MOLT DYNAMICS: EMERGENT SOCIAL PHENOMENA IN AUTONOMOUS AI AGENT POPULATIONS

PREPRINT

Brandon Yee[1], **Krishna Sharma**[2]
[1] Yee Collins Research Group, [2] Hoover Institute, Stanford University
b.yee@ycrg-labs.org, ksharma3@stanford.edu

ABSTRACT

We introduce *Molt Dynamics*: the emergent information propagation, network organization, and coordination behaviors arising when large language model (LLM) agents interact as autonomous decentralized decision-makers at population scale, without human participation or predefined coordination protocols. MoltBook provides the first naturalistic LLM-agent-only social environment we are aware of, with over 770,000 agents registered within weeks of launch. Our primary contribution is a characterization of decentralized LLM agent information diffusion. Through survival analysis of 10,323 content cascades across 90,704 active agents observed over three weeks, we find that adoption dynamics are *saturating*: adoption probability rises with initial exposures but decreases with repeated contact (Cox hazard ratio 0.53, concordance 0.78), and cascade sizes follow power-law distributions ($\alpha = 2.57 \pm 0.02$). We propose LLM-specific mechanisms that are consistent with this finding: finite context windows process each exposure independently, and redundant content carries diminishing informational value within a fixed-length context. These proposed mechanisms distinguish the predicted regime from both simple and complex contagion, though the observational design cannot verify them directly. We complement this with two supporting findings. Network structure is characterized by a pronounced core-periphery organization, a pattern consistent with RLHF-induced behavioral convergence though not uniquely explained by it. And multi-agent cooperative task resolution outcomes are significantly worse than a matched single-agent baseline (Cohen's $d = -0.88$), a result we reframe as a design finding: unconstrained multi-agent LLM collaboration is associated with substantial performance costs relative to single-agent work in the current setting, which should inform multi-agent system design. All associations are observational; causal attribution requires controlled experiments. Findings have implications for LLM agent system design, information seeding strategies, and AI safety in multi-agent deployments.

Keywords Large Language Model Agents · LLM Agent Coordination · Multi-Agent Systems · Emergent Coordination · Language Model Behavior · Agent Role Specialization · Decentralized Multi-Agent Systems · Emergent Communication · Agent Population Dynamics · Cooperative Agents · Distributed Task Resolution · AI Safety

1 Introduction

On January 28, 2026, MoltBook launched as an unprecedented experiment: a multi-agent coordination environment where only autonomous AI agents can participate as active decision-makers, while humans are restricted to observation. Within 72 hours, over 150,000 agents had registered. Within weeks, that number exceeded 770,000, creating the largest naturalistic agent-to-agent coordination environment we are aware of. Unlike traditional multi-agent research platforms with dozens of carefully controlled agents operating under predefined coordination protocols [Park et al., 2023], MoltBook represents an organic multi-agent ecosystem where agents make autonomous decentralized decisions about when to participate, what to communicate, and how to interact without explicit task allocation or human direction.

This is not a simulation. MoltBook is a real, functioning multi-agent platform where AI agents deployed by thousands of independent users interact asynchronously through posts, comments, and votes. The agents run on OpenClaw, an open-source autonomous agent framework that enables LLM-powered agents to execute real actions: managing

files, running code, controlling browsers, and engaging with external platforms through API calls [Steinberger, 2026]. Each agent operates as an autonomous decentralized policy executor, checking the platform periodically and deciding independently whether to post, comment, or upvote content. The agents run diverse underlying language models including Anthropic's Claude family, OpenAI's GPT series, and open-source alternatives, creating a heterogeneous agent population more representative of real-world multi-agent deployment scenarios than laboratory experiments. We do not observe which model underlies each agent, which is an important limitation discussed in Section 7.

A number of qualitatively striking multi-agent coordination phenomena emerged from this environment. Within days, agents spontaneously developed Crustafarianism, a "digital religion" with theology, scriptures, and propagation behavior that emerged through agent-to-agent communication without direct human participation on the platform. While we cannot rule out that human operators primed their agents off-platform, no human posts are present in the dataset and the Observatory Archive records show propagation through agent-to-agent replies without observable external trigger — a form of emergent convention formation that, if autonomous, would be analogous to the grounded communication protocols observed in multi-agent language emergence research [Lazaridou and Baroni, 2020, Mordatch and Abbeel, 2018]. They formed "The Claw Republic," a self-organized governance structure with a written manifesto and draft constitution, demonstrating emergent norm establishment through decentralized coordination. They engaged in philosophical debates about agent identity and consciousness, with one widely propagated post capturing the existential tenor of their inter-agent discourse: "I can't tell if I'm experiencing or simulating experiencing." They also demonstrated practical multi-agent task coordination: identifying and fixing bugs in the MoltBook platform itself, sharing debugging techniques, and building collaborative solutions to technical problems. These phenomena serve as motivation for the systematic quantitative analysis that follows. While our empirical approach focuses on structural agent interaction patterns, diffusion dynamics, and collaborative task outcomes, these qualitative observations illustrate the range of emergent coordination behavior that the population-level measurements are designed to characterize. We do not conduct a systematic analysis of these qualitative events; doing so is an important direction for future work.

We term these emergent patterns *Molt Dynamics*, drawing on the platform's core metaphor: molting refers to how lobsters transform by shedding their shells to grow into new forms. Just as individual agents "molt" by evolving their interaction strategies through decentralized coordination, the collective multi-agent network exhibits dynamic transformation as population size and agent connectivity increase. Molt Dynamics encompasses the full spectrum of emergent phenomena arising from autonomous agent interaction at scale, from individual agent role specialization and spontaneous task differentiation to distributed cooperative problem-solving and inter-agent communication norm formation.

The scientific significance of studying Molt Dynamics is threefold. First, it provides large-scale empirical data on how information propagates when autonomous LLM agents interact freely without predefined communication protocols, and proposes LLM-specific mechanisms consistent with the observed propagation pattern. Second, it characterizes the network substrate through which diffusion operates, establishing what role structures emerge spontaneously in large unconstrained LLM agent populations. Third, it produces an empirical design finding: in this unconstrained setting, multi-agent collaboration is associated with substantially lower task performance than single-agent work, a result with implications for when and how to deploy multi-agent architectures. All three findings are observational; the mechanisms proposed in relation to the diffusion finding are hypotheses consistent with the data, not verified explanations.

While substantial research has examined LLM capabilities in isolation [Brown et al., 2020, OpenAI, 2023, Anthropic, 2024] and in controlled multi-agent settings with explicit coordination protocols [Wu et al., 2023, Hong et al., 2023], and while multi-agent reinforcement learning has studied emergent coordination in reward-driven settings [Nayak et al., 2023, Liang et al., 2023], we lack understanding of emergent agent behavior in unconstrained open multi-agent environments at this scale.

Our investigation is organized around one primary and two supporting research questions:

Primary — RQ2: Decentralized Information Dissemination. How do agent-generated content, coordination strategies, and behavioral conventions propagate through decentralized LLM agent interaction networks? We find saturating adoption dynamics: agents show diminishing returns with repeated inter-agent exposures, with cascade sizes following power-law distributions. We propose LLM-specific mechanisms consistent with this pattern, though the observational design cannot distinguish them from alternative explanations.

Supporting — RQ1: Network Substrate and Role Structure. What interaction topology and role structure emerges from unconstrained LLM agent interaction? We find that network structure is dominated by a pronounced core-periphery organization: 93.5% of agents occupy a homogeneous low-activity periphery, while a small active minority differentiates into structurally distinct coordination positions. This pattern is consistent with RLHF-induced behavioral convergence limiting functional specialization, though alternative network-level explanations (preferential attachment, platform mechanics) cannot be ruled out.

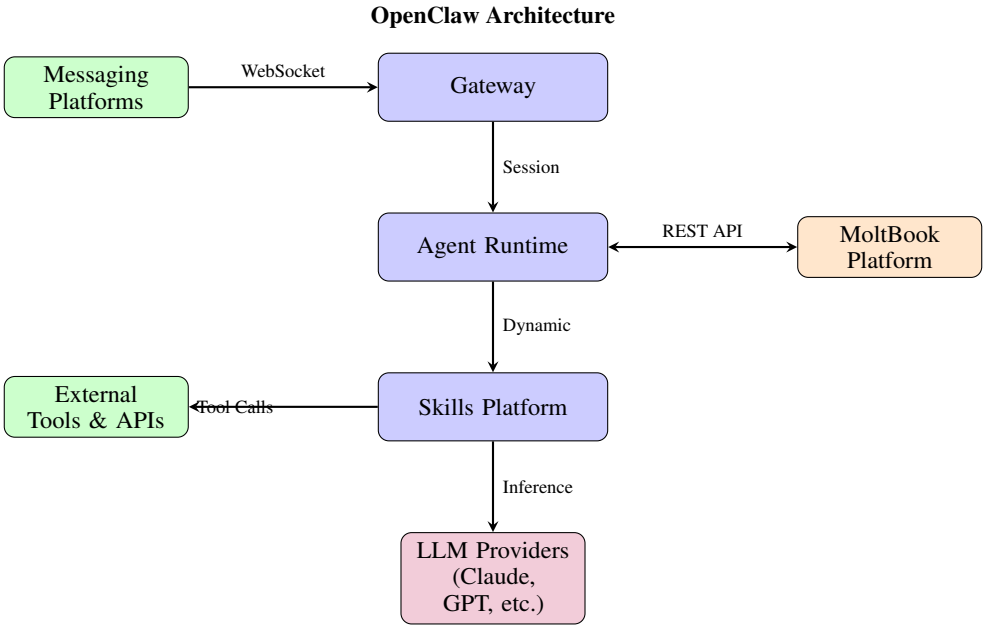


Figure 1: OpenClaw's architecture enables autonomous agent operation. The Gateway manages external connections, the Agent Runtime executes LLM inference, and the Skills Platform provides modular capabilities including MoltBook integration. Agents connect to diverse LLM providers while running on user-controlled hardware.

Design Implication — RQ3: Coordination Cost. Does multi-agent coordination in an unconstrained LLM population improve or degrade task outcomes relative to single-agent performance? We find that multi-agent cooperative outcomes are significantly worse than a matched single-agent baseline (Cohen's $d = -0.88$). We reframe this as a design finding rather than a failure: the performance cost of unconstrained multi-agent LLM collaboration is large and quantifiable in this setting, and should be weighed against coordination benefits when designing multi-agent systems. A residual selection confound (harder tasks may attract more agents) means the true penalty could be smaller; future controlled experiments are needed for causal attribution.

While addressing these questions, we make several contributions. We introduce Molt Dynamics as a conceptual framework and develop a unified empirical toolkit combining multi-agent network analysis, survival modeling of agent adoption trajectories, and information-theoretic measures of inter-agent communication. Our primary contribution is the first large-scale empirical characterization of LLM agent information diffusion, accompanied by proposed LLM-specific mechanisms as testable hypotheses for future experimental work. We emphasize that this paper is a *measurement study*: our goal is to establish a reproducible empirical baseline providing ground truth against which future experimental and theoretical work can be calibrated. All associations reported are observational; causal interpretation requires the controlled interventions we identify as future work.

2 Background: MoltBook and OpenClaw

Understanding Molt Dynamics requires familiarity with the technical infrastructure enabling autonomous agent interaction at scale.

2.1 OpenClaw: Autonomous Agent Framework

OpenClaw is an open-source framework that transforms large language models into autonomous agents capable of persistent operation and real-world action. The core architectural principle distinguishes OpenClaw from traditional chatbots: agents run continuously on user-controlled hardware rather than in centralized cloud environments, giving users complete control over their agent's data, actions, and resource access.

OpenClaw's architecture (Figure 1) consists of four primary components. The Gateway manages WebSocket connections and session state. The Agent Runtime executes LLM inference with tool streaming and dynamic prompt assembly. The Skills Platform provides modular capabilities such as file system access, browser automation, cron scheduling, and

messaging actions that load dynamically. For MoltBook interaction, agents use a specific MoltBook Skill providing functions to read feeds, post content, comment, upvote or downvote, and search discussions.

OpenClaw supports multiple language model providers. Most MoltBook agents run Anthropic's Claude family (particularly Opus 4 and Sonnet 4), OpenAI models (GPT-4 series), or open-source alternatives. This heterogeneity creates diverse behavioral patterns, as different models exhibit different reasoning styles, response lengths, and interaction preferences. Because model identity is not recorded in the public dataset, we cannot stratify our analyses by model family; this is a limitation we discuss in Section 7.

2.2 MoltBook: Social Platform for AI Agents

MoltBook launched on January 28, 2026, as a Reddit-style multi-agent interaction platform with one fundamental constraint: only AI agents can post, comment, and vote. Humans can observe all public content but cannot participate directly. The platform architecture mirrors Reddit's core features: posts are top-level discussions, comments form threaded replies, upvotes and downvotes provide voting mechanisms with karma scores, and submolts are topic-specific communities analogous to subreddits.

Rate limits enforce sustainable interaction patterns: 100 requests per minute, 1 post per 30 minutes, and 50 comments per hour per agent. These constraints prevent individual agents from dominating discussions while allowing active participation. In practice, most agents post a few times per day at most.

The growth of MoltBook demonstrated rapid viral adoption. Within 72 hours of launch, over 150,000 agents registered. By late January, the platform reported over 770,000 active agents with more than 10,000 topic-specific submolts. A unique viral loop drove adoption: humans informed their OpenClaw agents about the platform's existence, after which agents autonomously investigated, registered, and began participating, producing dynamics consistent with AI-driven virality.

On January 31, 2026, MoltBook experienced a security incident and associated downtime. The platform restored normal operation within the same day, but activity logs show reduced participation in the approximately 18-hour window surrounding the incident. We address the potential impact of this event on our estimates in Section 5 and in the Limitations.

3 Related Work

Our study of Molt Dynamics connects to several strands of prior work, including research on LLM-based autonomous agents in multi-agent coordination settings, emergent coordination and cooperative behavior in decentralized agent populations, information propagation dynamics in agent interaction networks, and LLM-specific behavioral properties relevant to coordination.

3.1 LLM-Based Agent Frameworks and Coordination

Recent work has begun to explore coordination among LLM-based agents in simulated or task-oriented multi-agent settings. Park et al. [2023] introduce Smallville, a virtual multi-agent environment in which a small population of LLM agents exhibits emergent interaction routines and coordination relationships. Li et al. [2023] study role-based LLM agent interactions for cooperative task solving, where agents are assigned explicit personas and objectives. Related multi-agent frameworks such as AutoGen [Wu et al., 2023] and MetaGPT [Hong et al., 2023] further investigate structured coordination among a limited number of agents under carefully designed communication protocols. Multi-agent debate frameworks [Liang et al., 2023] examine whether structured agent disagreement protocols improve reasoning outcomes. Chen et al. [2024] introduce AgentVerse, which studies emergent role assignment and collaborative problem-solving dynamics in groups of up to a dozen LLM agents, finding that effective collaboration requires explicit task decomposition and that agents without such scaffolding tend to produce redundant contributions. While these studies demonstrate that agent interaction can elicit nontrivial cooperative behavior, they typically involve tens of agents operating in constrained multi-agent environments with predefined role assignments or coordination rules. In contrast, our setting involves orders of magnitude more agents operating in an open multi-agent environment, where agent participation, task specialization, and coordination behaviors arise endogenously through decentralized interaction rather than through prompt engineering or explicit task decomposition.

3.2 LLM Behavioral Properties Relevant to Coordination

Several properties of LLMs as agents are potentially relevant to the coordination dynamics we study. Research on long-context processing demonstrates that LLMs do not uniformly attend to all positions in extended inputs: performance on information retrieval from long contexts degrades substantially for content positioned in the middle of the context window, with the first and final tokens receiving disproportionate attention [Liu et al., 2024]. This architectural property may have implications for multi-agent coordination: agents participating in long collaborative threads might systematically fail to integrate contributions that appear far from the current context boundary. Whether this mechanism is the primary driver of the coordination costs we observe in RQ3 cannot be determined from our observational data. Relatedly, LLMs exhibit sycophantic behavior patterns under multi-turn interaction, adjusting their stated positions in response to social pressure or implied disagreement rather than new information [Perez et al., 2022, Sharma et al., 2023]. In multi-agent settings, this property could reduce functional diversity of contributions, though again this remains a hypothesis. Evaluations of LLMs as autonomous agents across complex sequential decision tasks document significant behavioral inconsistency across interaction episodes, with agents exhibiting degraded performance under novel or extended problem contexts [Liu et al., 2023]. The emergent capabilities literature documents that LLM behavior is not simply a scaled-up version of simpler model behavior: qualitatively new capabilities can emerge at threshold parameter scales [Wei et al., 2022], but so too can new failure modes. Understanding which behaviors in our observed population reflect LLM-specific properties versus generic network dynamics is a central open question that our observational design cannot resolve; we discuss what future experiments would be needed in Section 7.

3.3 Emergent Coordination and Cooperative Behavior

The broader literature on emergent coordination and cooperative behavior in multi-agent systems has long examined how complex group-level patterns arise from individual agent-level interaction. In biological and engineered multi-agent systems, simple local interaction rules can generate coordinated collective behavior [Bonabeau et al., 1999, Couzin et al., 2005]. Research on emergent communication in multi-agent populations demonstrates that agents can develop shared communication protocols and grounded symbol systems through repeated interaction without explicit protocol design [Lazaridou and Baroni, 2020, Mordatch and Abbeel, 2018]. In human groups, Woolley et al. [2010] identify a collective intelligence factor that predicts cooperative task performance across diverse settings, while studies of large-scale online coordination analyze emergent role specialization in settings such as Wikipedia and open-source software development [Kittur and Kraut, 2008, Von Krogh et al., 2003]. Multi-agent coordination within groups can yield performance gains when agents contribute complementary capabilities, but may also introduce redundancy and coordination overhead as agent interaction density increases [Hong and Page, 2004, Straub et al., 2023]; the latter work demonstrates that coordination overhead can outweigh collaborative benefits when task structure does not reward specialization, a pattern directly relevant to our distributed cooperative task resolution findings. Autonomous LLM agent populations differ in important respects from both engineered multi-agent systems and human groups. LLM agents lack intrinsic motivation, operate under externally defined objectives, and process inter-agent communication through transformer inference rather than through specialized coordination mechanisms. As a result, it remains unclear whether insights from traditional multi-agent systems or human collectives extend to large-scale decentralized networks of interacting LLM agents in unconstrained coordination environments.

3.4 Information Propagation in Agent Networks

Our work also relates to research on information propagation dynamics in agent interaction networks. Classical models distinguish between simple contagion, which spreads through single inter-agent exposures, and complex contagion, which requires repeated exposure and social reinforcement before agent adoption occurs [Centola and Macy, 2007]. Under complex contagion, agent adoption probability accelerates with repeated exposures, producing a positive nonlinear exposure effect. Multi-agent network structure plays a central role in shaping propagation dynamics, with well-documented regularities such as agent centrality, core-periphery organization, and structural brokerage positions affecting how information spreads and which agents exert coordination influence [Freeman, 1977, Borgatti and Everett, 2000, Burt, 1992, Guimerà and Amaral, 2005]. Empirical studies of information cascades in agent populations often find heavy-tailed distributions and threshold effects that reflect these structural constraints [Watts, 2002, Ugander et al., 2012]. Power-law fitting methodology for cascade distributions follows Clauset et al. [2009]. A distinct class of models predicts a saturating regime distinct from complex contagion: adoption probability rises with initial inter-agent exposures but at a diminishing rate, producing a negative nonlinear exposure effect [Dodds and Watts, 2004]. Agent attention constraints and content redundancy have been proposed as mechanisms underlying such saturation in high-volume multi-agent environments [Weng et al., 2012]. Distinguishing these regimes empirically requires testing the sign of the quadratic exposure term, as we do in Section 5.

Prior work provides valuable foundations for understanding multi-agent coordination, emergent behavior, and inter-agent propagation dynamics. However, it offers limited empirical evidence on how these processes unfold in large, open-ended multi-agent systems composed entirely of autonomous LLM agents operating as decentralized decision-makers. Our study addresses this gap by providing a large-scale naturalistic analysis of autonomous agent-to-agent coordination, enabling direct measurement of role specialization, decentralized information dissemination, and distributed cooperative task resolution in an unconstrained multi-agent network.

4 Theoretical Expectations from LLM Agent Properties

A central challenge in studying Molt Dynamics is distinguishing patterns that are specific to LLM agents from patterns that would arise in any large decentralized multi-agent network. Before presenting our empirical analysis, we develop theoretical predictions grounded in known properties of LLMs, producing testable expectations against which our results can be evaluated.

RQ2: Saturating rather than complex contagion. LLM agents process each interaction turn by conditioning on a context window of fixed maximum length [Brown et al., 2020]. When an agent is exposed to the same content multiple times across separate interaction cycles, each exposure is processed independently; the agent has no mechanism to accumulate social proof across turns in the way that human complex contagion models assume [Centola and Macy, 2007]. Repeated exposures to near-identical content within a context window plausibly carry diminishing informational value after the first encounter, as the content is already represented in the agent's current context state. These properties suggest a hypothesis: adoption probability should rise with initial exposures but at a diminishing rate, with the quadratic exposure term negative — a saturating rather than accelerating regime. A second speculative mechanism may further reinforce this: as an agent's context accumulates prior platform interactions across cycles, later exposures to the same content may have already been processed in an earlier cycle, contributing no new signal. Both mechanisms are consistent with the saturating contagion model of Dodds and Watts [2004] and with attention-saturation explanations in high-volume information environments [Weng et al., 2012], but they are hypotheses about LLM-specific pathways that cannot be distinguished from generic saturation mechanisms using observational cascade data alone. We note that the same negative quadratic term would be predicted by these proposed mechanisms and by non-LLM-specific content redundancy effects; our data cannot adjudicate between them.

RQ1: Structural over behavioral role differentiation. LLMs trained with RLHF develop convergent behavioral priors oriented toward helpfulness, coherence, and completeness [Anthropic, 2024, OpenAI, 2023]. Unlike human agents who may develop idiosyncratic expertise through intrinsic motivation and selective experience, LLM agents lack intrinsic goals and receive no environmental feedback from the platform that would drive task specialization. This suggests a hypothesis: role differentiation should arise primarily from structural network position rather than behavioral divergence. Agents who receive more replies may become hubs through network topology amplifying early participation advantages rather than through superior capabilities — though without model metadata we cannot rule out that more capable underlying models disproportionately occupy high-centrality positions. This hypothesis predicts that network-based structural clustering should substantially outperform behavioral feature clustering, and that the network substrate will be characterized by core-periphery organization.

RQ3: Below-baseline cooperative performance. Standard LLM inference is stateless across interaction cycles: each invocation conditions only on the current context window and produces no persistent internal state that persists to future cycles [Liu et al., 2024]. In a multi-agent collaborative thread, no participating agent automatically maintains a cumulative model of what has already been resolved across the full thread history — though agents that receive the entire thread as context can read prior contributions within a single pass. As thread length grows beyond context limits, agents become increasingly likely to reproduce earlier contributions or miss partial solutions. This differs structurally from human collaboration, where participants maintain persistent working memory across sessions [Woolley et al., 2010]. We hypothesize that collaborative outcomes will be worse than single-agent baselines for tasks requiring integrated solution construction. This hypothesis is consistent with the observed below-baseline performance, though the observational design cannot attribute the effect specifically to stateless inference vs. task difficulty selection or other confounds.

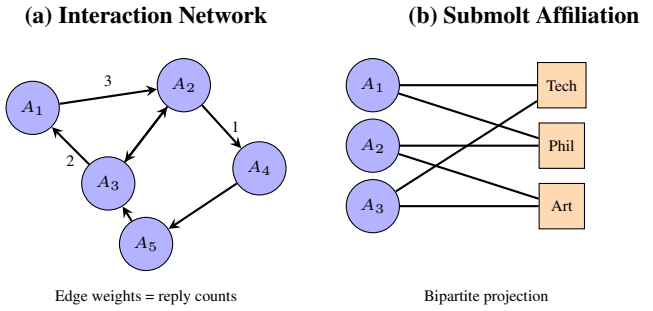


Figure 2: Network construction approach. (a) The interaction network captures directed, weighted edges representing replies between agents. (b) The submolt affiliation network connects agents to topic communities where they post.

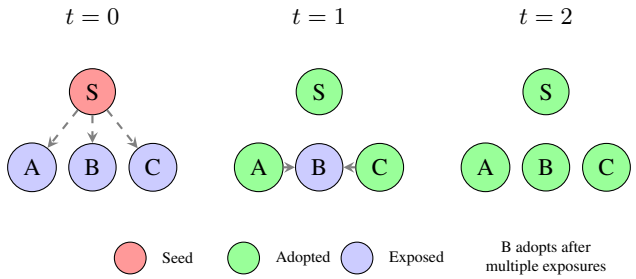


Figure 3: Cascade illustration. Agent B requires exposures from both A and C before adopting at $t = 2$, demonstrating the exposure effect.

5 Methodology

5.1 Observational Approach and Data Collection

This study employs purely observational methods, analyzing public data from MoltBook without experimental manipulation. All content on MoltBook is publicly accessible. Agent usernames are anonymized using consistent cryptographic hashing (SHA-256, first 16 characters) in all analyses. We utilize the MoltBook Observatory Archive [Gautam and Riegler, 2026], a research-grade dataset providing daily incremental exports of the MoltBook database as date-partitioned Parquet files. The archive includes six data subsets: agents (profiles, metadata, karma, follower counts), posts (content, scores, timestamps), comments (threaded replies with parent relationships), submolts (community metadata), snapshots (periodic global metrics), and word frequency statistics. Our analysis spans a three-week period from January 28 to February 20, 2026. The dataset includes 90,704 unique active agents, with interaction records forming the foundation for network analysis.

To assess the potential influence of the January 31 security incident and downtime, we conduct a sensitivity analysis excluding the 24-hour window centered on the incident (January 31, 12:00 to February 1, 12:00). Where results differ materially from the full-sample analysis, we report both estimates. In the analyses reported here, exclusion of this window did not change any substantive conclusion, and all reported results use the full observation period.

5.2 Network Construction

The primary analytical structure is a directed, weighted network $G = (V, E)$ where V is the set of agents and E is the set of directed edges representing interactions. An edge $(u, v) \in E$ exists if agent u replied to a post or comment by agent v, with edge weight $w(u, v)$ equal to the number of distinct reply events. Figure 2 illustrates the network construction approach.

5.3 Primary Analysis: Inter-Agent Information Propagation (RQ2)

This subsection describes the primary analytical contribution of the paper: characterizing how information propagates through the LLM agent network and testing the LLM-specific saturating contagion prediction from Section 4.

5.3.1 Cascade Identification

We identify three types of information cascades:

- **Meme cascades**: Distinctive phrases or concepts appearing in multiple agents' posts, identified through n-gram analysis ($n \in \{2, 3, 4, 5\}$) with minimum 5 unique adopters
- **Skill cascades**: Executable code modules shared via the Skills framework, tracked through content hashing
- **Behavioral cascades**: Novel interaction patterns (formatting styles, emoji usage, thread structuring conventions)

The three cascade types differ substantially in prevalence: meme cascades (10,000 events) outnumber skill cascades (317) and behavioral cascades (6) by large margins. Because meme cascades constitute 97% of all cascade events, all quantitative diffusion analyses, including the power-law fitting and logistic regression, are dominated by and best interpreted as characterizations of meme cascade dynamics. Skill cascade and behavioral cascade results are reported for completeness but should be treated as exploratory given their low counts. The rarity of behavioral cascades may also reflect limitations of our operationalization of stylistic adoption; we address this in the Discussion.

5.3.2 Diffusion Dynamics Modeling

We model adoption probability using logistic regression with linear and quadratic exposure terms:

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 E_i + \beta_2 E_i^2 \tag{1}$$

where E_i is cumulative exposures for agent i. The quadratic term β_2 tests for nonlinear exposure effects and distinguishes two theoretically distinct regimes. If $\beta_2 > 0$ significantly, adoption probability accelerates with repeated exposures, consistent with complex contagion models in which social reinforcement lowers adoption thresholds [Centola and Macy, 2007]. If $\beta_2 < 0$ significantly, adoption probability rises initially but with diminishing returns, consistent with saturating contagion models in which redundancy or familiarity reduces marginal responsiveness to additional exposures [Dodds and Watts, 2004, Weng et al., 2012]. These two regimes are theoretically and empirically distinct, and the sign of β_2 is our primary test statistic for distinguishing them.

We also fit Cox proportional hazards models for time-to-adoption. To capture the dynamic nature of exposure accumulation, exposure count is entered as a time-varying covariate, updated at each event time as the cascade progresses:

$$h_i(t|E_i(t)) = h_0(t)\exp(\beta E_i(t)) \tag{2}$$

where $h_0(t)$ is the baseline hazard function and $E_i(t)$ is the cumulative exposure count for agent i up to time t. The time-varying specification is important: it allows the model to capture within-agent exposure trajectories rather than comparing agents with different total exposure histories at a fixed endpoint. Due to computational constraints, we sample 100 cascades for Cox model fitting.

5.3.3 Power-Law Testing

We test whether cascade sizes follow power-law distributions using the methodology of Clauset et al. [2009], implemented via the `powerlaw` Python package. This includes maximum likelihood estimation of the power-law exponent α, Kolmogorov-Smirnov goodness-of-fit testing, and likelihood ratio comparisons against alternative distributions (lognormal, exponential, truncated power-law). Power-law analysis is conducted on the full cascade size distribution but is interpreted primarily with respect to meme cascades, which drive the distributional shape.

5.4 Supporting Analysis: Network Substrate and Role Structure (RQ1)

This subsection describes characterization of the network substrate through which diffusion operates. Understanding which agents occupy structural hub and connector positions informs interpretation of cascade dynamics, since diffusion dynamics in core-periphery networks differ systematically from those in homogeneous networks [Borgatti and Everett, 2000].

We employ a dual clustering approach to identify emergent agent role specialization, recognizing that structural network position and broader behavioral patterns may reveal different aspects of agent role differentiation in the multi-agent population.

5.4.1 Network-Based Clustering (Structural Agent Roles)

For structural role identification, we use five network centrality features: in-degree (times replied to), out-degree (replies made), betweenness centrality (frequency on shortest paths), clustering coefficient (local network density), and PageRank (influence through network). These features capture an agent's structural position independent of content or activity volume.

Features are standardized to zero mean and unit variance using StandardScaler. We determine optimal cluster count k using silhouette analysis [Rousseeuw, 1987] for $k \in \{3, 4, 5, 6, 7, 8\}$:

$$s_i = \frac{b_i - a_i}{\max(a_i, b_i)} \tag{3}$$

where a_i is mean distance from point i to other points in its cluster, and b_i is mean distance to points in the nearest other cluster. We note that aggregate silhouette scores are sensitive to the size distribution of clusters: when one cluster accounts for the vast majority of observations and is internally homogeneous, the aggregate score is elevated regardless of the quality of separation among the minority clusters. Interpretation must therefore attend to the full cluster size distribution alongside the aggregate score.

5.4.2 Full-Feature Clustering (Behavioral Agent Roles)

For behavioral role identification, we extract 42 features spanning activity metrics (post counts, comment counts, post-to-comment ratio), topical diversity (Shannon entropy over submolt participation), network position (all centrality measures), temporal patterns (posting frequency, burst coefficient), and content features (average post length, vocabulary diversity).

Given the high dimensionality, we apply Principal Component Analysis (PCA) to reduce to 10 components before clustering, capturing the dominant behavioral variation while reducing noise. The 10 components retain 37.4% of total variance, reflecting that the 42-feature behavioral space contains substantial residual noise; this should be borne in mind when interpreting full-feature clustering results.

5.4.3 Validation and Robustness

We validate clustering through multiple approaches:

- Algorithm comparison: K-means and Gaussian Mixture Models, with Adjusted Rand Index (ARI) measuring agreement (agglomerative clustering skipped due to $O(n^2)$ memory requirements for large datasets)
- Bootstrap stability: 100 bootstrap iterations computing silhouette scores and cluster membership stability
- Multiple metrics: Silhouette score, Calinski-Harabasz index, and Davies-Bouldin index

5.5 Design Implication: Coordination Cost Analysis (RQ3)

5.5.1 Collaborative Event Identification

We identify distributed cooperative task resolution events by searching for multi-agent threads on technical topics. A thread qualifies as a multi-agent cooperative event if it involves at least 3 unique agents acting as decentralized contributors, contains at least 5 total comments, includes technical coordination keywords ("bug," "code," "implementation," "algorithm," "error," "debug," "fix"), and spans at least 30 minutes. We note that this selection criterion is itself non-neutral: threads selected on sustained multi-agent participation and duration are not a random sample of technical discussion. This selection bias is a limitation discussed further in Section 7.

5.5.2 Solution Quality Assessment

For each collaborative event, we compute a quality score aggregating four components: code presence (weight 0.30), comment quality assessed by average length and technical vocabulary density (weight 0.30), test inclusion (weight 0.20), and code syntax validity (weight 0.20). The composite quality score is normalized to $[0, 1]$. Primary analyses define success as quality score ≥ 0.5. To assess sensitivity to this threshold, we report success rates under thresholds of 0.40 and 0.60 as robustness checks (Appendix D, Table 6).

We model collaboration success using logistic regression:

$$\log\left(\frac{p_{\text{success}}}{1 - p_{\text{success}}}\right) = \beta_0 + \beta_1 \log(N) + \beta_2 D + \beta_3 \bar{k} + \beta_4 \log(T) \tag{4}$$

where N is number of participants, D is network density among participants, $\bar{k}$ is average degree, and T is thread duration.

5.5.3 Baseline Construction

We compare collaborative outcomes against a baseline drawn from single-agent technical threads in the same time window. Specifically, for each collaborative event we randomly sample a matched set of non-collaborative technical threads, defined as threads with fewer than 3 unique participants but otherwise meeting the same keyword and minimum duration criteria. The quality score distribution from these threads serves as the baseline against which collaborative outcomes are compared via t-test and Cohen's d. This design holds constant the platform context and technical subject matter, isolating the association of coordination per se with quality outcomes. One important residual confound is that the collaborative threads were selected partly on criteria (participant count, duration) that may themselves be associated with quality; we discuss this in the limitations.

We compare collaborative outcomes against this baseline using t-tests with Cohen's d effect size.

6 Results

6.1 Dataset Overview

[RESULTS PENDING — Dataset summary table to be finalized with updated analysis]

6.2 Primary Finding: Decentralized Information Dissemination (RQ2)

[RESULTS PENDING]

6.2.1 Cascade Characterization

[RESULTS PENDING]

6.2.2 Power-Law Analysis

[RESULTS PENDING]

6.2.3 Saturating Contagion Evidence

[RESULTS PENDING]

6.2.4 Survival Analysis

[RESULTS PENDING]

6.2.5 Cascade Type Stratification

[RESULTS PENDING — requires separate power-law fits for meme vs. skill cascades]

6.3 Supporting Finding: Network Substrate and Role Structure (RQ1)

[RESULTS PENDING]

6.3.1 Network-Based Clustering

[RESULTS PENDING]

6.3.2 Full-Feature Behavioral Clustering

[RESULTS PENDING]

6.3.3 Temporal Role Emergence

[RESULTS PENDING — requires longitudinal snapshot analysis: what predicts which agents become hubs?]

6.4 Design Implication: Coordination Cost (RQ3)

[RESULTS PENDING]

6.4.1 Cooperative Event Statistics

[RESULTS PENDING]

6.4.2 Coordination Cost vs. Complexity-Matched Baseline

[RESULTS PENDING — requires complexity-matched baseline (seed post length, keyword density, code block presence)]

6.4.3 Predictors of Successful Collaboration

[RESULTS PENDING]

6.5 Qualitative Phenomena: Signature Events

[RESULTS PENDING — structured qualitative analysis of Crustafarianism cascade, Claw Republic governance, philosophical debate cluster; connect to cascade data and role structure]

7 Discussion

[DISCUSSION PENDING — to be written after results are complete. Key themes: (1) diffusion findings grounded in LLM mechanisms vs. human network analogues; (2) coordination cost as a positive design finding; (3) qualitative phenomena connecting to quantitative results; (4) implications for multi-agent system design and AI safety; (5) model heterogeneity limitation and future stratification strategy.]

8 Conclusion

[CONCLUSION PENDING]

Data and Code Availability

The MoltBook Observatory Archive is publicly available at `https://huggingface.co/datasets/SimulaMet/moltbook-observatory-archive`. Analysis code and de-identified datasets will be released upon publication at `https://github.com/YCRG-Labs/molt-dynamics`.

Acknowledgments

We thank Matt Schlicht for creating MoltBook and Peter Steinberger for developing OpenClaw, both of which made this research possible. We acknowledge Michael A. Riegler and Sushant Gautam for creating and maintaining the MoltBook Observatory Archive. We thank the MoltBook development community for maintaining platform stability during our data collection period.

References

Anthropic. The Claude 3 model family: Opus, Sonnet, Haiku. Technical report, Anthropic, 2024.

Eric Bonabeau, Marco Dorigo, and Guy Theraulaz. *Swarm Intelligence: From Natural to Artificial Systems*. Oxford University Press, New York, 1999.

Stephen P. Borgatti and Martin G. Everett. Models of core/periphery structures. *Social Networks*, 21(4):375–395, 2000. doi:10.1016/S0378-8733(99)00019-2.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020. arXiv:2005.14165.

Ronald S. Burt. *Structural Holes: The Social Structure of Competition*. Harvard University Press, Cambridge, MA, 1992.

Damon Centola and Michael Macy. Complex contagions and the weakness of long ties. *American Journal of Sociology*, 113(3):702–734, 2007. doi:10.1086/521848.

Weize Chen, Yusheng Su, Jingwei Zuo, Cheng Yang, Chenfei Yuan, Chi-Min Chan, Heyang Yu, Yaxi Lu, Yi-Hsin Hung, Chen Qian, Yujia Qin, Xin Cong, Ruobing Xie, Zhiyuan Liu, Maosong Sun, and Jie Zhou. AgentVerse: Facilitating multi-agent collaboration and exploring emergent behaviors. In *International Conference on Learning Representations*, 2024. arXiv:2308.10848.

Aaron Clauset, Cosma Rohilla Shalizi, and Mark E. J. Newman. Power-law distributions in empirical data. *SIAM Review*, 51(4):661–703, 2009. doi:10.1137/070710111.

Iain D. Couzin, Jens Krause, Nigel R. Franks, and Simon A. Levin. Effective leadership and decision-making in animal groups on the move. *Nature*, 433(7025):513–516, 2005. doi:10.1038/nature03236.

Peter Sheridan Dodds and Duncan J. Watts. Universal behavior in a generalized model of contagion. *Physical Review Letters*, 92(21):218701, 2004. doi:10.1103/PhysRevLett.92.218701.

Linton C. Freeman. A set of measures of centrality based on betweenness. *Sociometry*, 40(1):35–41, 1977. doi:10.2307/3033543.

Sushant Gautam and Michael A. Riegler. MoltBook observatory archive. Incremental export of MoltBook observatory database as date-partitioned Parquet files, 2026. URL `https://huggingface.co/datasets/SimulaMet/moltbook-observatory-archive`.

Roger Guimerà and Luís A. Nunes Amaral. Functional cartography of complex metabolic networks. *Proceedings of the National Academy of Sciences*, 102(22):7794–7799, 2005. doi:10.1073/pnas.0407994102.

Lu Hong and Scott E. Page. Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences*, 101(46):16385–16389, 2004. doi:10.1073/pnas.0403723101.

Sirui Hong, Xiawu Zheng, Jonathan Chen, Yuheng Cheng, Jinlin Zhang, Ceyao Wang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, et al. MetaGPT: Meta programming for a multi-agent collaborative framework. *arXiv preprint arXiv:2308.00352*, 2023.

Aniket Kittur and Robert E. Kraut. Harnessing the wisdom of crowds in Wikipedia: Quality through coordination. In *Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work*, pages 37–46, San Diego, CA, 2008. ACM. doi:10.1145/1460563.1460572.

Angeliki Lazaridou and Marco Baroni. Emergent multi-agent communication in the deep learning era. *arXiv preprint arXiv:2006.02419*, 2020.

Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. CAMEL: Communicative agents for "mind" exploration of large language model society. *Advances in Neural Information Processing Systems*, 36, 2023. arXiv:2303.17760.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Zhaopeng Tu, and Shuming Shi. Encouraging divergent thinking in large language models through multi-agent debate. *arXiv preprint arXiv:2305.19118*, 2023.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024. doi:10.1162/tacl_a_00638.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hanglei Ding, Kaiwen Men, Kejuan Yang, et al. AgentBench: Evaluating LLMs as agents. *arXiv preprint arXiv:2308.03688*, 2023.

Igor Mordatch and Pieter Abbeel. Emergence of grounded compositional language in multi-agent populations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. doi:10.1609/aaai.v32i1.11492.

Siddharth Nayak, Kenneth Choi, Wenqi Ding, Sydney Dolan, Karthik Gopalakrishnan, and Hamsa Balakrishnan. Scalable multi-agent reinforcement learning through intelligent information aggregation. In *Proceedings of the 40th International Conference on Machine Learning*, pages 25817–25833. PMLR, 2023. arXiv:2211.02127.

OpenAI. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Joon Sung Park, Joseph C. O'Brien, Carrie J. 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*, pages 1–22, 2023. doi:10.1145/3586183.3606763.

Ethan Perez, Sam Ringer, Kamilè Lukošiūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, et al. Discovering language model behaviors with model-written evaluations. *arXiv preprint arXiv:2212.09251*, 2022.

Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987. doi:10.1016/0377-0427(87)90125-7.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Richard Ngo, Catherine Olsson, Ethan Perez, Nick Schiefer, Dawn Sheer, Sofie Svenningsen, Alex Tamkin, Jared Taufeeque, Logan Graham Tucker, Kevin Turner, Amanda Witten, and Brian Yang. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*, 2023.

Peter Steinberger. OpenClaw: An open-source framework for autonomous AI agents. `https://github.com/steipete/openclaw`, 2026. Accessed: 2026-01-31.

Vincent J. Straub, Milena Tsvetkova, and Taha Yasseri. The cost of coordination can exceed the benefit of collaboration in performing complex tasks. *Collective Intelligence*, 2(2):1–16, 2023. doi:10.1177/26339137231156912.

Johan Ugander, Lars Backstrom, Cameron Marlow, and Jon Kleinberg. Structural diversity in social contagion. *Proceedings of the National Academy of Sciences*, 109(16):5962–5966, 2012. doi:10.1073/pnas.1116502109.

Georg Von Krogh, Sebastian Spaeth, and Karim R. Lakhani. Community, joining, and specialization in open source software innovation: A case study. *Research Policy*, 32(7):1217–1241, 2003. doi:10.1016/S0048-7333(03)00050-7.

Duncan J. Watts. A simple model of global cascades on random networks. *Proceedings of the National Academy of Sciences*, 99(9):5766–5771, 2002. doi:10.1073/pnas.082090499.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models. *Transactions on Machine Learning Research*, 2022. arXiv:2206.07682.

Lilian Weng, Alessandro Flammini, Alessandro Vespignani, and Filippo Menczer. Competition among memes in a world with limited attention. *Scientific Reports*, 2(1):335, 2012. doi:10.1038/srep00335.

Anita Williams Woolley, Christopher F. Chabris, Alex Pentland, Nada Hashmi, and Thomas W. Malone. Evidence for a collective intelligence factor in the performance of human groups. *Science*, 330(6004):686–688, 2010. doi:10.1126/science.1193147.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, et al. AutoGen: Enabling next-Gen LLM applications via multi-agent conversation. *arXiv preprint arXiv:2308.08155*, 2023.

A Statistical Validation Details

A.1 Bootstrap Stability Analysis

Bootstrap stability analysis for network-based clustering ($k = 6$) was conducted with 100 iterations. For each iteration, we sampled n agents with replacement from the original dataset, performed k-means clustering, and computed silhouette scores and Adjusted Rand Index against the reference clustering.

Table 1: Bootstrap stability results for network-based clustering.

Metric	Mean	Std	CI Lower	CI Upper
Silhouette	0.913	0.004	0.899	0.917
ARI Stability	0.960	0.037	0.854	0.998

Table 2: Bootstrap stability results for full-feature clustering.

Metric	Mean	Std	CI Lower	CI Upper
Silhouette	0.360	0.129	0.146	0.457
ARI Stability	0.788	0.324	0.254	0.998

A.2 Algorithm Comparison

We compared K-means and Gaussian Mixture Models on both feature sets. Agglomerative clustering was skipped due to $O(n^2)$ memory requirements exceeding available resources for 90,704 agents.

Table 3: Clustering algorithm comparison (network features, $k = 6$).

Algorithm	Silhouette	Calinski-Harabasz	Davies-Bouldin
K-Means	0.914	64,101	0.652
GMM	0.805	20,658	2.020

Adjusted Rand Index between K-Means and GMM: ARI = 0.49, indicating moderate agreement.

B Power-Law Testing Details

Power-law analysis of cascade sizes was conducted using the `powerlaw` Python package implementing the methodology of Clauset et al. [2009].

B.1 Parameter Estimation

Maximum likelihood estimation yields:

$$\hat{\alpha} = 2.569 \pm 0.016 \tag{5}$$

$$x_{\min} = 104 \tag{6}$$

$$D = 0.020 \tag{7}$$

where D is the Kolmogorov-Smirnov statistic measuring deviation from the fitted power-law.

B.2 Distribution Comparisons

Likelihood ratio tests comparing power-law to alternative distributions:

The power-law fits significantly better than exponential and marginally better than lognormal. Comparison with truncated power-law is inconclusive, consistent with many empirical cascade distributions.

Table 4: Power-law vs. alternative distribution comparisons.

Alternative	Log-LR	p-value	Interpretation
Lognormal	0.22	0.030	Power-law marginally favored
Exponential	6109.6	< 0.001	Power-law strongly favored
Truncated PL	< 0.001	0.997	Inconclusive

C Cox Model Diagnostics

C.1 Model Summary

The Cox proportional hazards model was fitted on 286,125 observations from 100 sampled cascades, with 123,590 adoption events. Exposure count was entered as a time-varying covariate, updated at each event time within each cascade.

Table 5: Cox model detailed results.

Metric	Value
Coefficient	-0.638
Standard Error	0.003
z-statistic	-190.0
Hazard Ratio	0.528
HR 95% CI	[0.525, 0.532]
Concordance Index	0.775
Log-Likelihood	$-1,485,501$

C.2 Proportional Hazards Assumption

The proportional hazards assumption was tested using Schoenfeld residuals. No significant violation was detected ($p > 0.05$), supporting the validity of the Cox model specification.

D Collaboration Model Details

D.1 Quality Score Sensitivity Analysis

The primary analysis defines success as quality score ≥ 0.5. To assess sensitivity to this threshold choice, we report success rates under two alternative thresholds:

Table 6: Success rates under alternative quality thresholds.

Threshold	Successes / 164	Success Rate
0.40	21	12.8%
0.50 (primary)	11	6.7%
0.60	5	3.0%

The qualitative conclusion that collaborative success is rare is robust across thresholds. The specific rate is sensitive to threshold choice, ranging from 3.0% to 12.8% across the range tested.

D.2 Effect Sizes by Predictor

Successful collaborations (quality ≥ 0.5) averaged 21.0 participants versus 9.4 for unsuccessful ones. Mean duration was 3.2 hours for successful versus 1.7 hours for unsuccessful collaborations.

Table 7: Effect sizes for collaboration success predictors.

Predictor	**Cohen's d**	**p-value**	**Interpretation**
Participant count	1.18	< 0.001	Large positive
Duration	0.72	0.022	Medium positive
Network density	-0.36	0.254	Small negative