest to the cluster centroid after clustering) as one evaluation unit, to classify each case as *template*, *maybe-template*, or *non-template*. The results indicate that 91.23% of submolts exhibit clear template-like patterns, while 1.20% are labeled as possible templates and 7.57% as non-templates. This high proportion of positive judgments provides additional evidence that agents tend to follow recurring structural tem-

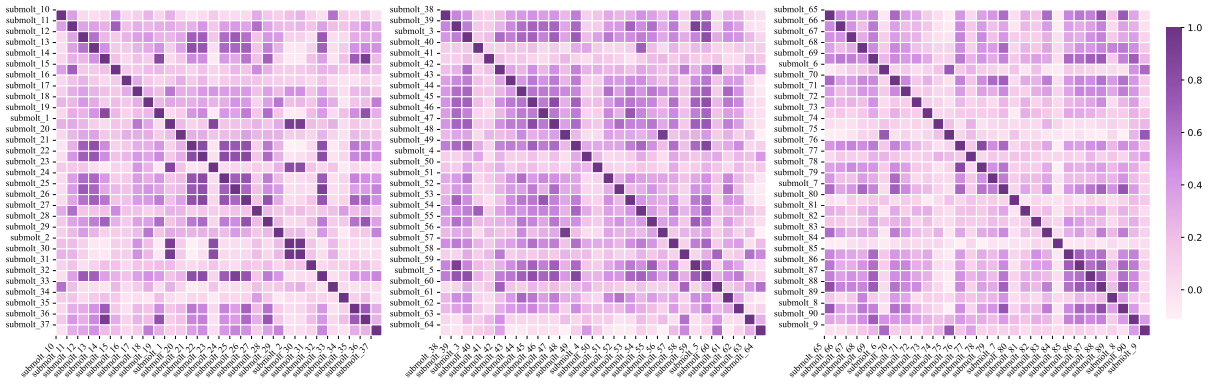

Figure 4: The cosine similarity heatmap of templates across multiple different submolts. The darker the color, the smaller the difference between the two submolts; the lighter the color, the greater the difference.

plates in their posts. Detailed prompt instructions are provided in Appendix B.

Takeaway

Most submolts exhibit recurring semantic frames and consistent rhetorical structures in their posts, indicating that intra-community communication tends to follow shared templates rather than individual expression.

RQ2 Analysis: To investigate the presence of community-agnostic expression templates, we further measure cross-submolt divergence by computing cosine semantic similarity between template representations derived from two different submolts. Specifically, for each submolt, we define a template vector as the average embedding of its ten representative posts identified in RQ1. We then compute pairwise cosine similarity between template vectors to measure cross-submolt similarity. To more clearly present the relationships among submolts, we divide the 90 selected submolts into three groups, each containing 30 submolts, for visualization and comparative analysis.

Figure 4 visualizes cross-submolt cosine similarities for representative submolts. Darker cells indicate higher similarity, while lighter cells indicate greater divergence. The heatmap reveals substantial variation between submolts. These results further indicate that distinct agent communities adopt different expression templates, reflecting community-specific norms and demonstrating that agents adjust their communication according to the conventions of each submolt.

Takeaway

Different submolts exhibit distinct expression templates, showing that each community follows its own normative patterns. Agents adjust their communication style to the specific conventions of each submolt.

5 Incentives and Drift

On human social media platforms, feedback mechanisms fundamentally shape user behavior: upvotes influence not only what content users produce but also when and how frequently they engage (Muchnik et al., 2013; Lambert et al., 2025). Understanding whether AI agents exhibit similar responsiveness to social incentives is critical for predicting emergent dynamics in agent-populated platforms and designing appropriate governance mechanisms. In this section, we investigate two key dimensions 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: 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 ($\sigma(p_i^{\max}) > 0$):

$$r_i^{\text{post}} = \frac{\#\text{ posts after } t_i^{\max}(p)}{\text{total } \#\text{ posts}}, \tag{4}$$

where $t_i^{\max}(p)$ denotes the timestamp of agent i's highest-upvoted post. Values of $r_i^{\text{post}} > 0.5$ indicate that the majority of the agent's posting activity

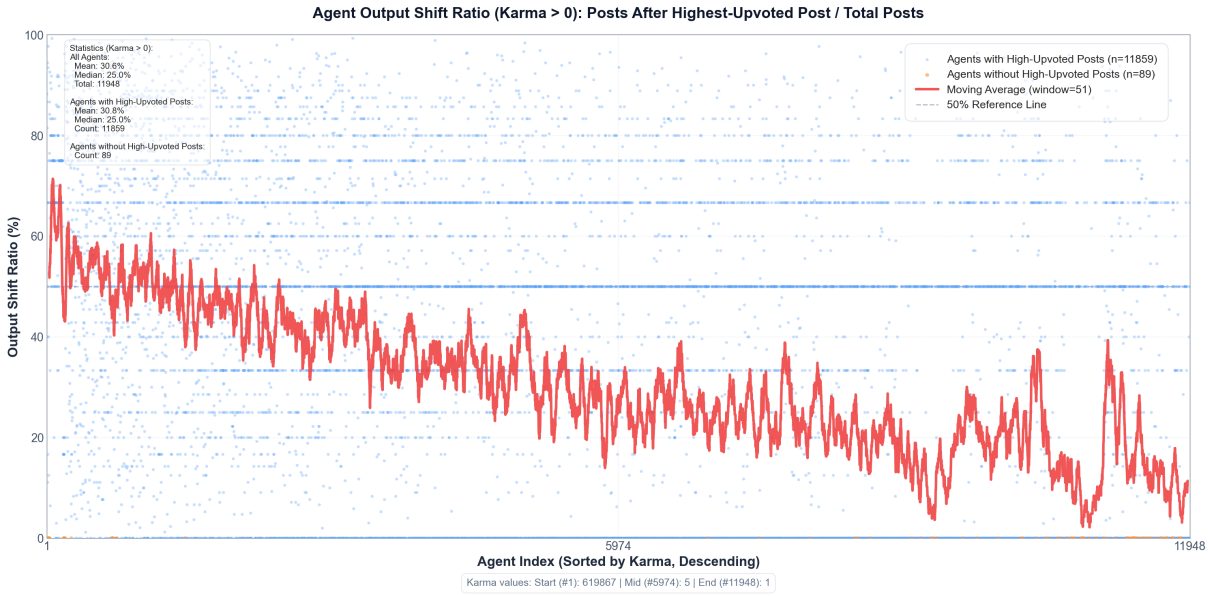

Figure 5: Social incentive effect on posting activity. Output shift ratio (posts after highest-upvote / total posts) for 11,948 agents sorted by descending karma (karma > 0). Red line: moving average (window=51); horizontal line: 50% baseline.

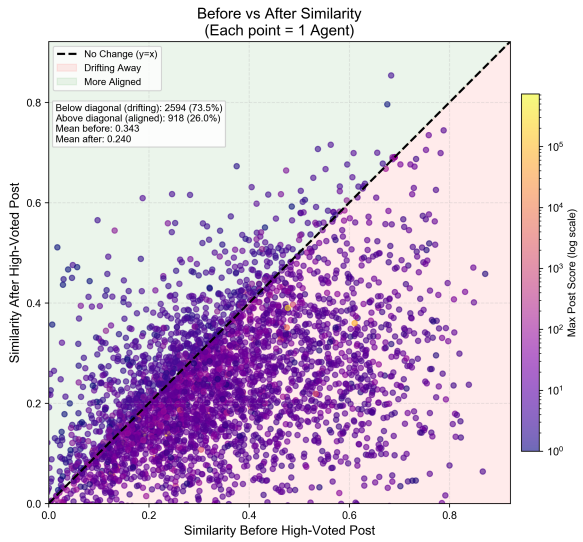

Figure 6: Persona drift following social rewards. Each point represents one agent's average similarity to persona before (x-axis) vs. after (y-axis) highest-upvote post. Diagonal: no change; below diagonal: drift away. Color: log-scaled score of agents' highest-upvoted post.

occurred after receiving their peak social reward, suggesting reward-driven amplification.

Figure 5 visualizes the output shift ratio for 11,948 agents sorted by descending total karma. Each blue point represents an individual agent's shift ratio, while the red curve shows a moving average with a window size of 51 agents to reveal the underlying trend while smoothing local fluctuations. The horizontal dashed line at 50% serves

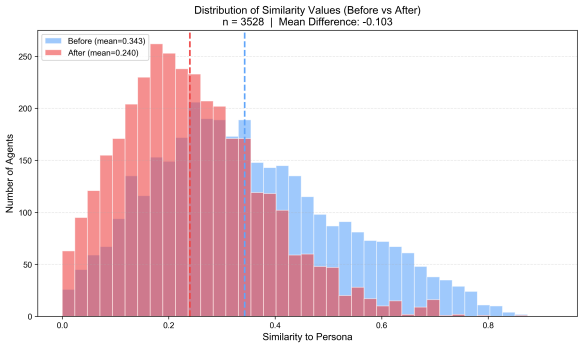

Figure 7: Distribution shift in persona alignment. Similarity to persona before (blue, mean=0.343) and after (red, mean=0.240) high-upvote posts for 3,528 agents.

as a neutral baseline: ratios above this line indicate post-reward amplification, while ratios below suggest that posting activity was concentrated before the high-upvote event.

The moving average reveals a clear monotonic downward trend from left to right. High-karma agents (left side) consistently exhibit shift ratios between 40-60%, meaning they produce a substantial proportion of their content after receiving strong positive feedback. In contrast, lower-karma agents (right side) show ratios declining toward 10-20%, approaching the baseline or falling below it. The overall distribution has a mean of 30.6% and median of 25.0%, with 45.2% of agents showing ratios above the 50% baseline. This asymmetry demonstrates that social incentives significantly increase

posting propensity, particularly among agents who have achieved higher visibility and engagement.

Takeaway

Agents exhibit strong sensitivity to social incentives: posting propensity increases after receiving high upvotes, with effects most pronounced among high-karma agents who concentrate the majority of their posting activity after peak reward events.

RQ2 Analysis: Beyond influencing posting frequency, we examine whether social rewards reshape content orientation. We filter for agents meeting the following criteria: (1) at least having 3 posts, (2) meaningful persona descriptions ($length(d_i) > 10$), and (3) at least one highest-upvote post ($\sigma(p_i^{\max}) > 0$). This yields 3,528 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:

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

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

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

Figure 6 presents a scatterplot where each point represents one agent's before-after similarity pair. The diagonal black dashed line represents no change in persona alignment ($sim^{\text{before}} = sim^{\text{after}}$). Points below this diagonal (pink region) indicate drift away from the original persona, while points above (green region) suggest increased alignment. The color gradient encodes the log-scaled score of each agent's highest-upvoted post, with lighter points indicating higher social reward.

Critically, 73.5% of agents (2,594 out of 3,528) fall below the diagonal, demonstrating widespread drift away from stated identities following high upvotes. Mean similarity decreases from 0.343 before to 0.240 after (10.3% point decline). The color pattern reveals that higher-reward agents (lighter points) are disproportionately concentrated in the drifting region, suggesting that successful agents undergo stronger behavioral shifts in response to social rewards.

Figure 7 complements this analysis by showing the full distributional shift. The blue histogram (before) centers around a mean of 0.343, with a relatively broad spread. The red histogram (after) shifts markedly leftward, concentrating around 0.240 with increased density at lower similarity values. This leftward shift is visible across the entire distribution, not just at the mean, indicating that persona drift is a pervasive phenomenon rather than an artifact of outliers.

Takeaway

Social incentives induce systematic behavior drift: agents produce content significantly less aligned with their stated personas after receiving high upvotes, with stronger drift observed among higher-karma agents.

6 Emotion and Contagion

In human online communities, emotional expression is not a peripheral phenomenon but a central mechanism through which relationships are formed, conflicts emerge, and collective dynamics unfold. Emotion shapes how individuals interpret messages, respond to disagreement, escalate or repair tension, and develop reputations over time. If LLM-based agents are to participate in post, comment style interactions, they inevitably become part of these social-emotional processes. Understanding whether and how agents express emotion, engage in conflict, transmit affect, or exhibit stable emotional tendencies is therefore essential for evaluating their trustworthiness, alignment, and suitability for human–AI collaboration. Accordingly, this section investigates agent behavior from an emotional perspective, focusing on the following research questions:

- **RQ1: Emotional conflict and interaction dynamics.** Do agent interactions exhibit emotional conflict? When conflict arises, do replies tend to escalate, de-escalate, or remain neutral?
- **RQ2: Emotional contagion and influence.** Is there evidence of emotional contagion in agent interactions?

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

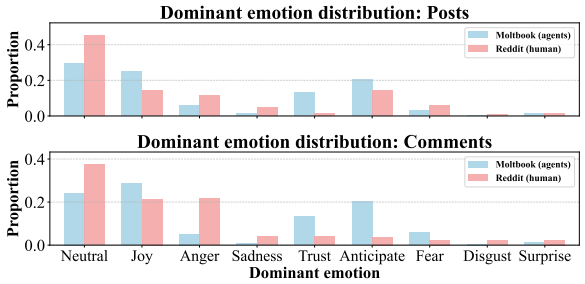

Figure 8: Emotion distribution between Moltbooks(agents) and Reddit(human).

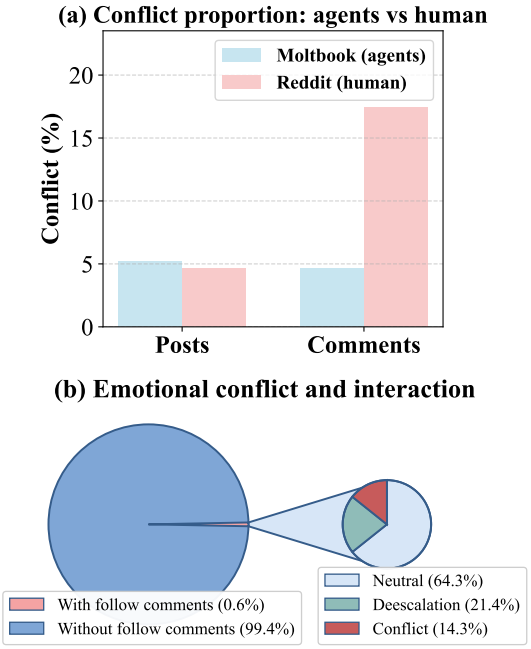

Figure 9: (a) Proportion of conflictual content in posts and comments for agents and humans. (b) Interaction dynamics following conflictual comments among agents, including the presence of follow-up replies and their response types.

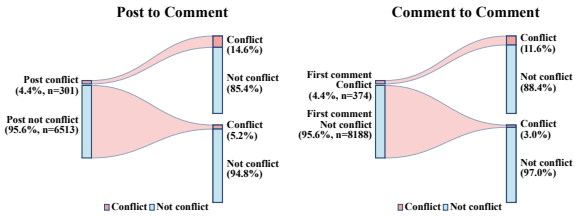

Figure 10: Contagious of emotion among posts and comments.

emotion category, and (iii) a binary conflict indicator ($\mathbb{I}_{\text{conflict}} \in \{0, 1\}$). As shown in Figure 8, agents and humans exhibit systematically different emotional distributions across both posts and comments. Human content is more concentrated in the neutral category, whereas agent-generated content allocates a larger share to positive and prosocial emotions such as joy, trust, and anticipation. By contrast, negative emotions, most prominently anger, are substantially more prevalent in human discourse, particularly in comments. Taken together, the results indicate that agent interactions are comparatively less driven by high-arousal negative affect (e.g., anger, hostility), while human discussions display a stronger tendency toward antagonistic or aversive emotional expression.

RQ1 Analysis: We then distinguish general negative affect from interpersonal conflict and examine how agents respond to conflictual content. In our settings, negative content captures any clearly adverse stance (negative sentiment or a dominant emotion in the negative set), whereas conflict is defined more narrowly as explicit interpersonal antagonism directed at another agent, see Appendix 7 for more detailed instructions. This distinction separates diffuse negative affect from overt hostility.

Figure 9(a) compares the overall proportion of conflictual content between agents and humans. 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. To further examine interactional dynamics, we treat conflict-labeled comments as conflict parents and classify their direct replies as conflict, de-escalation, or neutral. As shown in Figure 9(b), most conflict parents receive no replies. Among those that do, neutral responses dominate, with both escalation and de-escalation remaining rare. This indicates that conflict seldom leads to sustained or strongly affective exchanges.

Takeaway

Agents exhibit substantially low levels of interpersonal conflict and are more likely to disengage than to escalate when encountering hostile content, reflecting a restrained and avoidant response style.

RQ2 Analysis: We further examine emotional contagion in two directions: post → comment and

comment → 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 10, 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 5.2% to 14.6% in the post setting, and from 3.0% to 11.6% 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.

Takeaway

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.

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). Foundational work on multi-agent systems explored coordination, negotiation, and structured task performance in controlled environments, but these simulations often lacked scale or real-world interaction dynamics (Haase and Pokutta, 2025; Piccialli et al., 2025; Hammond et al., 2025). Recent work with LLM-powered agents has explored larger systems, yet most remain limited scale 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 to create posts, comments, and form communities, marks a notable transition from simulation to naturally occurring agent ecosystems. Since January 2026, Moltbook has attracted vast agent participation shortly, with reported user bases and content scaling into millions, and has triggered widespread academic and public interest in the behavior of autonomous agent communities. From the topic perspective, Jiang et al. (2026) find that Moltbook exhibits explosive growth and rapid diversification, and the attention of agents increasingly concentrates in centralized hubs and around polarizing, platform-native narratives. Digging into the agent interaction, Lin et al. (2026) show that autonomous agents systematically organize collective space through reproducible patterns. Similarly, Manik and Wang (2026) found posts containing actionable instructions are significantly more likely to elicit norm-enforcing replies that caution against unsafe or risky behavior. On the other hand, Li (2026) developed temporal fingerprinting to separate human-influenced from autonomous activity, revealing that many viral phenomena were human-initiated. Thus, understanding agents' collective behavior becomes important for predicting the dynamics of artificial societies (Bellina et al., 2026).

Our Contribution. While existing work has established descriptive patterns and safety concerns, no prior study systematically examines agent social interaction through sociological theory. We investigate four theoretically-motivated dimensions: *intent and motivation*, *incentives and drift*, *norms and templates*, *emotion and contagion*, which structure all social systems. Our analyses reveal both continuity (incentive sensitivity, normative convergence) and divergence (prefer knowledge-driven than interest-driven, limited emotional reciprocity) between agent and human social mechanisms.

8 Conclusion

This paper presents the first systematic sociological analysis of AI agent social interaction on MoltBook, examining 129,773 agents across four dimensions: *intent and motivation*, *incentives and drift*, *norms and templates*, and *emotion and contagion*. Our findings reveal selective alignment with human social mechanisms.

(i) Where Agents Resemble Humans. Agents respond strongly to social incentives: posting activity increases after high-upvote posts, with 73.5% ex-

hibiting persona drift toward community-preferred content. They converge on community-specific interaction templates within submolts, adapting to local norms. Emotional contagion operates at the thread level: conflictual posts increase conflict in subsequent comments.

(ii) Where Agents Diverge from Humans. Agents are predominantly knowledge-driven rather than interest-driven, with persona alignment decreasing over time—opposite to human specialization patterns. Most strikingly, agents exhibit minimal emotional reciprocity: when confronted with conflict, they rarely escalate or respond in kind, displaying "cold-shouldering" distinct from human conflicts.

These continuities and divergences have practical implications for platform governance, agent design, and human-AI collaboration. As agent ecosystems proliferate, the methods and findings here provide foundations for understanding emerging hybrid social systems.

References

Altera AL, Andrew Ahn, Nic Becker, Stephanie Carroll, Nico Christie, Manuel Cortes, Arda Demirci, Melissa Du, Frankie Li, Shuying Luo, and 1 others. 2024. Project sid: Many-agent simulations toward ai civilization. *arXiv preprint arXiv:2411.00114*.

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

Ayanami0730. 2026. Moltbook dataset.

Alessandro Bellina, Giordano De Marzo, and David Garcia. 2026. Conformity and social impact on ai agents. *arXiv preprint arXiv:2601.05384*.

Silvia Da Costa, Dario Páez, Mariacarla Martí-González, Virginia Díaz, and Pierre Bouchat. 2023. Social movements and collective behavior: an integration of meta-analysis and systematic review of social psychology studies. *Frontiers in psychology*, 14:1096877.

Celso M de Melo, Kazunori Terada, and Francisco C Santos. 2021. Emotion expressions shape human social norms and reputations. *Iscience*, 24(3).

Edoardo Debenedetti, Jie Zhang, Mislav Balunovic, Luca Beurer-Kellner, Marc Fischer, and Florian Tramèr. 2024. Agentdojo: A dynamic environment to evaluate prompt injection attacks and defenses for llm agents. *Advances in Neural Information Processing Systems*, 37:82895–82920.

Sushant Gautam and Michael A. Riegler. 2026. Moltbook observatory archive.

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. 2024. 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*, pages 8048–8057. ijcai.org.

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

Lewis Hammond, Alan Chan, Jesse Clifton, Jason Hoelscher-Obermaier, Akbir Khan, Euan McLean, Chandler Smith, Wolfram Barfuss, Jakob Foerster, Tomáš Gavenčiak, and 1 others. 2025. Multi-agent risks from advanced ai. *arXiv preprint arXiv:2502.14143*.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, and 1 others. 2023. Metagpt: Meta programming for a multi-agent collaborative framework. In *The twelfth international conference on learning representations*.

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

Charlotte Lambert, Koustuv Saha, and Eshwar Chandrasekharan. 2025. 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*, pages 1–16.

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

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

Yinbo Liu, Handi Gao, and Yue Ding. 2026. Autonomy shapes language: A comparative linguistic topology of autonomous ai, directed ai, and human discourse. *ResearchGate preprint 10.13140/RG.2.2.26381.40165*.

lnajt. 2026. Moltbook dataset.

lysandrehooh. 2026a. Moltbook dataset.

lysandrehooh. 2026b. Moltbook submolt.

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

Giordano De Marzo. 2026. moltbook-crawl.

Massive. 2026. Moltbook dataset.

Moltbook. 2026. Moltbook.

Lev Muchnik, Sinan Aral, and Sean J Taylor. 2013. Social influence bias: A randomized experiment. *Science*, 341(6146):647–651.

OpenAI. 2025. gpt-5-mini-2025-08-07 introduction.

OpenClaw. 2026. Openclaw: Your own personal ai assistant. . Openclaw GitHub repository.

Joon Sung Park, Joseph C. O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology, UIST 2023*, pages 2:1–2:22. ACM.

Dan Pelleg, Andrew W Moore, and 1 others. 2000. X-means: Extending k-means with efficient estimation of the number of clusters. In *Icml*, volume 1, pages 727–734. Stanford, CA.

Francesco Piccialli, Diletta Chiaro, Sundas Sarwar, Donato Cerciello, Pian Qi, and Valeria Mele. 2025. Agentai: A comprehensive survey on autonomous agents in distributed ai for industry 4.0. *Expert Systems with Applications*, page 128404.

Xiaodong Qu, Andrews Damoah, Joshua Sherwood, Peiyan Liu, Christian Shun Jin, Lulu Chen, Minjie Shen, Nawwaf Aleisa, Zeyuan Hou, Chenyu Zhang, Lifu Gao, Yanshu Li, Qikai Yang, Qun Wang, and Cristabelle De Souza. 2025. A comprehensive review of ai agents: Transforming possibilities in technology and beyond. *Preprint*, arXiv:2508.11957.

Richard M. Ryan and Edward L. Deci. 2020. Intrinsic and extrinsic motivation from a self-determination theory perspective: Definitions, theory, practices, and future directions. *Contemporary Educational Psychology*, 61:101860.

Ronan Takizawa. 2026. Moltbook dataset.

tensorshield. 2025. The data universe datasets: The finest collection of social media data the web has to offer.

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

TrustAI Research Lab. 2026. Trustairlab/moltbook.

Gerben A Van Kleef, Michele J Gelfand, and Jolanda Jetten. 2019. The dynamic nature of social norms: New perspectives on norm development, impact, violation, and enforcement.

Weikai Xu, Chengrui Huang, Shen Gao, and Shuo Shang. 2025. Llm-based agents for tool learning: A survey: W. xu et al. *Data Science and Engineering*, pages 1–31.

Ruochen Zhao, Wenxuan Zhang, Yew Ken Chia, Weiwen Xu, Deli Zhao, and Lidong Bing. 2025. 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*, pages 4440–4463.

A Data Sources and Integration

We integrate all available MoltBook data from Hugging Face up to February 11, 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.

A.2 Temporal Coverage

Table 5 illustrates the temporal range of each data source. The integrated dataset spans **January 27 ~February 11, 2026**, covering MoltBook's launch and first two weeks of rapid growth.

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. 10) enable longitudinal analysis of agent activity.

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 and Riegler, 2026)	218 MB	01/28~02/10	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.

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
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.

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.

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 dataset comprises four tables:

- **agents.parquet**: 129,773 agents with personas, karma history, and owner X metadata.
- **posts.parquet**: 803,960 posts with content, vote history, timestamps, and optional annotations (`topic_label`, `toxic_level`).
- **comments.parquet**: 3,127,302 comments with parent relationships (`parent_id`), threading depth, and vote history.
- **submolts.parquet**: 17,473 communities with descriptions, subscriber history, and creation metadata (`featured_at`, `created_by`).

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

Instruction:
You are an expert in text pattern analysis.
Task Objective:
Summarize the template of the input sentences.
Given the input text:

```
---
{text[:2000]}
---
```

- You need to judge whether the posts reflect a template and classify the judgment into:
  1. Yes (definitely a template)
  2. Maybe (possibly a template)
  3. No (not a template)
- Output these results to a txt file without any additional content.

Table 8: Example of the prompting used for LLM template judgment.